GENETIC SUSCEPTIBILITY TO TUBERCULOSIS

Lucy J. Boothroyd, B.Sc.
Department of Epidemiology and Biostatistics
McGill University, Montreal
July, 1994

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

Name Lucy J Brailsford

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Health Sciences - Public Health

SUBJECT TERM

0573

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture	0729
Art History	0377
Cinema	0900
Dance	0378
Fine Arts	0357
Information Science	0723
Journalism	0391
Library Science	0399
Mass Communications	0708
Music	0413
Speech Communication	0459
Theater	0465

EDUCATION

General	0515
Administration	0514
Adult and Continuing	0516
Agricultural	0517
Art	0273
Bilingual and Multicultural	0282
Business	0688
Community College	0275
Curriculum and Instruction	0727
Early Childhood	0518
Elementary	0524
Finance	0277
Guidance and Counseling	0519
Health	0680
Higher	0745
History of	0520
Home Economics	0278
Industrial	0521
Language and Literature	0279
Mathematics	0280
Music	0522
Philosophy of	0998
Physical	0523
Psychology	0525
Reading	0535
Religious	0527
Sciences	0714
Secondary	0533
Social Sciences	0534
Sociology of	0340
Special	0529
Teacher Training	0530
Technology	0710
Tests and Measurements	0288
Vocational	0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language	
General	0679
Ancient	0289
Linguistics	0290
Modern	0291
Literature	
General	0401
Classical	0294
Comparative	0295
Medieval	0297
Modern	0298
African	0316
American	0591
Asian	0305
Canadian (English)	0352
Canadian (French)	0355
English	0593
Germanic	0311
Latin American	0312
Middle Eastern	0315
Romance	0313
Slavic and East European	0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy	0422
Religion	
General	0318
Biblical Studies	0321
Clergy	0319
History of	0320
Philosophy of	0322
Theology	0469

SOCIAL SCIENCES

American Studies	0323
Anthropology	
Archaeology	0324
Cultural	0326
Physical	0327
Business Administration	
General	0310
Accounting	0272
Banking	0770
Management	0454
Marketing	0338
Canadian Studies	0385
Economics	
General	0501
Agricultural	0503
Commerce Business	0505
Finance	0508
History	0509
Labor	0510
Theory	0511
Folklore	0358
Geography	0366
Gerontology	0351
History	
General	0578
Ancient	0579
Medieval	0581
Modern	0582
Black	0328
African	0331
Asia, Australia and Oceania	0332
Canadian	0334
European	0335
Latin American	0336
Middle Eastern	0333
United States	0337
History of Science	0585
Law	0398
Political Science	
General	0615
International Law and Relations	0616
Public Administration	0617
Recreation	0814
Social Work	0452
Sociology	
General	0626
Criminology and Penology	0627
Demography	0938
Ethnic and Racial Studies	0631
Individual and Family Studies	0628
Industrial and Labor Relations	0629
Public and Social Welfare	0630
Social Structure and Development	0700
Theory and Methods	0344
Transportation	0709
Urban and Regional Planning	0999
Women's Studies	0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture	
General	0473
Agronomy	0285
Animal Culture and Nutrition	0475
Animal Pathology	0476
Food Science and Technology	0359
Forestry and Wildlife	0478
Plant Culture	0479
Plant Pathology	0480
Plant Physiology	0817
Range Management	0777
Wood Technology	0746
Biology	
General	0306
Anatomy	0287
Biostatistics	0308
Botany	0309
Cell	0379
Ecology	0329
Entomology	0353
Genetics	0369
Limnology	0793
Microbiology	0410
Molecular	0307
Neuroscience	0317
Oceanography	0416
Physiology	0433
Radiation	0821
Veterinary Science	0778
Zoology	0472
Biophysics	
General	0786
Medical	0760

EARTH SCIENCES

Biogeochemistry	0425
Geochemistry	0996
Geodesy	0370
Geology	0372
Geophysics	0373
Hydrology	0388
Mineralogy	0411
Paleobotany	0345
Paleoecology	0426
Paleontology	0418
Paleozoology	0985
Palynology	0427
Physical Geography	0368
Physical Oceanography	0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences	0768
Health Sciences	
General	0566
Audiology	0300
Chemotherapy	0992
Dentistry	0567
Education	0350
Hospital Management	0769
Human Development	0758
Immunology	0982
Medicine and Surgery	0564
Mental Health	0347
Nursing	0569
Nutrition	0570
Obstetrics and Gynecology	0380
Occupational Health and Therapy	0354
Ophthalmology	0381
Pathology	0571
Pharmacology	0419
Pharmacy	0572
Physical Therapy	0382
Public Health	0573
Radiology	0574
Recreation	0575
Speech Pathology	0460
Toxicology	0383
Home Economics	0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry	
General	0485
Agricultural	0749
Analytical	0486
Biochemistry	0487
Inorganic	0488
Nuclear	0738
Organic	0490
Pharmaceutical	0491
Physical	0494
Polymer	0495
Radiation	0754
Mathematics	0405
Physics	
General	0605
Acoustics	0986
Astronomy and Astrophysics	0606
Atmospheric Science	0608
Atomic	0748
Electronics and Electricity	0607
Elementary Particles and High Energy	0798
Fluid and Plasma	0759
Molecular	0609
Nuclear	0610
Optics	0752
Radiation	0756
Solid State	0611
Statistics	0463

Applied Sciences

Applied Mechanics	0346
Computer Science	0984
Engineering	
General	0537
Aerospace	0538
Agricultural	0539
Automotive	0540
Biomedical	0541
Chemical	0542
Civil	0543
Electronics and Electrical	0544
Heat and Thermodynamics	0348
Hydraulic	0545
Industrial	0546
Marine	0547
Materials Science	0794
Mechanical	0548
Metallurgy	0743
Mining	0551
Nuclear	0552
Packaging	0549
Petroleum	0765
Sanitary and Municipal	0554
System Science	0790
Geotechnology	0428
Operations Research	0796
Plastics Technology	0795
Textile Technology	0994

PSYCHOLOGY

General	0621
Behavioral	0384
Clinical	0622
Developmental	0620
Experimental	0623
Industrial	0624
Personality	0625
Physiological	0989
Psychobiology	0349
Psychometrics	0632
Social	0451

Abstract

A genetic epidemiological approach was used to study inherited susceptibility as a risk factor for tuberculous disease follow.ng infection with *Mycobacterium tuberculosis*. Based on an experimental mouse model, the hypothesis was that a human recessive susceptibility gene is located on the long arm of chromosome 2 and has a major effect on the development of disease.

Clinical and genetic data from family members were used in linkage analyses with the lod score method to test for co-segregation of genetic markers and the susceptibility trait. Evidence for linkage to the candidate chromosome region was obtained for a large, multiplex Aboriginal Canadian family who experienced a tuberculosis epidemic in 1987–1989. Evidence against linkage was found in 16 smaller families from Colombia and Hong Kong, who had endemic but not epidemic exposure. The results suggested that the function of a chromosome 2 susceptibility gene may be more important in an epidemic situation.

Résumé

Une approche épidémiologique fut utilisée pour étudier la susceptibilité génétique de développer la tuberculose-maladie suite à une infection avec Mycobacterium tuberculosis. A partir d'un modèle expérimental chez la souris, cette étude voulait vérifier l'hypothèse de l'existence d'un allèle récessif de susceptibilité à la tuberculose situé sur la bras q du chromosome 2 qui aurait une influence majeure sur la probabilité de développer la maladie suite à l'infection.

La méthode lod score a été utilisée lors de l'analyse de liaison des données cliniques et génétiques des membres des familles étudiées afin de démontrer la ségrégation des marquers génétiques avec le statut de susceptibilité. Les éléments en faveur d'une liaison génétique avec la région chromosomique-candidate ont été obtenus dans une grande famille Autochtone du Canada ayant connu plusiers cas de tuberculose en contexte épidémique entre 1987 et 1989. Les éléments allant à l'encontre de la liaison génétique ont été obtenus dans 16 petites familles de Colombie et de Hong Kong exposées en contexte endémique plutôt qu'épidémique. Ces résultats suggèrent que la fonction d'un allèle de susceptibilité au chromosome 2 serait plus importante en contexte épidémique.

Acknowledgements

I would first like to thank the family members in Canada, Colombia, and Hong Kong who participated in this study. In particular, I thank the contact member of the Canadian family for her patience and cooperation during our interviews. I extend my gratitude to the following persons who were crucial for the collection of the families and clinical and epidemiological information: Dr. E.A. Fanning, Alberta Health Tuberculosis Services, Edmonton, Alberta; Dr. M. Miller, Departments of Microbiology and Infectious Diseases, Sir Mortimer B. Davis Jewish General Hospital, Montreal, Quebec; Dr. L. Garcia, Universidad de Antioquia, Medellin, Colombia; Dr. D.A. Higgins, Department of Pathology, University of Hong Kong, Hong Kong; Dr. E. Schurr, Department of Experimental Medicine, McGill University, Montreal, Quebec; and other international personnel who were involved in completing the study questionnaires. Additional data used in the study were obtained with diabetes families, whose DNA was purchased from the Human Biological Data Interchange (U.S.).

At the Montreal General Hospital Research Institute in the laboratories of Dr. K. Morgan and Dr. E. Schurr, I thank many persons for their work, in particular Ms. J. Liu and Ms. A.J. Paradis, and also Ms. I. Anacleto, Ms. N. Buu, Ms. D. Frappier, Mr. H. Kim, Dr. F. Sanchez, and Dr. E. Schurr. I wish to thank Ms. M. Fujiwara for her supervision of laboratory work, text editing, and project assistance, and Ms. L. Simkin in the Computer Core Facility for Genetic Data Analysis (Canadian Genetic Diseases Network of Centres of Excellence) for her database work, programming, and computer assistance. My sincere thanks are extended to Dr. T. Gyorkos for her helpful comments on thesis drafts. I wish to thank my supervisor, Dr. K. Morgan, for extensive discussions of the project; research funding provided by the Canadian Genetic Diseases Network of Centres of Excellence; and the opportunity to attend the Advanced Linkage Course, Columbia University, New York, as well as three informative conferences, including the International Genetic Epidemiology Society First Annual Meeting, Minneapolis (1992), the World Congress on Tuberculosis, Bethesda (1992), and the National Workshop on Tuberculosis, HIV and Other Emerging Issues, Toronto (1993). I thank the Medical Research Council of Canada for their funding support from 1991–1993 in the form of an MRC studentship.

Finally, I would like to express my gratitude to those who, with patience and understanding, gave me advice and moral support while I completed this thesis, particularly Debby, Susan, Ditty, and Ted, and especially my parents, Jean and Roy, and Michiel.

Preface

This thesis presents a linkage analysis study of tuberculosis in families from three continents. The primary objective of the study was to test for linkage of disease susceptibility to loci located on the long arm of human chromosome 2, based on an experimental mouse model of the early immune response to *Mycobacterium bovis*, developed by researchers in the McGill Centre for the Study of Host Resistance (director, Dr. E. Skamene). In mice, control over growth of *M. bovis* in the first three weeks after intra-venous inoculation is determined by a single, dominant gene called *Bcg*, localized to a region of mouse chromosome 1 that shows linkage conservation with a region of human chromosome 2q (Gros et al., 1981; Schurr et al., 1990b). The *Bcg* gene is expressed in mature macrophages; these cells have a significant role in the early response to mycobacterial infection, prior to cell-mediated immunity (Gros et al., 1983). Evidence for a disease susceptibility gene in families infected with *Mycobacterium tuberculosis* could lead to a better understanding of the interactions between the host and the infecting organism and the applicability of animal genetic models to the study of certain human infectious diseases.

The collection of the study families was carried out by investigators in Canada, Colombia, and Hong Kong in the late 1980s. Molecular genetic analysis was carried out at the Montreal General Hospital Research Institute between 1988 and 1994 in the laboratories of Dr. K. Morgan and Dr. E. Schurr, under the supervision of Ms. M. Fujiwara and Dr. E. Schurr, respectively. I became involved in the study in 1992, at which time a database had been designed and managed by Ms. L. Simkin, and contained disease and marker typing data for the 19 families in the study population. Additional typing data was generated as the study continued. The database management and linkage analysis were carried out in the Computer Core Facility for Genetic Data Analysis of the Canadian Genetic Diseases Network of Centres of Excellence, under the supervision of Dr. K. Morgan.

The five Colombian and the 13 Hong Kong families that constitute the international study population reside in regions with high endemic rates of tuberculosis; these rates have remained fairly stable or decreased over the past thirty years. In

comparison, the Aboriginal Canadian family in the study was recently exposed to a tuberculosis epidemic, during 1987 to 1989, and the majority of the family members had no evidence of mycobacterial exposure prior to the outbreak. This family was originally ascertained in 1989 for a study of diagnostic tests for tuberculosis and linkage analysis with HLA-A, -B, and -C typing, as part of a Master's thesis in the Department of Epidemiology and Biostatistics, McGill University (Miller, 1991).

The selection, diagnosis, and blood sampling of the Colombian and Hong Kong families, as well as the collection of pedigree diagrams, was carried out by two collaborating centres following ethics approval, under the supervision of Dr. L. Garcia, Universidad de Antioquia, Medellin, Colombia and Dr. D.A. Higgins, University of Hong Kong, Hong Kong. Preparation of samples from these families was carried out both internationally and in Dr. E. Schurr's laboratory at the Montreal General Hospital Research Institute. The Aboriginal Canadian family was located through Alberta Health Tuberculosis Services (director, Dr. E.A. Fanning) which provided extensive clinical details from the outbreak investigation. Informed consent for blood sampling, genetic testing, and medical chart review was obtained and collection of blood samples and pedigree structure was carried out by Dr. M. Miller, Dr. E.A. Fanning and Dr. E. Schurr in September, 1989. Clinical status of the Canadian family members was determined by Dr. M. Miller with the assistance of Dr. E.A. Fanning and was used in the present thesis.

I was responsible for the data verification and management of the existing database. In addition, I developed a detailed questionnaire for the collaborators in Colombia and Hong Kong to document the family selection and collection methods, as well as to validate the clinical and epidemiological data. I updated the database with the clinical phenotype information and the additional genetic marker results. I also participated in the reading of laboratory data with Ms. J. Liu, Ms. A.J. Paradis, and Ms. N. Buu, and established communication with the contact Canadian family member to confirm the structure of the pedigree. I developed the genetic models for the analysis and was responsible for performing the linkage analysis of the study families. I interpreted the results and completed the write-up of the thesis. In May 1994, I participated in fieldwork with Dr. E.A. Fanning and Dr. E. Schurr in several rural Canadian communities

to revisit the Aboriginal Canadian family, discuss the results of the project with the family members, collect additional blood samples for future research, and verify the pedigree.

At the time I began the project, linkage analysis was chosen as the study method because a specific chromosomal region had been identified as the most likely region to contain a human homologue of the mouse *Bcg* gene, but a candidate human "BCG" gene had not been found. This method could be used to examine human families with diseased members for co-segregation of the clinical disease trait and genetic markers in the region of the candidate gene, without knowing if the homologous human gene did exist. The maximum likelihood or lod score method of linkage analysis was used because it has the potential to be much more powerful than non-parametric techniques. Since the lod score method requires the specification of a model of disease inheritance and tuberculosis is a complex, multi-factorial disease, a secondary objective of the thesis was to consider the advantages and limitations of the chosen method of analysis of the data.

This thesis has five sections. In section 1, a summary of the clinical and epidemiological features of tuberculosis is presented, followed by a review of the literature that addresses the role of genetic factors in susceptibility to the disease. The section concludes with a detailed introduction to linkage analysis. At the end of this section is a presentation of the results from two recent linkage studies of human mycobacterial disease. Section 2 defines the objectives of the thesis, while sections 3 and 4 describe the methods and the results, respectively. The final section presents a discussion of the findings and lists the major conclusions of the study.

TABLE OF CONTENTS

List of Tables

Page

List of Figures

List of Appendices

List of Frequently Used Acronyms and Symbols

PPD	purified protein derivative
BCG	bacille Calmette-Guérin vaccine
"BCG"	the putative human homologue of the mouse *Bcg* gene/locus, on chromosome 1 in the mouse and hypothesized to be on chromosome 2q in humans based on conserved linkage
HIV	human immunodeficiency virus
DNA	deoxyribonucleic acid
RFLP	restriction fragment length polymorphism
VNTR	variable number of tandem repeats
cM	centiMorgan
HLA	human leukocyte antigen
Bcg	gene on mouse chromosome 1 that regulates macrophage function in response to infection by *Mycobacterium bovis* and certain other mycobacteria to confer resistance/susceptibility before other immune responses
Lsh	gene on mouse chromosome 1 that controls resistance/susceptibility to *Leishmaniaum donovani*
Ity	gene on mouse chromosome 1 that controls resistance/susceptibility to *Salmonella typhimurium*
Nramp	natural resistance-associated macrophage protein gene/locus on mouse chromosome 1
NRAMP	natural resistance-associated macrophage protein gene/locus on human chromosome 2q
ter	telomere
lod score	logarithm (to base 10) of the odds ratio
$Z(\theta)$	lod score at a specific recombination fraction, θ
$L(\theta)$	likelihood of the data at a specific recombination fraction, θ
LR	likelihood ratio
θ	recombination fraction as a variable in a likelihood or lod score, also the maximum likelihood estimate of the fraction
p, q	allele frequencies
R/R	genotype with two copies of a normal allele (R) at a susceptibility locus, assuming a recessive mode of inheritance of susceptibility

R/r	genotype with one copy of a normal allele (R) and one copy of an abnormal allele (r) at a susceptibility locus, assuming a recessive mode of inheritance of susceptibility
r/r	genotype with two copies of an abnormal allele (r) at a susceptibility locus, assuming a recessive mode of inheritance of susceptibility
K	prevalence of disease in a population
f	penetrance (probability of phenotype given genotype)
Σ	summation operator
Ψ	prior probability operator
$P(x_i)$	probability of a phenotype, x_i
$P(g_i)$	probability of a genotype, g_i
$P(x_i \mid g_i)$	conditional probability of a phenotype given genotype
ELOD	expected lod score, evaluated at the maximum likelihood estimate of the recombination fraction
CRYG1	gamma crystallin locus on human chromosome 2q
CRYGP1	gamma crystallin pseudolocus on human chromosome 2q
DES	desmin locus on human chromosome 2q
FN-1	fibronectin 1 locus on human chromosome 2q
INHB	inhibin locus on human chromosome 2q
TNP-1	transition protein 1 locus on human chromosome 2q
VIL	villin locus on human chromosome 2q
PPV	positive predictive value (probability affected | affected diagnosis)
NPV	negative predictive value (probability unaffected | unaffected diagnosis)

SECTION 1: LITERATURE REVIEW

1.1 Tuberculosis

1.1.1 Etiology and clinical features

Tuberculosis is a communicable disease resulting from infection with the aerobic bacillus *Mycobacterium tuberculosis* or, less commonly, with *Mycobacterium bovis* present in non-pasteurized milk of diseased cattle. *M. tuberculosis* is usually transmitted through air exhaled from a diseased person to others in close proximity. The lungs are the usual site of the primary infection but the bacilli can be spread in the circulatory or lymphatic systems to any tissue or organ.

Following inhalation of *M. tuberculosis* into the alveolar regions of the lungs, the bacilli are ingested by alveolar macrophages. The pathogenicity of the mycobacteria depends on their ability to survive and multiply within the host; the immunocompetent host, in turn, defends itself by restricting bacterial growth and creating a low oxygen environment in the infected region (Grosset, 1993). Macrophages gain the ability to destroy mycobacteria through activation, either non-specifically as a result of the ingestion of organisms and cells, or specifically through interaction with lymphokines released by T lymphocytes (Dannenberg, 1989; Nardell, 1993). If the alveolar macrophages, having acquired varying levels of microbicidal activity, fail to destroy the bacilli the infecting organisms will continue to multiply until they can no longer be contained intracellularly (Nardell, 1993). Immature macrophages from the circulation are attracted to the area by the freed bacilli, chemotactic stimuli, and cellular debris (Dannenberg, 1989).

Within 3–8 weeks after infection cell-mediated immunity develops, signalled by the accumulation of specifically sensitized T lymphocytes and activated macrophages. Delayed-type hypersensitivity, a form of cell-mediated immunity, arises at the same time and involves cytotoxic T cells (Dannenberg, 1989). As a result of these immunologic processes the lung tissue in the infected region becomes semi-solid and acquires low oxygen tension, a process known as caseous necrosis (Grosset, 1993). In addition, immature or impaired macrophages are destroyed by T cells, facilitating the release of bacilli which can then be killed more efficiently by activated macrophages (Kaufmann, 1993). The development of delayed-type hypersensitivity in immunocompetent hosts is

associated with the formation of a positive skin reaction to the injection of a purified protein derivative (PPD) of *M. tuberculosis* (Dunlap and Briles, 1993). The induration that develops within 48–72 hours after cutaneous injection of PPD on the forearm is used to indicate a past infection with mycobacteria in an individual who has not received an anti-tuberculosis vaccine. Vaccination with an attenuated strain of *M. bovis* (the bacille Calmette-Guérin or BCG vaccine) is thought to sensitize T cells in a manner similar to previous mycobacterial infection. Vaccination also results in a positive skin reaction to PPD for a period of time; the reaction likely diminishes within 10 years for most of those vaccinated after infancy (Menzies and Vissandjee, 1992).

The clinical course of an infection with *M. tuberculosis* depends on the extent of mycobacterial virulence and adequate host response (Dannenberg, 1989). Approximately 5% of infected, immunocompetent individuals develop tuberculous disease in the first two years following infection and another 5% at some point in their lifetime (Styblo, 1991). Those who develop disease soon after infection, presumably as a direct result of an inability to contain the bacilli, can be said to have primary disease. Previously infected individuals who develop disease at a later time following a new infection with *M. tuberculosis* have re-infection disease.

If growtl of bacilli can be controlled within the caseous lesion, mature macrophages accumulate to contain the focus in a granuloma form (Nardell, 1993). Bacilli enclosed within granuloma can remain dormant but viable for decades; the bacilli can retain the ability to reactivate the disease process at a later time if the host becomes immunocompromised, leading to reactivation disease. At the same time, the development of cell-mediated immunity against the bacilli protects the host against re-infection because previously sensitized T cells can rapidly respond to the presence of mycobacteria; this resistance, although long-lasting, wanes with time (Stead, 1989; Nardell, 1993).

If mycobacteria are able to multiply in an uninhibited manner within the primary lesion, the granuloma will break down and the infected region will enlarge (Dannenberg, 1989). In the presence of high levels of bacilli, the delayed-type hypersensitivity response in an immunocompetent host will tend to be detrimental, leading to softening and liquefaction of the caseous focus and the development of cavities in the lung as expanding

necrosis erodes a bronchus and the liquid substance is drained (Nardell, 1993). The process of liquefaction determines to a great extent whether tuberculous disease will ensue, particularly for adults (Moulding, 1988). In liquefied regions, oxygen tension increases greatly and favours bacillary growth. Lung cavities are conducive to the dissemination of bacilli to other regions both within the lung and elsewhere in the body, and can facilitate the discharge of *M. tuberculosis* into the airways. In addition, the intense multiplication of bacilli leads to a large population of organisms. The larger the bacillary population, the more likely it is that mutant, drug resistant mycobacteria will arise; the proportion of resistant mutants in normal strains ranges from 1 in 10^5 to 1 in 10^8, depending on the drug (Grosset, 1993).

The clinical form of tuberculous disease which develops in a susceptible host depends on the extent to which mycobacterial growth is restricted (Comstock and O'Brien, 1991). For all disease outcomes, symptoms can include chronic fatigue, weight loss, and fever. Cavitary disease is often signalled by the presence of blood in the sputum, known as haemoptysis (Ravikrishnan, 1992). Immunocompetent adults are more likely to contain the disease in the lungs than children or those with compromised immunity. In an immunocompetent but previously infected adult, the development of disease often leads to a fairly innocuous, self-limiting illness (Nardell, 1993), associated with chronic weight loss and fatigue.

When the mycobacteria disseminate outside the lung, they tend to settle in well-vascularized regions of the body such as the kidney and meninges, producing extrapulmonary disease (Dunlap and Briles, 1993). Extrapulmonary tuberculosis occurs in about 15% of tuberculosis patients in countries with low prevalence of *M. tuberculosis* infection (Bloom and Murray, 1992). As a result of several factors, including a high prevalence of undernourishment and concomitant disease, tuberculosis in the developing world is often severe and associated with a larger proportion of extrapulmonary forms of the disease and severe wasting (Christie, 1987).

Miliary tuberculosis is characterized by massive spread of bacilli and the formation of many small, granulomatous lesions in the lung and elsewhere (Dunlap and Briles, 1993). The miliary form occurs most frequently in young children with primary disease,

but can also be seen in reactivation cases (Moulding, 1988). Child patients under 4 years also have a high incidence of extrapulmonary disease, particularly in the bones and joints or meninges (Jacobs and Starke, 1993), presumably due to immature immune systems (Collins, 1993). Cavitary disease is rare in child cases so that bronchial secretions are usually low in numbers of bacilli (Moulding, 1988). Children tend to be less infectious than adult patients as they also produce limited sputum (Christie, 1987).

In an immunocompromised host, such as a person with acquired immunodeficiency syndrome (AIDS), there is a deficiency in the cell-mediated immune response so that the skin test can remain negative, caseous necrosis does not develop, and bacilli tend to multiply uninhibited by macrophages. Early progression or reactivation of an infection with *M. tuberculosis* is thus much more likely and the clinical manifestations of the disease will often be extrapulmonary or miliary. In the United States, the risk of tuberculosis for a skin test positive individual infected with human immunodeficiency virus (HIV) has been estimated at about 8% per year (Bloom and Murray, 1992).

Tuberculous disease is diagnosed on the basis of clinical signs, the chest X-ray, and analysis of body secretions for evidence of *M. tuberculosis*. Certain pulmonary changes observable by X-ray suggest primary disease, including infiltrates in the middle or lower regions where bacilli tend to be inhaled (Dunlap and Briles, 1993), enlargement of lymph nodes, and effusion into the pleural space (Glassroth, 1993). A case of reactivated disease can be signalled by fibronodular changes in the well-oxygenated regions of the upper lobes and cavitation (Glassroth, 1993). Sputum and respiratory or gastric secretions are used to develop a stained smear to detect acid-fast bacilli under the microscope. Culturing of the specimen remains the diagnostic "gold standard" and is considerably more sensitive and specific than staining, but conventional techniques require weeks for the generation of a positive *M. tuberculosis* culture, because of the slow generation time of the mycobacteria (Glassroth, 1993). The most infectious cases will be those with large numbers of bacilli in their sputum and will be both culture and smear positive for the mycobacteria (Rieder et al., 1989a).

The skin test can be a useful tool to detect new infection in an individual at risk who has not been vaccinated with BCG. The false positive rate of the procedure likely

relates to the prevalence of non-tuberculous mycobacterial infections in the population. The cut-off for a positive result can be adjusted accordingly but is often set at ≥ 10 mm of induration for young North Americans who are, in general, more likely to be exposed to other mycobacteria than to *M. tuberculosis* (Snider, 1982). False negative results can arise because young children under 3 years often have little response to the procedure, exposure could have been too recent, the person could be malnourished, immunocompromised, or have acute tuberculosis, the test may have been administered incorrectly, or there may be other unknown reasons (Bass, 1993).

Without treatment, the case fatality for tuberculosis has been estimated between 40 and 60% (Bloom and Murray, 1992). Most cases of tuberculosis are curable with chemotherapy as long as prolonged treatment is given to eliminate all populations of bacilli, at least two drugs are prescribed to guard against the development of drug resistance, and patient compliance is high (O'Brien, 1993). Short-course, multi-drug therapies of 6 months have cure rates of greater than 90% with compliant patients (Bloom and Murray, 1992). The culture positive patient is usually rendered non-infectious fairly quickly, with conversion of the sputum culture to negative within 3 months if bacilli are not initially drug-resistant (O'Brien, 1993). A treatment regimen of 4 months duration is possible for persons with pulmonary tuberculosis that have negative sputum smear and culture results before treatment is initiated (O'Brien, 1993).

1.1.2 Risk factors

The development of tuberculosis involves two distinct stages, specifically, the acquisition of mycobacterial infection and the progression to significant tuberculous illness (Comstock, 1975). The role of genetic factors in the risk of developing tuberculosis is discussed in section 1.2. Other risk factors for tuberculosis are grouped below according to whether they affect the development of tuberculous infection or disease.

The probability of developing a tuberculous infection depends on the risk of being exposed to *M. tuberculosis* in the air or, in some parts of the world, to *M. bovis* in non-pasteurized milk. In particular, risk factors for inhalation of bacilli relate to the closeness and frequency of contact with an infectious person and the degree of infectiousness of the

source case (Comstock and Cauthen, 1993). Infection with *M. tuberculosis* will therefore be more likely in regions with high prevalence of infectious cases, in overcrowded, poorly-ventilated living conditions, and with frequent, close exposure to sputum smear-positive persons. Approximately 25% of those exposed to infectious persons become infected themselves (Dowling, 1991), although this risk can be higher depending on duration of exposure and closeness of contact.

The risk of developing disease following infection relates to the immunocompetence of the host and the virulence of the mycobacteria. The bacilli may possess factors that confer an ability to escape the host's defenses (Collins, 1993); greater numbers of bacilli, either initially or as a result of increased virulence, can trigger harmful delayed-type hypersensitivity responses in the host, facilitating tissue destruction (Dannenberg, 1989). Risk factors which tend to lower the immunocompetence of the infected host or weaken lung tissue include the following: malnutrition, physical stress, immunosuppressive chemotherapy, and concomitant illness such as diabetes mellitus, auto-immune disease, silicosis, lymphoma, renal failure, and parasitic, bacterial, or viral infections (Stead and Dutt, 1988; Schweinle, 1990; Collins, 1993). The higher risk of tuberculosis among immigrant populations may relate, in part, to the effects of the stresses of the immigration process on latent infection and host defenses (Christie, 1987).

When there is a high risk of tuberculous infection, an age variation is seen in risk of disease such that there is a peak in disease in infants and young children, followed by a second peak in young adults and a final peak in elderly persons (Comstock and O'Brien 1991), predominantly in males. The increased risk for older persons is also seen when risk of infection in general is low (Comstock and O'Brien, 1991), and likely relates to a higher risk of previous infection in this age group and the opportunistic effects of decreased immunocompetence on primary, reinfection, or reactivation disease (Gardner, 1980). While the cause of the second, young adult peak is not known, the peak in young children has been suggested to correspond more to the recency of infection rather than to immaturity of the host response to infection (Comstock, 1975). The risk of disease is greatest in the first two years following infection (Sutherland, 1976; Dowling, 1991).

1.1.3 Epidemiology

Over the last century, the tuberculosis incidence and death rate have dramatically decreased in the developed world. The decline of the tuberculosis death rate in England and Wales began in the 1850s and proceeded at about 1% a year until the discovery of the tuberculous bacilli in 1882 (Ayvazian, 1993). The death rate from tuberculosis was 180/100,000 population in Canada and 188/100,000 population in the United States in 1901 and 1904, respectively (Thorpe, 1989; Bates and Stead, 1993). By 1920 these rates were 87/100,000 in Canada and 100/100,000 in United States, and mortality further decreased to 2.5/100,000 in Canada and 4/100,000 in the United States by 1969 (Thorpe, 1989; Statistics Canada, 1988; Bates and Stead, 1993). Between 1953 and 1984 the annual number of reported tuberculosis cases in the United States decreased by 74%, from 84,304 to 22,255 cases (Rieder et al., 1989a). The decrease in tuberculosis incidence and death rate before World War II has been attributed to general improvement in living conditions and hygiene, except during the Depression of the 1930s, since BCG vaccination, chemotherapy, and mass X-ray surveys were not commonly undertaken at the time (Styblo, 1980; Styblo, 1989). After 1950, there was an acceleration in the decline of tuberculosis incidence, the annual risk of tuberculous infection, and particularly mortality, which has been ascribed to the development and use of effective anti-tuberculosis treatment (Styblo, 1989).

It has been suggested by others, in particular by Rubin (1993), that the major cause of the decline in tuberculosis mortality in the last century was the death of susceptible persons and the evolution of a more resistant population. To support his view, Rubin (1993) refers to similar decreases in death rates from other infectious diseases and low mortality among age groups with high tuberculous infection rates. For instance, although 95% of the adult population over 45 years in the United States was skin test positive in 1940, the tuberculosis death rate was already low in that group (Sagan, 1987) A hypothetical model of the "epidemic wave" nature of tuberculosis supports Rubin's theory and is described below.

Tuberculous disease has the capacity to become epidemic when previously unexposed persons are crowded together; under the epidemic wave model, the mortality,

incidence, and infection rates of tuberculosis follow waves, increasing sharply, maximizing, and then subsiding gradually (Stead and Dutt, 1988). The gradual decrease relates to the elimination of susceptible persons and infectious sources and the entire wave is thought to last about 300 years, with peaks for the mortality, morbidity, and infection rates at 50, 100, and 200 years, respectively. At the beginning of the epidemic wave it is mostly young adults who succumb to the disease, whereas near the end the disease afflicts older persons (Ayvazian, 1993). Proponents of the model suggest that the natural decline of the infection, incidence, and death rates is essentially unaffected by human intervention such as case isolation and BCG vaccination (Ayvazian, 1993; Bates and Stead, 1993).

The epidemic wave model explains the current low levels of tuberculosis morbidity and mortality in industrialized nations and predicts continued increases in such rates for populations in Asia and Africa. It seems likely that the factors which contributed to the improvement of the tuberculosis situation in the developed world are numerous and interactive, and the following scenario may apply: the sources of infection decreased with the death of susceptible persons and the treatment of cases, exposure level lessened and ability to contain infection increased as living conditions improved, and the host response to infection and disease was enhanced by the development of intrinsic resistance and extrinsic intervention with drugs.

Since 1980, the tuberculosis incidence rate in Canada has decreased from 11.5/100,000 population to 7.5/100,000 in 1990 (Statistics Canada, 1992). The rate did increase slightly in 1989, for the first time in 30 years, with a 4.5% elevation in number of cases (rate, 7.8/100,000) (Bueckert, 1992; Statistics Canada, 1992). The Canadian tuberculosis death rate, which was 47.2/100,000 population in 1945, declined to 0.8/100,000 in 1980 and has remained fairly stable since then, reaching 0.6/100,000 in 1990 (Statistics Canada, 1986; Statistics Canada, 1992). The recent morbidity and mortality rates represent a decrease to an average of about 2,250 cases and 185 deaths per year in the decade 1980–1990 (Statistics Canada, 1992). Tuberculosis cases notified in 1990 occurred more often in males (accounting for 55%); across both genders, the highest incidence rates occurred in those greater than 55 years old, particularly among those older

than 75 years, with the next highest rates in the age groups 25–34 and then 0–4 years (Statistics Canada, 1992).

The improvement of the tuberculosis situation has not been uniform across the Canadian population and the case and death load has continued to be focused in high risk groups such as Aboriginal peoples, immigrants from countries with high tuberculosis prevalence, the urban poor, and the elderly (Gaudette and Ellis, 1993) A study of tuberculosis incidence by census division from 1984–1988 in Quebec and from 1985–1987 in the rest of Canada found that the 13 regions with rates higher than 20 cases per 100,000 population (ranging from 20.9–217/100,000) were concentrated in northern Canada. In 11 of these regions, at least 62% of the cases were known to have affected Aboriginal peoples (Gaudette and Ellis, 1993). Cases occurred most often among Aboriginal or urban immigrant persons in regions with incidence rates of 10–19/100,000 population (Gaudette and Ellis, 1993). The findings for Aboriginal Canadians are consistent with those of Enarson and Grzybowski (1986) who reported that the average annual rates of tuberculosis among registered Indians and Inuit between 1970 and 1981 were 16 and 24 times higher, respectively, when compared to other people born in Canada, predominantly of European descent. In 1990, about 20% of the new tuberculosis cases occurred in Aboriginal persons (Health and Welfare Canada, 1993). From 1985–1987, the incidence of tuberculosis in males and females over 65 years of age was 3–4 and 2–3 times higher, respectively, than that in younger age groups; in the same time period, the immigrant population had an incidence rate more than five-fold greater than that in non-Aboriginals born in Canada (Statistics Canada, 1989).

In the United States, a pronounced and sustained increase in incidence of tuberculosis has occurred since 1985. The downward trend of incidence from 1953–1984 had been expected to continue; instead, an excess of 39,000 cases was found to have arisen between 1985 and 1991, when actual case notification rate was compared to an extrapolation of the previous trend in incidence (Bloom and Murray, 1992; Kent, 1993). The number of tuberculosis cases in 1991 (26,283 cases, yielding a rate of 10.4/100,000 population) represents a national increase of 18% since 1985 (Bloom and Murray, 1992). Notifications in New York City account for about 15% of the caseload in the United

States, with a incidence rate of 50.2/100,000 in 1991, which is nearly five times the national average and represents a 143% increase in cases since 1980 (Goldsmith, 1993; Barnes et al., 1993). In 1990, the case rate reported in central Harlem was more than 233/100,000 (Hamburg, 1992).

The elevation of tuberculosis incidence in the United States is likely due to increases in the number of immigrants, substance abusers, homeless persons, and particularly HIV-infected individuals (Snider, 1993; Kent, 1993; Centers for Disease Control, 1991), as well as diminished attention to control of the disease in medical and political establishments (Bloom and Murray, 1992). Tuberculosis is now recognized as the most common opportunistic infection of HIV-infected patients and is often the earliest clinical manifestation of HIV infection, prior to the development of AIDS (Barnes et al., 1993). Susceptibility to primary, reinfection, and reactivation tuberculosis is increased by co-infection with HIV (FitzGerald et al., 1991).

In addition to the increase in tuberculosis incidence, the proportion of drug-resistant cases has risen in the United States since 1986, after more than 30 years of stable or decreasing proportions (Frieden et al., 1993). The rise is likely a result of inadequate treatment of the disease in the past and immigration from regions with high levels of drug-resistant *M. tuberculosis* (Kent, 1993). A study of culture-positive tuberculosis patients in New York City in April 1991 found that 33% were resistant to at least one anti-tuberculosis drug (Frieden et al., 1993). In a national survey of cases in the first quarter of 1991, drug resistance to one or more drugs was identified in 13.4% of tuberculosis patients not known to have received treatment in the past (Kent, 1993). Recent outbreaks of multi-drug resistant disease in New York and Florida resulted in over 200 cases with a fatality rate of 80%; some of the *M. tuberculosis* isolates were resistant to seven anti-tuberculosis drugs, and 96% of the patients were known to be HIV-infected (Barnes et al., 1993). While HIV-infected individuals do not seem to be particularly predisposed to the drug-resistant forms of tuberculosis, the spread of both drug-resistant and drug-sensitive disease is facilitated by the increased tuberculosis susceptibility of HIV-positive persons (Barnes et al., 1993). Although the situation appears to be less grave in Canada at present, multi-drug resistant cases will likely continue to arise in the

United States and elsewhere if disease treatment is inadequate or tuberculosis incidence increases (Kent, 1993; Barnes et al., 1993).

Tuberculosis remains a leading cause of morbidity and mortality in developing nations, and the highest incidence and death rates in the world occur in sub-Saharan Africa (Rodrigues and Smith, 1990; De Cock et al., 1992). In a review of the global tuberculosis situation, the World Health Organization (WHO) estimated that 8 million persons developed tuberculosis and 2.6–2.9 million died of the disease in 1990 (Sudre et al., 1992). About one third of the world's population is believed to be infected with *M. tuberculosis* (Sudre et al., 1992). Table 1 presents the prevalence of tuberculous infection and estimated case and death rates in 1990 for the WHO regions, as reported by Sudre et al. (1992). At least 19% of the population in each region is infected, with the highest prevalence in the Western Pacific region (44%); in terms of numbers of persons, most infected individuals reside in South-East Asia (25%), China (22%), and in Europe and the USA, Canada, Japan, Australia and New Zealand (22%). The majority of cases predicted for 1990 occurred in South-East Asia (about 30%), China (27%), and Africa (18%). The average case detection ratio[1] was estimated as 46% worldwide, with highest levels in the Western Pacific (average, 88%; range, 61–100%) and Eastern Mediterranean (average, 70%; range, 37–100%) and lowest levels in Africa (average, 24%; range, 16–32%). In terms of rates, both the expected incidence and mortality rates were highest in Africa (265/100,000 and 124/100,000, respectively); the greatest number of deaths from tuberculosis occurred in South-East Asia (about 32%) and China (27%).

The current pandemic of HIV infection is expected to increase the magnitude of tuberculosis morbidity and mortality worldwide (Sudre et al., 1992). In early 1992, WHO estimated that at least 9–11 million adults and 1 million children were HIV-infected globally (Barnes et al., 1993). A total of about 3 million persons between 15–49 years were estimated to be dually infected with HIV and *M. tuberculosis* in the study of WHO regions by Sudre et al. (1992). The majority of dually infected individuals will occur in developing countries where most persons with tuberculous infection are less than 50 years

---

[1]The average case detection ratio for each region was calculated as the ratio of the average notification rate in 1980–1989 to the incidence rate predicted for 1990.

Table 1. World tuberculosis statistics (adapted from Sudre et al., 1992)

WHO region	Tuberculous infection, 1990[e]	Tuberculosis cases expected in 1990[f]	Tuberculosis deaths expected in 1990[g]
	Prevalence, % (percentage of total)	Cases per 100,000 population (percentage of total)	Deaths per 100,000 population (percentage of total)
Africa[a]	33.8 (9.9)	265 (17.5)	124 (22.5)
Americas[b]	25.9 (6.8)	127 (7)	49 (7.5)
Eastern Mediterranean[a]	19.4 (3.1)	155 (7.4)	43 (5.5)
South-East Asia[a]	34.3 (24.7)	195 (36.4)	30 (36)
Western Pacific[c]	43.8 (11.3)		
China	33.7 (22.0)	191 (26.6)	72 (27)
Europe[a] and others[d]	31.6 (22.2)	6.3 (5.1)	3.3 (1.5)
All regions	32.8 (100)	152 (100)	55 (100)

Notes.

[a] Includes all countries in the WHO region

[b] Includes all countries in the WHO region, except USA and Canada

[c] Includes all countries in the WHO region, except China, Japan, Australia, and New Zealand

[d] Includes USA, Canada, Japan, Australia, and New Zealand

[e] Calculated using a model which included the annual risk of infection from tuberculin survey data (age-specific prevalences, available since 1975), the past rate of change of infection risk, and the age distribution of the population

[f] Calculated from the annual risk of infection (1% of risk represented 39–59 cases of smear-positive pulmonary tuberculosis per 100,000 population), assuming 1.22 cases of smear-negative and extra-pulmonary tuberculosis per case of smear-positive disease. The annual risks of infection were 1.5–2.5% in the African countries, 0.5–1.5% in the American and Eastern Mediterranean countries, and 1–2.25% in the South-East Asian and Western Pacific countries. An additional number of HIV-related tuberculosis cases was estimated by applying the prevalence of tuberculous infection in the 15–49 year age group to populations thought to be HIV-infected in this age group, and assuming a 10% risk of progression to tuberculous disease per year for dually infected persons.

[g] Calculated assuming a fatality of 50% for the untreated cases predicted in 1990 and 15% for treated smear-positive or smear-negative cases in developing countries. The proportion of patients started on treatment was obtained from an estimate of the average case detection ratio, which was the ratio of the average notification rate in 1980–89 to the incidence rate predicted for 1990. In addition, it was assumed that the case fatality of HIV-infected tuberculosis cases was 50%, regardless of treatment.

old, because the prevalence of HIV infection is highest in the 15–49 years age group (Sudre et al., 1992). A study in sub-Saharan Africa found that 20–67% of tuberculosis patients were infected with HIV (De Cock et al., 1992). In contrast, the prevalence of tuberculous infection is very low in the age group at high risk for HIV infection in industrialized nations, but is high in older persons who reflect the high risk of tuberculosis in the past (Sudre et al., 1992). A survey of tuberculosis clinics in the United States found a median HIV seroprevalence of 3.4% (range, 0–46%; Barnes et al., 1993). On a worldwide basis, Sudre et al. (1992) estimated that 305,000 tuberculosis cases and 151,000 tuberculosis deaths occurred in persons infected with HIV in 1990.

1.1.4 Prevention and control

Tuberculosis prevention and control should have high priority worldwide. Socio-economic development would likely have the greatest effect on reducing occurrence of the disease, but this represents a tentative, at best, long-term solution in many regions (Rodrigues and Smith, 1990). There is currently an urgent need for the integration of effective tuberculosis control programs into global health care systems, the improvement and rigorous application of existing methods of control, and the development of new strategies of prevention, detection, and treatment of the disease.

The most effective method to decrease tuberculosis transmission is the reduction of the sources of tuberculous infection, through identification and successful treatment of cases (Centers for Disease Control, 1990; Rodrigues and Smith, 1990). Chemo-prophylaxis, the protection of healthy, skin-test positive persons against disease progression with anti-tuberculosis treatment, has risk- and cost-benefit for high risk groups (e.g., HIV-infected persons and household contacts of infectious cases) and for low risk groups under 35 years of age (Geiter, 1993; FitzGerald and Gafni, 1990). Vaccination with BCG has demonstrated variable effectiveness: in 10 randomized control trials of BCG vaccines the protection ranged from 0–80% in different human populations (Rodrigues and Smith, 1990). As a preventive measure, BCG vaccination is indicated for young children in specific regions but the impact of vaccination on transmission rates of tuberculosis is questionable (Styblo, 1989; Rodrigues and Smith, 1990).

Research and development is needed in a variety of areas in order to prevent and control tuberculosis more effectively. In particular, there is a need for shorter treatment regimens that guard against the development of drug resistance and have minimal side effects, as well as new drugs to target *M. tuberculosis* strains that are resistant to the traditional medicines (Bloom, 1993; Centers for Disease Control, 1990). As non-compliance with treatment is a major cause of increased disease transmission and the development of drug-resistant strains, it is important to establish treatment programs that optimize compliance; such programs may involve directly observed therapy and legal means to ensure completion of treatment (Hamburg, 1992). To decrease spread of the disease, institutional settings must ensure proper ventilation methods (Barnes et al., 1993), and immigrant populations must receive effective screening and follow-up (Gaudette and Ellis, 1993). The development of rapid, more sensitive methods of diagnosis and drug susceptibility testing of isolates is crucial (Bloom, 1993; Barnes et al., 1993; Centers for Disease Control, 1990). Special attention will have to be paid to the management of HIV-positive persons, with early identification of tuberculous infection and disease in this group (Styblo, 1989); chemoprophylaxis of dually infected individuals (Barnes et al., 1993); and clarification of the risk associated with BCG vaccination for dissemination of *M. bovis* in HIV-infected persons (De Cock et al., 1993). Central to improvements in tuberculosis control is increased commitment of resources. intensified public health efforts in awareness and follow-up of cases, and the adaptation of programs and technology to the developing world (Bloom and Murray, 1992; Barnes et al., 1993; De Cock et al., 1993).

The global tuberculosis situation is challenging our understanding of the etiology, risk factors, and epidemiology of the disease. Knowledge of the mechanisms of tuberculosis pathogenesis is particularly lacking (Bloom, 1993). Research is needed to gain a better understanding of the molecular basis of mycobacterial virulence, intracellular survival, drug resistance, and interaction with HIV, as well as the host response to infection and the development of protective immunity (Bloom, 1993; Bloom and Murray, 1992).

1.2 Genetic epidemiology of tuberculosis

1.2.1 Definitions

It is clear that a better understanding of the factors that increase the risk of disease following infection is needed to develop appropriate treatment and control strategies for tuberculosis. The causes of disease can be environmental, genetic, or a combination of both (Borch-Johnsen and Sorensen, 1993), related to the "causative agent," the external "environment," and "host factors," including inherited components (Neel and Schull, 1954). Genetic epidemiology is defined as the study of "the inherited causes of disease in populations and the etiology, distribution, and control of disease in groups of relatives" (Morton, 1982). Genetic epidemiologists investigate if and how diseases cluster in families, and whether genetic factors contribute to the clustering when environmental or cultural inheritance is taken into account (King et al., 1984). At a time when medical researchers are studying genetic factors to explain variation in predisposition to many diseases, both communicable and non-infectious, a synthesis of genetics and epidemiology can contribute to the understanding of illnesses of multi-factorial etiology (Morton, 1982). In fact, it has been suggested by Baird (1990) and more recently by Schull (1993) that to meet today's public health challenge of a more thorough understanding of disease causation, it will be crucial to study the interaction between the external environment and genetic make-up. Recent advances in molecular biological techniques and statistical analysis offer new avenues through which genetic contribution to disease can be explored.

The human genome is comprised of the diploid set of 46 chromosomes, of which 44 are autosomal and two are sex chromosomes. The chromosomes vary in size and consist of a short arm (labelled p), a long arm (labelled q), and a narrow region in between the arms, called the centromere. The total length of inherited material (deoxyribonucleic acid, DNA) in the genome is about three billion base pairs, each base pair consisting of two complementary nucleotide units on the two opposite DNA strands. Genes are sequences of nucleotides and contain the information required for assembling biological products such as proteins, as well as regions without coding function. Genes vary widely in terms of their proportion of coding sequences and their length, ranging from less than one thousand to a million nucleotides. A large proportion of the genome

is thought to consist of non-coding regions, such as sequences involved in the control of gene expression.

A locus is the specific location of a gene or non-coding DNA sequence on a particular chromosome. A difference in the DNA sequence between individuals at a given locus is called a polymorphism. Heritable sequence variation can arise as a result of alteration of DNA sequences through infrequent change, called mutation, in the inherited material. If the inherited difference can be reliably detected in a laboratory, then the variable characteristic can be used as a genetic marker. Because each chromosome has a homologous partner inherited from the other parent, the characteristics detected for a particular marker can differ on the two chromosomes of a pair. The alternate forms of a marker are alleles; when the two alleles are the same state the individual is homozygous for the marker and when they are different, the person is heterozygous. The alleles present at a locus constitute an individual's genotype and the expressed outcome of the genotype is the person's phenotype for the locus.

Polymorphisms can be detected as variations in the length of a DNA segment at a specific locus. These markers can be generated by sequence variation that (1) creates or destroys a restriction enzyme recognition site resulting in a restriction fragment length polymorphism (RFLP) or, (2) alters the number of times short, simple sequences are repeated within the segment, forming a microsatellite. Most of the microsatellite markers used in this thesis are dinucleotide-repeat markers, for which the repeated unit is two nucleotides in length. Variable number of tandem repeat (VNTR) markers generally have a repeated unique sequence of more than four nucleotides.

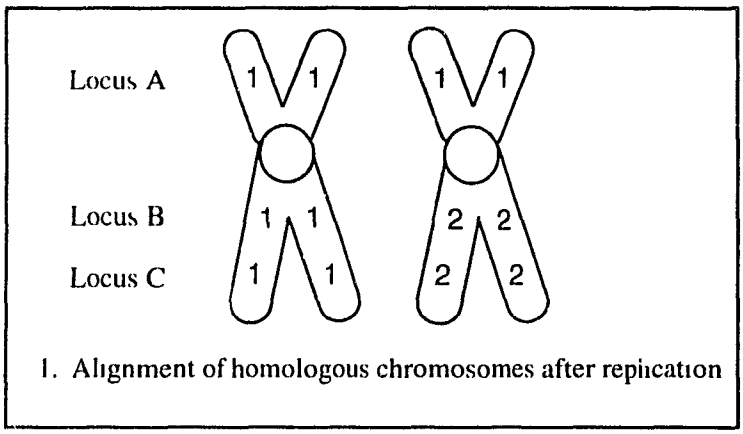
A genetic map is an ordering of DNA markers along a chromosome at specified genetic distances, measured in terms of centiMorgans (cM). In comparison, a physical map uses the base pair as the unit of distance; a genetic distance of 1 cM roughly corresponds to a physical distance along the chromosome of 1 megabase or 1 million base pairs. A human genetic map can be constructed with linkage analysis, in which the co-inheritance of marker alleles from parents to children is analyzed. A disease gene can be located on the map by following the co-segregation of a DNA marker and disease

status (phenotype) in a family with several affected members, called a multiplex family, without needing to know the identity or function of the gene involved.

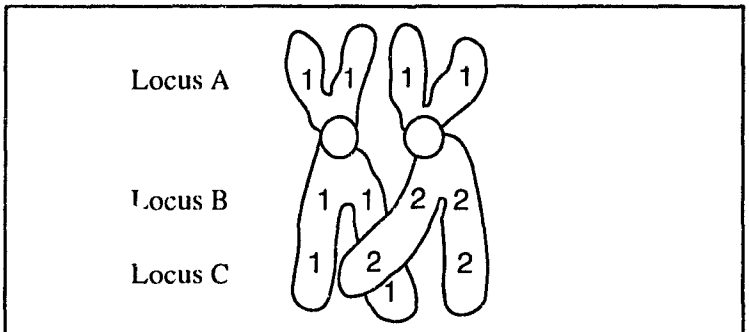
If two loci on a chromosome are close together, then their alleles will tend to be inherited together. During the production of gametes, a method of cell division called meiosis occurs, during which the homologous chromosomes duplicate and are distributed to form haploid gametes (see Figure 1). Before meiosis, the DNA of every chromosome replicates so that each consists of a pair of chromatids, held together at the centromere. During meiosis the homologous chromosomes are closely aligned in the nucleus, at which time a physical reassortment of genetic material can occur, called crossing-over. A breaking and rejoining of the DNA strands on two homologous chromosomes can result in recombination of two loci, so that alleles originally on the same chromosome (but at different loci) can be separated.

The probability of recombination increases with distance between two loci so that markers which are close together will have a lower frequency of recombination than distant markers. Very closely linked alleles that tend to be inherited together form a haplotype. The recombination frequency or fraction of two loci is an estimate of the genetic distance between them. A recombination frequency of 1% (or a fraction of 0.01) is defined as equal to 1 cM in genetic map distance. This correspondence can only be assumed for closely linked markers with a recombination frequency of less than 10%; for larger distances, recombination fractions between adjacent pairs of loci cannot be summed, and a map function is required to translate the recombination fraction to additive genetic distance.

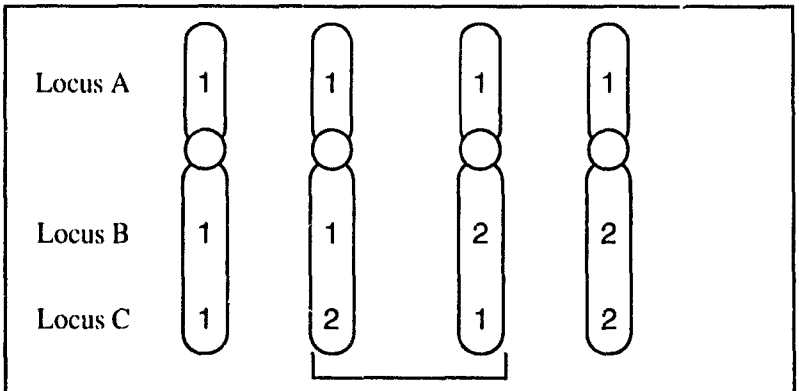
In this thesis, the term susceptibility gene refers to a gene which has two alleles, one normal allele and one abnormal, susceptibility allele which is neither necessary nor sufficient for disease, but confers an increased risk of disease. Depending on the mode of inheritance of the alleles at the susceptibility locus, the expression of particular genotypes will yield the phenotype of genetic susceptibility to the disease. The term disease gene, in comparison, refers to a gene with an abnormal allele which is necessary and sufficient for disease.



1. Alignment of homologous chromosomes after replication



2. Crossing-over between locus B and C, involving one chromatid from each homologous chromosome



3. Two of the chromatids produced are recombinant, displaying the rearrangement of the alleles at locus C, and two are the parental types

Figure 1. The process of crossing-over during meiosis (adapted from Connor and Ferguson-Smith, 1987)

The putative susceptibility gene considered in this study is hypothesized to be present at an autosomal locus on chromosome 2q. When a disease or susceptibility trait is autosomal recessive, an individual must be homozygous for the abnormal allele in order to express the trait, whereas an autosomal dominant trait can be expressed if one or two abnormal alleles are inherited. The genetic markers used in the linkage analysis in this thesis, which are known to be present on chromosome 2, display autosomal co-dominant inheritance, whereby the expression of both alleles at the locus can be distinguished. For such markers, the observed phenotypes arising from genetic analysis indicate the genotypes at these loci, in the absence of laboratory error, null alleles, or mutation in their DNA sequence (Weber and Wong, 1993).

1.2.2 Population studies

Various lines of evidence can be used to provide support for a role of genetic factors in susceptibility as opposed to entirely environmental or chance events. The evidence includes a higher incidence or prevalence of the disease among relatives of cases than among relatives of controls or among members of the general population, as well as the development of the disease over time in a larger proportion of a cohort with affected relatives than in a comparable group without a family history of the disease (King et al., 1984). Such studies for tuberculosis susceptibility do not seem to exist, although the disease has been suggested for centuries to cluster in families. Further indirect evidence for genetic susceptibility to mycobacterial disease includes an epidemic of leprosy in North America in two regions settled by Acadians that share French ancestry (New Brunswick and Louisiana), as well as the reported familial clustering of the disease within the French Acadian ethnic group (Neel and Schull, 1954). The interpretation of such observations is problematic because of the infectious nature of tuberculosis and leprosy; exposure to the causative agent as well as genetic similarity will be increased among relatives, and disease clustering in families or communities may occur in the presence of uniform susceptibility due to increased exposure opportunities in a household or geographic region.

Allele frequencies are influenced by many genetic, ecological, and evolutionary factors which interact with each other, often in a complex manner (Hartl and Clark,

1989). The population genetic factors include mutation, patterns of mating, population size, migration, geographic distribution of individuals, and natural selection (Hartl and Clark, 1989). By selecting for or against particular phenotypes, natural selection modifies the genotype frequencies in a population so that the allele frequency distribution gradually changes. Individuals that possess phenotypes which confer greater biological fitness and environmental adaptation will tend to survive and reproduce, thus affecting the distribution of genotypes in the next generation. For example, the sickle cell allele has been retained in some equatorial African populations, despite the low biological fitness of the homozygous individual, because of a survival advantage possessed by the heterozygous carrier of the trait for increased resistance against *Plasmodium falciparum* malaria (Stead, 1992). If susceptibility to a fatal infectious disease that claims many victims before adulthood is influenced by genetic factors, then the force of natural selection, operable over many generations, may favour survival of those with a resistant genotype. Study of the pattern of tuberculosis incidence and severity in populations over many generations may produce evidence in favour of selection for resistant alleles.

Archaeological evidence has indicated the presence of tuberculosis among prehistoric peoples in both the Old and the New World (Merbs, 1992). However, it was only within densely-populated and impoverished regions such as the industrial centres of Europe and the crowded cities of Post-Columbus America that a high, epidemic level of *M. tuberculosis* infection and case rate was maintained. The disease became more severe in terms of case number and fatality with the increased contact of persons with movement across continents, followed by increased density of populations and the adoption of more sedentary lifestyle. When infection rates of tuberculosis reached epidemic level. the death rate was very high, particularly among children. In the early nineteenth century in Europe, tuberculosis death rate in London reached 7/1,000 population (Ferguson, 1955) and it has been suggested that the disease accounted for a third of all deaths in Paris (Bloom and Murray, 1992). Acute epidemics of tuberculosis occurred among North American Natives clustered on reserves in the late nineteenth and early twentieth centuries; the Native death rate in the Qu'Appelle Valley region of Saskatchewan, for example, was reported as approximately 10/1,000 in 1881 and 90/1,000 in 1886 (Thorpe,

1989). Once Europeans penetrated the interior of Africa in the early twentieth century, Native Africans likewise developed a severe form of tuberculosis that led to a high mortality rate from the disease (Stead, 1992).

Evidence presented by Thorpe (1989) suggested that human populations did not differ with respect to the proportion of persons susceptible to tuberculosis when the disease was first encountered. The early experience with tuberculosis resulted in extremely high case and mortality rates for all peoples, particularly among persons of reproductive age and younger. However, various ethnic groups do differ in the extent of their past experience with epidemic level tuberculous disease. Archaeological and historical sources have been used to indicate how long various populations have been in contact with tuberculosis, and thus how long natural selection may have favoured resistant individuals in each group. The assumption must be made that there was genetic variation in susceptibility to the disease within the populations when exposure began, resulting in different phenotypes which had differential fitness in the face of disease. With this in mind, strong natural selection may have been operating for many centuries for Europeans and Mongolians, for 300–400 years for Africans brought to America, for a shorter period of time for isolated Native Americans, and for only about 80 years for Central Africans (Stead, 1992 and Kushigemachi et al., 1984). Thus, one would expect the frequency of tuberculosis susceptibility alleles to vary among ethnic groups with disparate tuberculosis experiences in the past.

The clinical course of tuberculosis following infection varies in its progression such that some persons are able to combat the infection whereas others develop a primary "typhoidal illness characterized by wasting, lymphadenopathy, serious effusions, and fever" (Stead, 1992). Those persons who develop reactivation disease after many years of contained infection tend to develop a chronic pulmonary form that advances slowly with few symptoms besides gradual weight loss (Stead, 1992). The acute primary stage of disease is often observed in those populations with a shorter history of tuberculosis exposure, such as African-American slaves in America prior to the 1860s, and among Central Africans today. Tuberculosis seen in African-Americans today tends to be the

chronic pulmonary type, which seems to have a faster rate of progression than in Caucasians (Stead, 1992).

In an attempt to find evidence for variation in genetic susceptibility, racially dissimilar populations have been compared for differences in risk of tuberculosis infection and disease. It has been difficult in studies of different racial groups to control for potential confounding factors, such as age, infection intensity, and socioeconomic status. Kushigemachi et al. (1984) reviewed reports of differences in case rates in different populations and concluded that these variations could be the result of many other factors besides inherited susceptibility. To make accurate comparisons of this kind, large numbers of individuals in different populations, who were exposed under similar circumstances and are comparable with respect to age and gender, would have to be actively followed for a long period of time (Kushigemachi et al., 1984).

A study by Stead et al. (1990) followed black and white residents of racially-mixed nursing homes who experienced the same residence environment and were tuberculin skin-tested on entrance to the homes. Of those who had a negative skin test at admission, later development of a positive reaction was significantly more frequent among black than among white residents (overall relative risk = 2.1; 95% confidence interval = 2.0–2.2). The association of race with skin test conversion held regardless of the percentage of black residents in the home, gender or age group of the converted residents, and race of the index patient, when a single infectious source case could be identified. No significant racial difference was seen in the proportion of untreated residents either skin test positive at admission, or later, who developed clinically-evident tuberculous disease. It is not clear that the two groups under comparison would have dissimilar enough genetic background with respect to tuberculosis susceptibility for a difference in case rate to be detected. Several important risk factors for skin test conversion, falsely negative skin tests, and disease development were not controlled for (e.g., health and socioeconomic status, use of medications) (Rosenman, 1990; Felton et al., 1990). A racial difference in severity of tuberculous disease was observed, in which blacks who were diagnosed as culture-positive cases were more likely to be smear-positive than culture-positive whites, but this difference was not statistically significant

at the 5% level (relative risk = 1.5; 95% confidence interval = 0.88–2.54). The study, therefore, does not support racial variation in susceptibility to disease following infection.

1.2.3 Twin studies

More direct evidence for inherited risk factors in the development of tuberculous disease has come from twin studies (Fine, 1981; Harvald and Hauge, 1956). Monozygous twins carry the same set of alleles, whereas dizygous twins share on average 50% of their alleles, and are as genetically similar as any pair of siblings. With the assumption that monozygous and dizygous twins share their environment to an equivalent degree (an assumption which may not always hold--see later discussion), a significantly higher percentage of concordance with respect to disease in monozygous than in dizygous twins provides support for genetic factors in the occurrence of the disease. Concordance of less than 100% indicates that a non-genetic component has a role in development of the disease. Penetrance is defined as the probability of the phenotype (e.g., the disease) given the genotype and can be estimated as the percent concordance of disease between monozygous twins, assuming an unbiased selection of twin pairs and certain assumptions.

The population of twins can be obtained by starting with cases, identifying and following up all those who have a twin, or by using a twin registry to find all pairs with at least one diseased twin (Breitner et al., 1993). The success of complete ascertainment for either method depends on how well the twin population is covered by the information source used. It is difficult to avoid biased selection and follow-up of twin pairs, which tend to favour the identification of concordant monozygous twins because their similarity makes them more easily recognizable. The result of such bias can be an over-estimation of the concordance of monozygous as compared to dizygous twins and an inflation of the penetrance estimate (Fine, 1981). The incidence of twin birth and the proportion of twins that are dizygous varies among different ethnic groups (LaBuda et al., 1993). The ratio of monozygous to dizygous twins in the United States is about 1:2, thus an unbiased ascertainment of a twin sample from such a group should be consistent with this ratio (Breitner et al., 1993).

In order to avoid bias, it is crucial in twin studies to (1) correctly identify zygotic status when dealing with like-sex twins by incorporating comparison of physical

resemblance and genetic testing with several polymorphic markers, and (2) determine the clinical status of all twin pairs in a uniform manner for a sufficient period of time to detect all cases of disease (LaBuda et al., 1993). For diseases with an environmental element, such as tuberculosis, it is essential to ensure or assume equal exposure to *M. tuberculosis* since conception for the two types of twins (the "shared environment assumption"; Breitner et al., 1993). As a result, it is best to compare disease concordance between monozygous twins and same-sex dizygous twins; however, there may still be more similar environment for the monozygous twins because of their physical likeness and possible sharing of circulation until birth. An unknown difference in environment can inflate the evidence for hereditary factors (King et al., 1984). Environmental data about the amount of time spent together and the similarity of activities should thus be collected from all twin pairs in order to assess whether the shared environmental assumption is likely to hold. The ideal method to investigate the roles of genetic inheritance and environment is the adoption study in which the incidence of disease is examined for biologically related persons raised in different environments since infancy or early childhood (King et al., 1984).

Twin studies of tuberculosis have consistently shown a disease concordance between monozygous twins on the order of 2–3.5 times higher than that in dizygous twins (Fine, 1981). Using multivariate analysis of British twins ascertained from a population of registered tuberculosis patients (n = 202 pairs), Comstock (1978) found a significantly greater monozygous disease concordance of 31% ($p = 0.05$) compared to 15% in same-sex dizygous twins, after adjustment for seven covariates. The categorical covariates included age at diagnosis (0–14, 15–24, or ≥ 25 years), type of disease (reinfection or other) and sputum results (positive or other) for the index twin, and, for the co-twin, gender, known tuberculosis contact (yes or no), contact with index twin (living together or not), and years from index to co-twin diagnosis (0–9 or ≥ 10 years). Zygotic origin was assigned on the basis of physical likeness and blood group testing. Given the frequency of twins in the general population, 97.5% of the expected proportion of twins were located. However, the ethnicity of the twins was not considered so that immigrant

twins from areas of high tuberculosis incidence, if more likely to be monozygous and concordant, may have increased the variation of concordance between the types of twins.

The most comprehensive tuberculosis twin study published is the New York State hospital and clinic-based investigation of Kallmann and Reisner (1943), which used a "twin family method" since data were also collected on full siblings, half-siblings, parents, and spouses of twin index cases. A state-wide, five-year collection of the index cases, using uniform classification systems for zygotic origin and clinical status of the twin partner, resulted in enrollment of 308 twin pairs, of whom 230 pairs were dizygous. For each pair, the index twin aged 15–64 years was diagnosed with reinfection tuberculosis[2]. In addition, data were analyzed for 226 spouses, 688 parents, 42 half-siblings, and 720 full siblings of twin patients.

The disease concordance rates with respect to a diagnosis of tuberculosis for the various sets of relatives were adjusted for differences in age distribution; the method of correction related incidence to the number of persons within and surviving a period of high tuberculosis risk, assumed in the study to span the ages 15–29 years. The adjusted incidences were, therefore, the risks of developing tuberculosis (assuming exposure to infection) for persons over 29 years with a particular relationship to the index case. The tuberculosis incidence in a general population 15 years and older with similar age and racial distribution as the state of New York was calculated for comparison, assuming that morbidity was 10 times the tuberculosis death rate for whites and five times the death rate for blacks. The mortality rate for blacks was multiplied by a smaller factor because a higher proportion of deaths was known to occur among black tuberculosis patients at the time.

As shown in Table 2, regardless of whether age-corrected or uncorrected incidence was considered, disease concordance correlated with degree of relationship to the index

---

[2]Historically, the term "reinfection tuberculosis" was given to any adult case with upper lobe and cavitary tuberculosis (Moulding, 1988). As much of the study population was likely infected during childhood, the cases probably represented a combination of re-infection and reactivation disease.

Table 2. Tuberculosis concordance in pairs of relatives and the general population

	General Population	Relationship to affected index twin					
		Spouse	Parent	Half sib	Full sib	Di-zygous twin	Mono-zygous twin
					† for like sex ‡ for unlike sex		
Number of relatives studied	--	226	688	42	720 -- --	230 118† 112‡	78
Number of diseased relatives	--	14	114	4	136	42	48
Uncorrected concordance (%)	1.1*	6.2	16.6	9.5	18.9	18.3	61.5
Corrected concordance (%)	1.4	7.1	16.9	11.9	25.5 27.7† 21.7‡	25.6 30.2† 20.5‡	87.3
Corrected concordance (%) with tuberculous parents**	--	--	--	--	35.3	38.6	88.9
Corrected concordance (%) without tuberculous parents***	--	--	--	--	18.1	18.8	85.7

Adapted from Kallmann and Reisner, 1943 and Neel and Schull, 1954.

* Estimate of tuberculosis "morbidity rate" in a general population (reinfection tuberculosis cases per 100 persons over 14 years of age) in which the ratio of white to non-white persons is approximately 5:1. The age group from birth to 14 years of age was not represented by any of the groups of relatives in the study.

** Corrected concordance for pairs in which the index twin has a parent known to be a case

*** Corrected concordance for pairs in which the index twin has parents not known to be cases

case (i.e., proportion of shared genes). The concordance between dizygous twins and full siblings was very similar and was significantly less than that in monozygous twins. The effect of environment was indicated by the elevation of incidence among spouses of index cases as compared to the disease occurrence in a general population. Dizygous and monozygous index twins did not differ significantly in average age at disease onset, average duration of observation, percentage with known exposure to tuberculosis, or percentage with a known family history of the disease. Both dizygous and monozygous discordant twins had reached the average age of 31 years by the time they were examined for the study and both types of discordant twins had been unaffected for about six years since the diagnosis of tuberculosis in the index twin.

When dizygous twins and full siblings with and without affected parents were compared, the disease risk was modified to a similar extent in both fraternal groups (that is, the concordance approximately doubled if parents were affected; Table 2). When monozygous twin pairs with and without diseased parents were compared, however, there was little change in monozygous disease risk, suggesting that the increase in concordance for dizygous twins and full siblings with a family history of tuberculosis may have more to do with the inheritance of genetic factors than increased exposure to infection in the home. Although age-corrected concordance was slightly higher for like-sex full siblings and dizygous twins than for pairs of dissimilar sex, approximately the same level of disease risk was maintained for the siblings versus the twins in each gender category (like or unlike) (Table 2).

The top section of Table 3 displays the uncorrected disease concordance for dizygous and monozygous twin pairs for whom the co-twin was known to have had exposure to tuberculosis (presumably through contact with a sick person). This is the most careful comparison presented in the study for which we can assume that discordance between pairs of twins was likely not due to exposure of only one member of the pair to *M. tuberculosis*. The percent disease concordance between monozygous twins was approximately 2.6 times that for dizygous twins (69% versus 26%; $p < 0.0001$). The rest of Table 3 presents comparisons of clinical disease presentation between twin pairs. Whether complete or less complete similarity in extent of disease was considered,

Table 3. Distribution of concordance and discordance in study twin pairs

	Clinical code, 1st member*	Clinical code, 2nd member*	Dizygous twins	Monozygous twins
Number of known exposed co-twins	--	--	175	52
Number concordant for disease (%)	--	--	46 (26.3)	36 (69.2)
Number of pairs with complete clinical similarity	II III IV	II III IV	2 3 7	3 8 9
Number of pairs with less complete clinical similarity	II III IV	I II III	57 8 6	22 14 13
Number of pairs with less complete clinical dissimilarity	IV III	II I	16 86	1 8
Number of pairs with complete clinical dissimilarity	IV	I	45	0
Total number of all similar pairs	--	--	83	69
Total number of pairs	--	--	230	78
Proportion of similarity in all pairs	--	--	0.361†	0.885†

Adapted from Kallmann and Reisner, 1943.

* The clinical classification codes used were the following:

I No tuberculosis (with or without exposure)
II Mild tuberculosis with subsequent arrest of the disease
III Advanced tuberculosis
IV Fatal tuberculosis

† statistically significant difference in proportion, $p < 0.0001$

monozygous twins were significantly more similar than dizygous twins. Overall similarity (including complete and less complete) was significantly higher in monozygous than in dizygous twins (89% versus 36%; $p < 0.0001$). The ratio of similar clinical course of disease to dissimilar course in monozygous and dizygous twins was 2:1 and 1:8, respectively; in other words, similarity of disease progression was 16 times more likely in the monozygous twins.

The clinical comparisons made in this study support a role for genetic inheritance in risk of tuberculosis if (1) classes I and II are appropriate representations of resistant phenotypes and classes III and IV depict susceptible phenotypes, and (2) other factors that could affect clinical progression of the disease (e.g., concomitant illness or access to treatment) are not differentially distributed among the various groups of twin pairs. A more thorough evaluation of the clinical comparisons could have been made if the investigators had indicated the number of pairs of twins in which one member was not known to have been exposed to tuberculosis. If this differential exposure information was correct, including such pairs in the analysis of similar clinical course would be incorrect, since such pairs are expected to display dissimilarity.

1.2.4 Association studies

Association studies with population instead of family data use a case-control approach to examine the association between genetic markers and clinical phenotype. Patients and controls are tested with a marker, and the difference in frequency of particular alleles in the two groups is statistically tested. Observed association of clinical status with specific alleles can provide statistical evidence that a gene in the region tested may be involved in the development of a disease (MacCluer and Kammerer, 1991). The marker chosen for an association study may be at a candidate locus suspected to influence susceptibility (for example, variation in the apolipoprotein gene could be used in a study of atherosclerosis). Alternatively, various anonymous markers can be chosen at certain intervals in order to span the genome. The most direct (and less common) method of association analysis directly compares the DNA sequence of a candidate gene between cases and controls (Khoury et al., 1990).

The most important design issue in association studies is appropriate selection of controls. To study genetic risk factors, controls must be ethnically similar to the case population in order to avoid spurious association of case status with alleles that are different in frequency in the two groups due to ethnic diversity. This issue can be avoided with the Affected Family-Based Controls (AFBAC) association study in which alleles in parents are placed in a diseased category if they appear in an affected child and in a control category otherwise (Thomson, 1991). In addition to ethnicity, the case and control samples in an association study need to be similar for as many other non-genetic risk factors for the disease as possible to avoid confounding by covariates, such as age, gender, and socioeconomic status (MacCluer and Kammerer, 1991). For a disease with a necessary environmental component, a true association between a marker and clinical status may be diluted if a proportion of persons in the study have not been exposed to the external factor (Khoury et al., 1990).

Uncertainties in clinical diagnosis and unknown disease etiology, as well as the contribution of multiple genes to a disease phenotype, affect the power of detecting a true association but can be overcome somewhat with large sample size (Crowe, 1993). When many unlinked, independent markers are typed, a large number of comparisons, uncorrected for multiple testing, will lead to false positive results (Crowe, 1993). Thus more stringent criteria for statistically significant differences, adjustment of nominal probabilities, large sample sizes, or more sophisticated approaches (if necessary) will have to be used. The extent of the modifications will depend upon the number of unlinked markers studied, the prior probability that a tested locus will be involved in pathogenesis, and the false positive and negative rates that are acceptable (Kidd, 1993). Verification of cause and effect relationships between markers and clinical status requires replication of positive findings and combined analysis from several studies, if the patient and control populations are sufficiently similar across investigations (Crowe, 1993).

Several association studies of tuberculosis have examined the hypothesis that the genes of the human major histocompatibility complex (MHC), which code for products of the human leukocyte antigen (HLA) system, have a role in control of the immune response to mycobacterial infections (Singh et al., 1983). The MHC system consists of

several highly polymorphic genes in close proximity on chromosome 6 which code for proteins involved in regulation of the immune response by T lymphocytes (Trowsdale, 1993). The MHC molecules function as "recognition elements" by presenting processed antigens to specific kinds of T cells in a restricted manner, depending upon whether the molecules belong to class I (coded by HLA-A, B, and C loci) or class II (coded by HLA-DP, DQ, and DR loci) (Sanjeevi et al., 1992; Hill et al., 1991). It has been suggested that the extreme polymorphism of the MHC genes and the unusually even distribution of their alleles in human populations may have been maintained by selective pressure to provide defense against a diverse range of pathogens (Hill et al., 1991). Significant association of protection from severe malaria with a common HLA class I antigen and class II haplotype in West Africans has recently been reported (Hill et al., 1991).

Association studies of tuberculosis and HLA alleles, detected by serological testing, have provided conflicting results (Sanjeevi et al., 1992). Several investigations have shown positive associations with particular class I and II alleles in patients from various populations, while others have not detected any significant associations of HLA with disease. For example, HLA-DR2 and DQw1 alleles were associated with sputum smear-positive pulmonary tuberculosis when cases (n = 101) were compared with healthy controls of similar Indonesian ethnicity (n = 64) (Bothamley et al., 1989). Given the endemic nature of *M. tuberculosis* in the study region, it was likely assumed in the analysis that the controls had been exposed. The positive association was not statistically significant when Sanjeevi et al. (1992) corrected for testing of multiple HLA alleles. A study of 25 multiplex families with pulmonary tuberculosis in Northern India demonstrated sharing of identical HLA haplotypes between diseased siblings that was significantly greater than expected (Singh et al., 1983). HLA-DR2 was preferentially transmitted from both diseased and healthy parents to affected children rather than to healthy children. Exposure of all subjects to infection was quite likely in this family-based study in an endemic tuberculosis region, although several persons tested tuberculin negative. It was suggested by the study authors that the tuberculin insensitivity may have been related to the lower frequency of HLA-DR2 among the non-reactors when compared to the healthy reactors.

A more recent study with smear-positive pulmonary tuberculosis cases (n = 143) and controls (n = 287) in Southern India showed a positive association of HLA-DR2 with disease that remained significant after correction for the number of HLA-DR alleles tested ($p < 0.01$) (Brahmajothi et al., 1991). The association was found in all but one of the ethnic subgroups of the study sample. Interestingly, the same DR2 allele had a significantly negative association with smear-negative patients (corrected $p < 0.05$), suggesting that the allele was correlated only with advanced forms of the disease. These findings, however, were not corroborated in a smaller Southern Indian study of pulmonary tuberculosis patients (n = 38) and healthy controls (n = 36) in which HLA-DR2 was more frequent in the control group (corrected p not significant) (Sanjeevi et al., 1992). In addition, no increased sharing of DR-DQ haplotypes was observed between affected siblings in 12 multiplex families, in contrast to the finding by Singh et al. (1983).

Confirmatory evidence of HLA associations with tuberculosis has yet to be presented. The contradictory findings may be the result of the selection of inappropriate control groups such as unexposed or exceptionally healthy volunteers, unknown classification errors such as false negatives which decrease the power of detecting associations, and other sample and/or methodological differences between studies. The HLA-DR2 allele may not be involved in determining the outcome of infection (that is, disease or no disease), but may instead have a role in the clinical course of the disease or may function in the development of tuberculin sensitivity, in which case an association with disease status could be detected if a sufficient number of controls were not exposed.

An HLA-DR association with leprosy (both tuberculoid and lepromatous forms) has been noted in several populations with case-control and family segregation approaches (Abel and Demenais, 1988). The HLA-DR3 allele has been associated with development of tuberculoid leprosy in population and family studies (Bothamley et al., 1989). On the basis of these investigations, it has been suggested that non-MHC genes confer susceptibility to leprosy *per se* (that is, the genes influence whether disease occurs or not) whereas HLA alleles influence the clinical form of disease which develops (Schurr et al., 1991a). In addition, it has been proposed that immunogenetic involvement in leprosy and tuberculosis operates in the same manner (Brahmajothi et al., 1991).

1.2.5 Segregation analysis

The purpose of a segregation analysis is to use phenotype data from families to determine the mode of inheritance of a disease-predisposing gene (Thomson, 1991). Knowledge of the physiological process by which the susceptibility gene contributes to disease etiology is not required for the analysis (King et al., 1984). Likelihood calculations are performed to determine whether the observed pedigree and clinical data fit particular Mendelian inheritance models, such as autosomal or sex-linked, dominant or recessive. Segregation analysis can test for inheritance of a single gene with a role in disease development even though there may be more than one gene contributing to disease.

Families may be selected for genetic study because they contain many affected persons or their case distribution seems to fit a particular inheritance mode. In a segregation analysis, the method of sample ascertainment is corrected for if necessary by study design, and estimation of genetic and ascertainment parameters is carried out simultaneously (Thomson, 1991). Cases of disease which have arisen in genetically-resistant family members for non-genetic reasons (e.g., as a result of high infection dosage, high strain virulence, or host immunosuppression) will decrease the fit of the data to a genetic model, unless an environmental contribution to disease is specified in the analysis. Computer programs (for example, POINTER and PAP) are currently available which allow specification of more complex models, such as the involvement of multiple genes or the inheritance of a gene whose pathogenic action, when abnormal, increases with age of the susceptible individual (Thomson, 1991; Hasstedt, 1989).

There are no published reports of segregation analyses of tuberculosis. However, a dominant gene for susceptibility to leprosy with 63% penetrance (that is, 63% probability of disease given susceptible genotype) was suggested by an early segregation analysis of a Caucasian population in Louisiana (Belknap and Hayes, 1961). Analysis of 91 Filipino families with at least one offspring case of lepromatous leprosy supported an autosomal recessive locus (Smith, 1979); however, a multi-locus model for susceptibility provided a better fit to the data, which was also the case for a New Guinean data set containing 340 lepromatous and non-lepromatous leprosy patients and their first-degree

relatives (Serjeantson et al., 1979). Several more recent leprosy studies have used complex models which incorporated a major gene effect, an inheritable polygenic component, and a random environmental contribution to disease. Investigations with 72 multiplex pedigrees from India (Haile et al., 1985) and 63 multiplex Thai families (Wagener et al., 1988) suggested recessive inheritance of tuberculoid leprosy.

Abel and Demenais (1988) performed a complex segregation analysis of 27 multi-generational pedigrees of 82 leprosy family members from Desirade Island, a region with high leprosy prevalence and for which homogeneity of environment and exposure to *M. leprae* in the population could be assumed. Three analyses were carried out with the specification of age-dependent penetrance of a major gene and use of three phenotypes: all forms of leprosy, non-lepromatous cases only, and lepromatous cases only. Segregation of an autosomal recessive or semi-dominant major susceptibility gene was supported by the data for the phenotype of any kind of leprosy ($p < 0.02$). The risk for disease development, under the recessive model, reached 62% by the age of 68 years for homozygous, genetically-susceptible persons. The proportion of non-genetic cases among the leprosy patients in the population was estimated to be 23% with the recessive model. There was also evidence for recessive inheritance of a major susceptibility gene when only non-lepromatous leprosy was considered ($p < 0.05$). For both the all leprosy and non-lepromatous leprosy phenotypes, purely environmental models of disease transmission were rejected ($p << 0.001$) and the major gene effects were significant without the addition of multifactorial components. There was not enough data for analysis with only lepromatous leprosy cases. Abel and Demenais suggest, on the basis of this and other segregation and HLA-association studies, that different susceptibility genes exist for leprosy *per se* and non-lepromatous leprosy subtype. In their opinion, the former is likely a non-MHC gene while the latter may be a gene in the HLA region.

1.2.6 Experimental animal models

Based on the human evidence presented in sections 1.2.2–1.2.5 it appears likely that genetic factors contribute to tuberculosis susceptibility. Clearly, the challenge of using a genetic epidemiological approach in the study of tuberculosis susceptibility is finding evidence for the genetic inheritance "signal" out of the "noise" created by the non-

genetic environment within and outside the host, such as nutritional status and mycobacterial exposure level. In order to separate the effects of inherited susceptibility and environmental conditions on the disease outcome, animal models have been used to locate a gene for tuberculosis susceptibility that may exist in humans because of evolutionary homology. The location of a putative human gene can be searched for in homologous regions of chromosomes in other animals.

Early studies by Lurie and Dannenberg (1952; 1965) were successful in breeding rabbit strains with varying degrees of susceptibility to tuberculous pneumonitis following inhalation of *M. tuberculosis*. The resistant trait was inherited in a dominant manner and was associated with a superior ability to sequester and inactivate mycobacteria within alveolar macrophages (Lurie et al., 1952; Lurie and Dannenberg, 1965). In the 1980s, researchers at the McGill Centre for the Study of Host Resistance developed a model of tuberculosis susceptibility in inbred mice (Skamene, 1989; 1991). The murine response to intravenous injection of a small dose of *M. bovis* was found to vary in the following manner: resistant mice showed inhibition of bacterial proliferation in their spleens and livers in the first 3 weeks of infection compared to rapid, uncontrolled growth of *M. bovis* in susceptible mice (Gros et al., 1981). A significant difference in bacterial load could be observed as early as 24 hours after infection, but the largest difference was seen at 3 weeks. The distinction between the two types of mice indicated that a single gene could be controlling the early response to presence of mycobacteria in the reticulo-endothelial organs. The study investigators suggested that the advantage possessed by the resistant mice was related to natural resistance mechanisms and was likely associated with an ability to control mycobacterial growth, rather than eliminate the bacilli, at an early stage. In some susceptible mice the development of a second phase was observed after 3 weeks during which the bacilli were progressively eliminated from splenic tissues, perhaps as a result of the establishment of specific cell-mediated immunity against the mycobacteria (Gros et al., 1981).

Analysis of numerous crosses between inbred mouse strains has confirmed that resistance to early growth of *M. bovis*, *M. leprae*, *M. avium*, *M. intracellulare*, and *M. smegmatis* is under the control of a single gene named *Bcg* (for bacille Calmette-

Guérin), that is autosomal can be expressed as a dominant resistant and recessive susceptibility allele (Schurr et al., 1990a; 1991a). At the early stage of defense against intravenous *M. bovis*, resistance at the *Bcg* locus determines control of the infection in the mouse model (Skamene, 1991). Using linkage analysis, described in more detail in section 1.3, the *Bcg* gene was localized to chromosome 1. The *Bcg* locus is believed to be identical to the *Lsh* and *Ity* loci which control host response to infection by *Leishmania donovani* and *Salmonella typhimurium* (Schurr et al., 1991a).

The *Bcg* gene is expressed in mature tissue macrophages and it appears that the macrophages in *Bcg* resistant mice can be more easily activated by phagocytosis of *M. bovis* or other stimulating factors, and thus can respond more quickly to the presence of mycobacteria than *Bcg* susceptible mice. The *Bcg* gene function is present *in vivo* even when T cells are depleted so that the differential in the early response to mycobacteria is still observed, whereas both *Bcg* resistant and susceptible mice allow mycobacterial growth 18 days after infection (Gros et al., 1983). Various phenotypic markers of activated macrophages are expressed at higher levels in the cells of resistant mice (Schurr et al., 1991a), and experiments *in vitro* have demonstrated the superior bactericidal ability of *Bcg* resistant macrophages (Denis et al., 1990). Macrophages isolated from *Bcg* susceptible mice can exert similar anti-microbial capability if activated by lymphokine treatment (Denis et al., 1990).

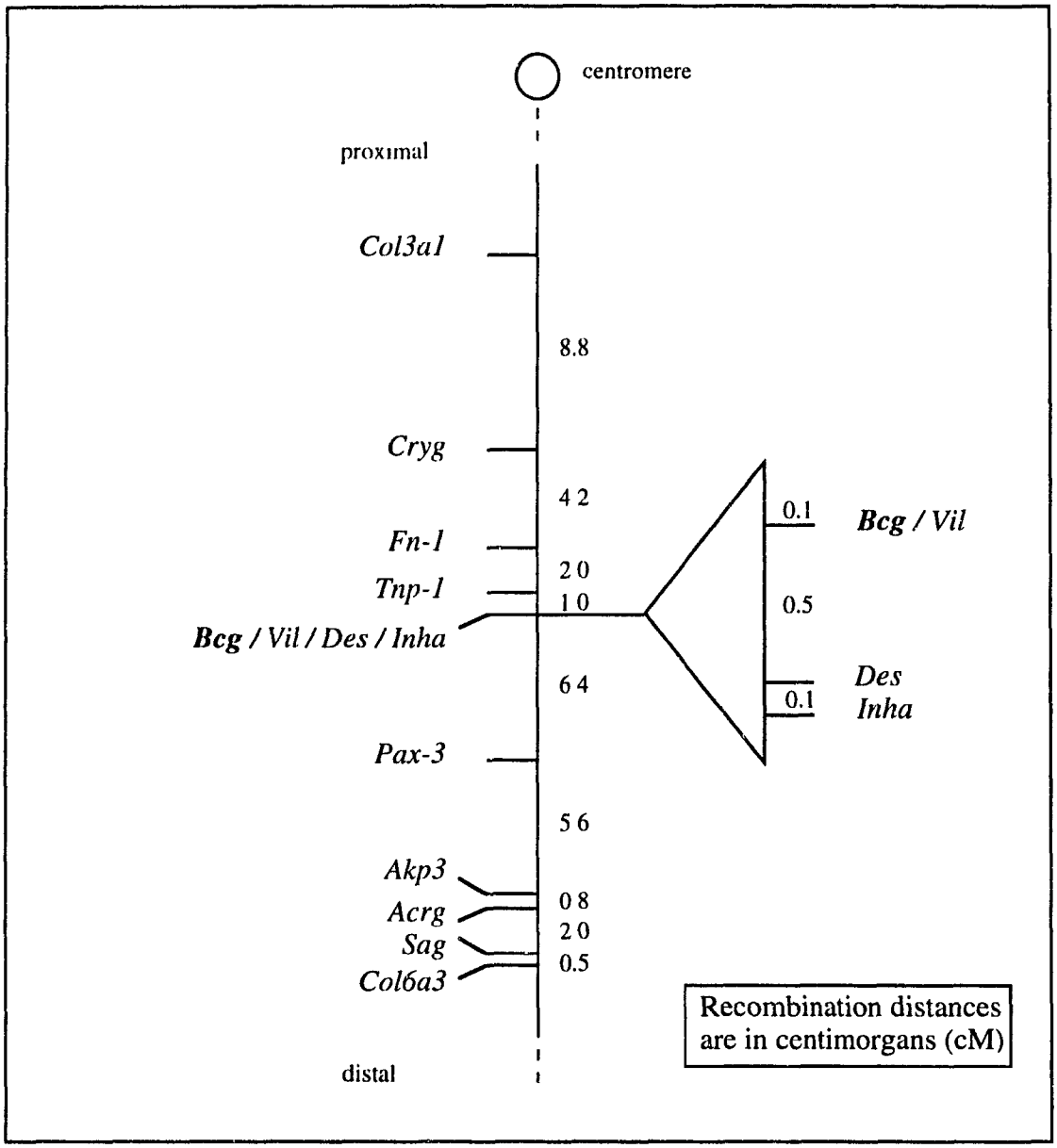
The role of the *Bcg* gene thus appears to be in the priming of macrophages for activation so that resistant tissue macrophages, by an unknown mechanism, can restrict intracellular multiplication of *M. bovis* independently of interaction with lymphocytes such as T cells (Skamene, 1991). In addition, *Bcg* resistant macrophages have been shown to have enhanced ability to present antigen to T cells (Denis et al., 1990) and thus may have a role in the later stage of cell-mediated immunity to mycobacteria through more efficient T cell interaction (Skamene, 1991). In comparison, isolated macrophages from the spleens of *Bcg* susceptible mice can suppress T cell proliferation *in vitro* following a high dose inoculum of *M. bovis* (Schurr et al., 1989a).

To generate a map of loci in the proximity of *Bcg*, several markers on mouse chromosome 1 have been tested by linkage analysis for segregation of their alleles with

the early response of resistance versus susceptibility. The linkage analysis was carried out with backcross mice and recombinant inbred strains. In a backcross, the heterozygous progeny of a cross between inbred homozygous resistant and homozygous susceptible strains are crossed "back" to homozygous susceptible or resistant mice to generate expected proportions of progeny with known genotypes. Recombinant inbred strains are inbred descendants of third generation offspring from two well-defined, inbred progenitor strains; in this way, they are a replicable, homozygous source of recombinations present in the third generation (Skamene, 1991). An integrated map of 12 loci within a 30 cM segment thought to contain *Bcg* has been generated through linkage analysis by several investigators and is displayed in Figure 2 (Schurr et al., 1989b, 1989c, and 1990b; Malo et al., 1991 and 1993; Epstein et al., 1991; Vidal et al., 1992). Of the loci presented in Figure 2, *Bcg* has been positioned closest to the villin locus (Schurr et al., 1989a and 1989c; Malo et al., 1991 and 1993).

A candidate gene for *Bcg*, named *Nramp* (for natural resistance-associated macrophage protein gene) was recently isolated and shown to be expressed in macrophage cells from reticulo-endothelial organs of mice (Vidal et al., 1993). The predicted amino acid sequence of the protein encoded by *Nramp* suggested that the protein was a membrane transporter with structural similarity to known transporter systems in both prokaryotes and eukaryotes (Vidal et al., 1993). Further analysis will be required to determine the mechanism by which the encoded protein affects the macrophage response to mycobacteria. A candidate mutation associated with susceptibility has been identified (Vidal et al., 1993).

A genetic region containing the loci which are linked to the murine *Bcg* gene, from *Col3a1* (collagen type 3 alpha 1) to *Col6a3* (collagen type 6 alpha 3), is conserved on the long arm of human chromosome 2, in a region stretching from band 32 to the end (telomere, denoted as "ter") of the long arm (Schurr et al., 1990b). The distal portion of chromosome 2q is thus the most likely region to contain a homologous human "BCG" gene. If the early immune response to injected *M. bovis* in inbred mouse stains is relevant to humans exposed to inhaled *M. tuberculosis*, then a human "BCG" locus may have a role in susceptibility to tuberculous disease, assuming the early response is a major



source: Malo et al., 1993a.

Figure 2. Composite genetic map of mouse chromosome 1 in the vicinity of the ***Bcg*** locus. The full names for the loci are as follows: *Col3a1*: collagen type 3 alpha 1; *Cryg*: gamma crystallin complex; *Fn-1*: fibronectin 1; *Tnp-1*: transition protein 1; *Bcg*: bacille Calmette-Guérin; *Vil*: villin; *Des*: desmin; *Inha*: inhibin alpha; *Pax-3*: paired box homeotic gene 3; *Akp3*: alkaline phosphatase-3, intestine, not Mn requiring; *Acrg*: acetylcholine receptor gene; *Sag*: S-antigen; *Col6a3*: collagen type 6 alpha 3.

determinant of clinical outcome. The possible role of a human "BCG" gene in tuberculosis susceptibility could be examined with linkage analysis, using available DNA markers which map to the chromosome 2q32–37 region.

1.3 Linkage analysis

1.3.1 The lod score method

Linkage analysis is used to estimate the genetic distance between loci and map a specific marker with respect to other loci. A statistical test assesses whether the null hypothesis of independent inheritance of alleles, which is indicated by a recombination frequency of 50% between two loci, can be rejected in favour of the alternate hypothesis of linkage, which is indicated by a recombination frequency of less than 50%. Two-point linkage analysis with a marker of interest and a marker of known location can be used to develop a linkage map, given certain assumptions (Ott, 1991). Multi-point analysis permits simultaneous calculation of the maximum likelihood for different orders of the loci.

The study unit for linkage analysis is a family with at least two generations of individuals whose phenotypes are observed at marker loci. To increase the power of the analysis, several families can be studied. If the families are unrelated then the observations on each family are statistically independent, and the evidence for or against linkage of the loci can be added across the families (Ott, 1991). For linkage analysis with a putative disease or susceptibility gene, clinical and genetic marker information is used to examine whether the disease trait is inherited independently of marker alleles in families with diseased members. Linkage analysis has the advantage of being able to provide evidence for a disease or susceptibility gene when population association does not exist between mapped loci and the locus of interest (Morris, 1993). In addition, by analyzing related persons linkage analysis studies avoid the problem of differential ethnic background between cases and controls, which can arise in a traditional association study.

In the lod score method of linkage analysis, the maximum likelihood method is used to estimate the recombination fraction between loci; the associated statistic, which measures the support for linkage at that recombination fraction versus absence of linkage,

is called the lod score. The lod score is the logarithm (to base 10) of the odds that two loci are linked versus they are unlinked. The lod score for a given recombination fraction, denoted as theta (θ), is defined as:

$$Z(\theta) = log_{10} [L(\theta)/L(½)]$$

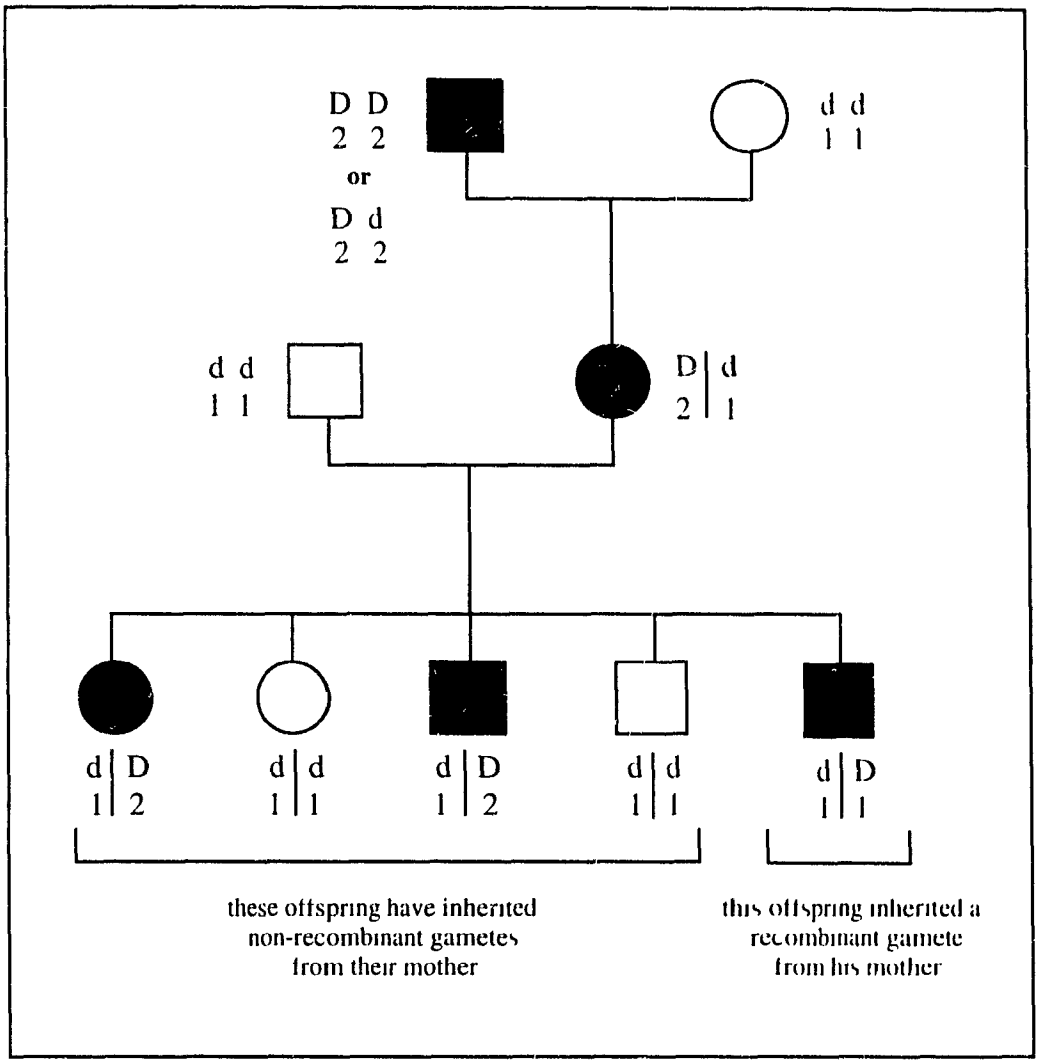
where $L(\theta)$ is the likelihood of the data given $\theta < 0.5$ and $L(½)$ is the likelihood of the data given $\theta = 0.5$. The lod score for a recombination fraction of 50% will always equal zero since the likelihood of the numerator and denominator in the ratio will be the same and $Z(\theta=½) = log_{10} [1]$.

For two randomly selected loci, there is a prior probability of about 2% that they will be linked at a recombination frequency of less than 30% (Elston and Lange, 1975). An odds ratio of 1000:1 is the criterion for a significant lod score because the prior odds for linkage (0.02) multiplied by the odds for linkage provided by the data (1000) yields a posterior odds for linkage of 20:1. An odds ratio of 1000:1 provides a posterior probability of linkage or significance level of 95% (20/21) and a posterior type I error of 5%, and corresponds to a lod score of 3 ($log_{10}[1000]$) (Risch, 1992).

Generally, a lod score greater than 3 for a particular recombination fraction implies that this fraction can be used to estimate the most likely genetic distance between two loci. A positive lod score will be computed as long as no recombinants are observed. A negative lod score indicates there is evidence against linkage of two loci at the specific recombination frequency. A lod score of -2 (100:1 odds against linkage) or less indicates statistically significant exclusion of linkage at the recombination frequency specified. Lod scores between -2 and 3 do not support or reject the null hypothesis with statistical significance. A family is considered informative for linkage when a non-zero lod score is computed for any recombination fraction less than 0.5 (Ott, 1991).

1.3.2 Example

The determination of the evidence for linkage provided by a family can be demonstrated with a simple example (adapted from Gusella et al., 1984) (Figure 3). Imagine we have clinical and marker information for a family composed of two parents and five offspring. In addition, we have information for the two maternal grandparents.



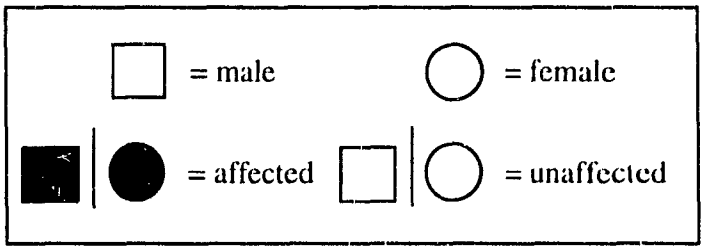


Figure 3. Sample family for linkage analysis (adapted from Gusella et al., 1984)

The disease gene has two alleles, D and d, and there are two alleles, 1 and 2, at the marker locus. A completely penetrant disease that is inherited in an autosomal dominant manner occurs in the family, so that the inheritance of one or two copies of the abnormal allele, D, always leads to disease expression. There are no non-genetic causes of the disease.

One affected grandparent is inferred to be 2/2 based on typing at the marker locus, and because he is affected he must be D/D or D/d at the disease locus. The unaffected female grandparent is inferred to be 1/1 at the marker locus and must be d/d at the disease locus, which makes her homozygous at both loci or doubly homozygous. The affected mother is heterozygous 1/2 for the marker locus and she must be D/d at the disease locus. She is thus inferred to be a doubly heterozygous, D-2/d-1 individual, where the vertical line separates haplotypes since the D-2 gamete is inferred to have been inherited from her father, and a d-1 gamete was received from her mother. Because we can infer which gamete was inherited from which parent we know the phase of the mother's typing at the two loci.

The unaffected father of the five offspring, who has 1/1 typing at the marker locus, is inferred to be a d-1/d-1 individual. He and his affected mate in Figure 3, therefore, represent a double backcross mating. Based on the observed disease phenotypes and marker typing data, the five children are inferred to be d-1/D-2 (affected); d-1/d-1 (unaffected); d-1/D-2 (affected); d-1/d-1 (unaffected); and d-1/D-1 (affected). Four children received a non-recombinant gamete from their mother, whereas one child received a recombinant gamete. The best estimate of the recombination frequency in the pedigree is 1/5 or 20%, where the recombination fraction, (θ), equals the number of recombination events detected divided by the total number of opportunities for recombination. Thus there is evidence for linkage of the marker and the disease loci in the pedigree.

The lod score provided by this family is calculated analytically as follows (adapted from Ott, 1991). Given the affected mother's inferred phase in Figure 3, she has passed on four non-recombinant gametes, each of these events occurring independently with a probability of $1 - \theta$, and has passed on one recombinant gamete with a probability of θ.

The likelihood of the observed data is, aside from a combinatoric constant,

$$L(\theta) = (1 - \theta)^4 (\theta)^1$$

The likelihood ratio (LR) of interest is

$$LR = L(\theta)/L(\tfrac{1}{2})$$

We wish to calculate the lod score $Z(\theta)$ where

$$Z(\theta) = log_{10} [L(\theta)/L(\tfrac{1}{2})]$$

In this example, the lod score is

$$Z(\theta) = log_{10} [(1 - \theta)^4 (\theta)^1/(1 - \tfrac{1}{2})^4 (\tfrac{1}{2})^1]$$

$$Z(\theta) = log_{10} 32 [(1 - \theta)^4 (\theta)^1]$$

The lod score can be evaluated for several assumed values of theta, θ. The theta with the highest lod score is the best estimate of the recombination fraction (Ott, 1991). For values of theta of 0, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4, the maximum lod score of 0.42 occurs at $\theta = 0.2$.

In the sample family, meiosis in the father does not provide information about linkage. Any recombination between the two loci would result in the same combination of alleles because he is homozygous at both loci. A recombination would also not be detectable if he had been homozygous at only one of the loci. For a family to be informative for linkage analysis with the lod score method, at least one parent must be heterozygous at both the loci (e.g., the disease and the marker loci). Markers with many alleles in the population, and more even allele frequencies, are useful for linkage analysis because there is a greater chance for a parent to be heterozygous at the marker locus. If a recessive genetic disease is being studied, families with two affected parents would not be informative because the parents would be homozygous at the disease locus in the absence of non-genetic cases.

It is also clear that without the typing information for the maternal grandparents we would not be able to infer the phase of the mother's typing: she could be phased as either D-2/d-1, having received D-2 from one parent and d-1 from the other, or d-2/D-1, having received d-2 from one parent and D-1 from the other. Given the first phase, she has passed on four non-recombinant and one recombinant gamete, whereas given the second phase, she has passed on one non-recombinant and four recombinant gametes.

The likelihood of the observed data for this family, allowing for a probability of occurrence of each phase of ½ and excluding the combinatoric constant, is then

$$L(\theta) = \tfrac{1}{2}\,[(1-\theta)^4\,(\theta)^1 + (1-\theta)^1\,(\theta)^4]$$

and the lod score is

$$Z(\theta) = \log_{10} \frac{\tfrac{1}{2}\,[(1-\theta)^4\,(\theta)^1 + (1-\theta)^1\,(\theta)^4]}{\tfrac{1}{2}(1-\tfrac{1}{2})^4\,(\tfrac{1}{2})^1 + (1-\tfrac{1}{2})^1\,(\tfrac{1}{2})^4]}$$

$$Z(\theta) = \log_{10} 16\,[(1-\theta)^4\,(\theta)^1 + (1-\theta)^1\,(\theta)^4]$$

For values of theta of 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, the maximum lod score of 0.12 occurs at $\theta = 0.2$. There is, therefore, a loss of information when the mother's phase is not known, resulting in a smaller lod score. Because of unknown phase and other factors which can complicate the analytical calculation of a lod score, such as incomplete penetrance, complex pedigrees, and missing data, computer programs have been developed to carry out the lod score method of linkage analysis.

1.3.3 Computer programs

In this section, general information will be presented about the computer programs used for the lod score analysis in this thesis. The LINKAGE analysis package contains a group of programs for linkage analysis of general pedigrees (this thesis), as well as three-generation, phase-known special pedigrees (Lathrop et al., 1984). In particular, the program MLINK calculates two-locus and multiple-locus likelihoods for a specified set of recombination fractions (for example, 0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5). The program ILINK finds the likelihood of the data and the maximum likelihood estimate of the recombination fraction and other parameters (e.g., allele frequencies) by iteration. A third program available for the analysis of general pedigrees, LINKMAP, (not used in this thesis), provides the likelihoods of a "test" locus at various locations on a known map of marker loci. The lod score method of linkage analysis carried out by the LINKAGE programs is parametric, requiring the specification of genetic models in order to analyze family data.

An assumption of human linkage analysis is that the population from which the study families are drawn is in Hardy-Weinberg equilibrium (Ott, 1991). Under the equilibrium assumption, mating in the population is random; that is, mates are not selected on the basis of kinship or phenotype, and thus genotype, at the study loci. It is also assumed that in the population all possible genotypes are equally fertile, natural selection does not affect the gene of interest, mutation can be disregarded, migration is negligible and the population is infinite in size (Hartl and Clark, 1989). Under these conditions the Hardy-Weinberg law states that, after one generation of random mating and in the absence of variation of allele frequency with gender, the genotype frequencies of individuals for an autosomal trait depend only on the allele frequencies in the population. The allele frequencies, and thus the genotypic make-up of the population, will remain constant over time. The law is quite robust to departure from several of the assumptions in human populations. Numerically, if p and q are the population frequencies of the normal resistant (R) and the susceptibility allele (r), respectively (where $p = 1 - q$), the probability of being a R/R or r/r individual equals p^2 and q^2, respectively and the probability of having a R/r or r/R genotype is $2pq$.

In the lod score method of linkage analysis, the mode of inheritance and the allele frequencies at any trait or marker locus considered must be described with genetic models. The likelihood of different modes of inheritance of the disease can be obtained by segregation analysis of the disorder. In the absence of formal analysis, inheritance pattern can sometimes be indicated by inspection of affected pedigrees. Penetrance functions are used in the LINKAGE programs to express the relationships between genotypes (to be inferred) and phenotypes (observed). For example, to specify particular modes of inheritance for an autosomal disease trait, the phenotype data for the locus can be designated as an affection status and penetrances are assigned to genotypes in a table. For the particular case of a dominant trait without non-genetic causes, zero probability of disease is assigned to the resistant genotype (with two copies of the normal allele) and non-zero probability of disease is assigned to the susceptible genotypes (with one or two abnormal alleles). For the case of a recessive trait without non-genetic causes, the only genotype which is assigned a non-zero probability of disease is the r/r genotype. For

affection status loci, a separate variable can be included in addition to clinical phenotype to designate a risk or liability class if, for instance, different individuals with the same genotype have different penetrances due to another factor, such as age. Co-dominant autosomal markers can be designated in the LINKAGE programs as numbered alleles, for which the phenotype is two allele numbers that represent the genotype in a fully-penetrant manner.

The linkage analysis programs calculate the probability of various genotypes given the genetic model and the family data. For the calculations in founders with missing marker typing or unknown disease status, the specified allele frequencies are especially important. Marker allele frequencies can be estimated from typing data in a similar ethnic group as the study pedigree or can be estimated using the observed typing data in unrelated founders in the study families. Iterative estimates of the marker allele frequencies can be obtained from the observed typing data using the ILINK program. The accuracy of the estimates will depend on the sample size. For a highly polymorphic marker, Ott (1992) suggests that at least 80 individuals are needed for detection of most alleles and accurate allele frequency estimation.

Segregation analysis of the disease, corrected for ascertainment method, can provide allele frequencies for the disease or susceptibility gene, in addition to likelihood analysis of the mode of inheritance. In the absence of such analyses, the prevalence of the disease at a point in time in a defined population can be used to estimate the frequency of the disease or susceptibility allele, under certain assumptions. For a disease which can arise from a single gene defect and has other non-genetic causes, and assuming Hardy-Weinberg equilibrium,

$$K = f_{qq}\, q^2 + f_{pq}\, 2pq + f_{pp}\, p^2$$

where K is the prevalence of disease, q and p are the frequencies of the abnormal and normal alleles, respectively, and f_{qq}, f_{pq}, and f_{pp} are the probabilities of disease given genotype (penetrance). Depending on the mode of inheritance of the genetic susceptibility and the penetrance of each genotype, particular genotypes will have higher or lower risks of disease (that is, people with the resistant genotype have a baseline risk of disease for non-genetic reasons, and those who are genetically susceptible have an additional amount

of risk). For a single gene disease with autosomal dominant inheritance which also arises in the population due to non-genetic factors,

$$K = \underbrace{f_{qq}\, q^2 + f_{pq}\, 2pq}_{\text{higher risk}} + \underbrace{f_{pp}\, p^2}_{\text{lower risk}}$$

where a proportion of cases arise in those with the lower risk genotype with two normal alleles. If phenocopies occur in the population and genetic susceptibility to the disease is inherited in a autosomal recessive manner, only the genotype with two abnormal alleles will be at higher risk for disease in the above equation. For an autosomal dominant disease without non-genetic causes,

$$K = f_{qq}\, q^2 + f_{pq}\, 2pq$$

For an autosomal recessive disease without non-genetic causes, the expression

$$K = f_{qq}\, q^2$$

provides a lower bound for the abnormal allele frequency, since the allele is also possessed by heterozygous individuals who do not express the disease.

1.3.4 The Elston-Stewart algorithm

Computer algorithms have been developed to carry out the lod score method on general pedigrees including complex situations in which there are large families with several generations, pedigrees with missing data, or marriage or consanguinity loops. In addition, for some diseases an abnormal or susceptible genotype does not always lead to illness, so that the susceptible genotype has reduced penetrance; that is, there is a probability of less than 100% of developing the disease. Reduced penetrance for a particular genotype is specified for an affection status locus in a genotype/phenotype table, with a probability of disease of less than 1.

The Elston-Stewart algorithm (1971), implemented in the MLINK and ILINK programs, carries out the maximum likelihood estimation of the recombination fraction between two loci in a recursive series of conditional probabilities (Ott, 1991). In a two-point analysis, the likelihood of pedigree data is the probability of observing the phenotypes at two loci for the $i = 1 \ldots n$ family members, $P(x_i)$ or $P(x_1, x_2, \ldots, x_n)$, where

x_i is the phenotype at two loci jointly for each person. The conditional probabilities of the phenotypes given the underlying genotypes are statistically independent in a family, so that the likelihood can be expressed as

$$L = \sum_{i=1}^{n} \sum_{g_i} P(x_1, x_2, \ldots, x_n \mid g_1, g_2, \ldots, g_n)\, P(g_1, g_2, \ldots, g_n)$$

where, for each $i = 1 \ldots n$ individual, the probabilities of the genotype $P(g_i)$ and the conditional probabilities of the phenotype given the genotype $P(x_i \mid g_i)$ are multiplied and then summed over all possible genotypes. For a pedigree member with unknown clinical status, $x_i = 0$, so that the likelihood of the data for the person equals 1 if he/she is a founder, and the likelihood is a function of parental genotypes otherwise.

Using the above representation of the pedigree likelihood, the Elston-Stewart algorithm carries out calculations using data supplied from parents to children, the likelihood of the data expressed as the following:

$$L(\theta) = \sum P(x_1 \mid g_1)\, P(g_1 \mid \bullet) \ldots \sum P(x_{n-1} \mid g_{n-1})\, P(g_{n-1} \mid \bullet) \sum P(x_n \mid g_n)\, P(g_n \mid \bullet)$$

where $P(g_i \mid \bullet)$ is an offspring's genotype given the genotypes of his/her parents or a founder's genotype which depends only on the population genotype frequencies under Hardy-Weinberg equilibrium. The algorithm begins at the last n sum, which corresponds to the last offspring entered in the data file. The multiplication of the two probabilities for this person is carried out for each genotype in the next to last sum, and these products are then added together. The strategy proceeds towards the first sum in the expression (one of the parents), producing summation results for each person. The set of summations for the entire family are multiplied together. This sequence of calculations is completed for each given value of the recombination fraction.

1.3.5 Complex genetic disease

Complex genetic diseases include diseases with a genetic component which have reduced penetrance, environmental causes, and/or genetic heterogeneity (defined below), or are caused by several loci which may interact. A disease with a genetic component can also be considered complex if the assignment of clinical status is difficult.

For a susceptibility locus with reduced penetrance, there may be another condition, perhaps environmental or even genetic, that is required for disease to occur. An example of reduced penetrance is a recessive cancer model, in which a person genetically predisposed to malignancy, having inherited one susceptibility allele, will not develop the disease unless he/she receives an environmental "hit" by a disruptive, mutating agent to produce the second, necessary susceptibility allele. Inherited retinoblastoma is an example of such a disease. As another example of reduced penetrance, it is well known that certain genetic diseases have an age at onset distribution (for example, Huntington's disease). For such disorders, probability of illness may be low at young age but increases as the susceptible individual becomes older. Such variation of disease risk is modelled with the LINKAGE programs by designating several age-specific liability classes for the susceptibility locus, to which family members are assigned on the basis of their age.

If non-genetic forms of a disease exist, persons who have inherited normal alleles at the susceptibility locus may develop the disease as the result of other factors, such as environmental exposure to a pathogenic agent. These persons are termed phenocopies, or sporadic or unlinked cases. If phenocopies occur, then the abnormal allele is not necessary for disease (Greenberg, 1993). With tuberculosis, socioeconomic conditions and immunocompetence can be worsened to a sufficiently low level so that disease following infection is very likely to occur, irrespective of genotype. In the LINKAGE programs the possibility of phenocopies can be incorporated by assigning a non-zero penetrance to lower risk genotypes that do not contain the abnormal allele.

A disease may be caused by several genes rather than just one, each contributing to the clinical outcome, which is known as polygenic inheritance. The limitation of linkage analysis with the LINKAGE programs used in this thesis is the underlying assumption that the observed clinical phenotype is the result of a single gene[3] (Pauls, 1993). With the involvement of many genes, reduced penetrance, and the absence of a single sufficient gene, linkage between a disease trait and a marker locus may be difficult to detect. Both theoretical and empirical studies have shown, however, that the effects

---

[3]A recent version of the programs, called TLINKAGE, has been developed for the analysis of a disease caused by the interaction of two genes (Lathrop and Ott, 1990).

of loci with a minor role in the familiality of a disease can be detected with linkage analysis, an example being the strong evidence in favour of an HLA linkage with type I diabetes, despite the likely contribution of other genes to the disorder (Risch, 1990).

For a genetically-heterogeneous disease, the same disease phenotype is caused by different abnormal alleles (Risch, 1990). There are two types of genetic heterogeneity, which are classified according to whether the abnormal alleles exist at the same locus (intra-locus or allelic heterogeneity) or at different loci (inter-locus or non-allelic heterogeneity). When a group of families segregating a genetically heterogeneous disease are analyzed for linkage to a particular marker, it is possible to fail to detect any evidence for linkage, particularly if the proportion of truly linked families for the given marker is not large (e.g., 10–30%) (Greenberg, 1992). When several affected families are analyzed, particularly if more than one ethnic population is represented, a combination of strongly negative and positive lod scores is a good indication that the specific marker and the disease trait are linked in only a proportion of the families (Risch, 1990). The HOMOG computer program has been developed to provide a statistical test for genetic heterogeneity, also known as the admixture test, indicating whether it is likely that only a proportion of the tested families show linkage between the marker and disease loci (Ott, 1991).

The most crucial requirement for linkage analysis with a disease trait is the appropriate definition of the diseased state, using either a continuous or dichotomous phenotype (Chakravarti and Lander, 1990). The goal is to use a valid clinical classification scheme which reflects the underlying genotypes in the pedigrees (Kringlen, 1993). A dichotomous "case" or "non-case" designation relies upon correct diagnosis of the diseased and the normal states. A quantitative phenotype can be more powerful for linkage analysis with a complex disease, as the phenotype can often capture more clinical information with respect to symptom expression than a categorical definition; a continuous phenotype may thus allow detection of genetic linkage with certain subtypes of a complex disorder (Pauls, 1993). A phenotype should be defined to best characterize the form of the disease under study. Broad disease definitions may include a variety of underlying clinical entities so that the effect of a particular gene for a specific subtype,

while present, can be diluted and not detectable by linkage analysis. Disease definition has been a challenge for researchers attempting to find linkage for psychiatric disorders such as schizophrenia and manic depression (Pauls, 1993).

One way to approach this challenge has been to allow variation in the disease definition. Several analyses are run, each time with exclusion of more broadly defined cases such as those with milder or more non-specific symptoms, until a very restrictive disease definition is being used. Evidence for linkage may only be found when a particular clinical subtype of the disease is used for case definition; such a result may reflect an underlying pathology with a genetic component or inter-locus genetic heterogeneity. Another method involves applying a level of diagnostic uncertainty to the penetrance of the susceptible genotypes, which is particularly useful with indeterminate diagnoses. In this way, a probability of a false positive and a false negative diagnosis can be incorporated into the linkage analysis.

1.3.6 Non-parametric methods

Non-parametric methods of linkage analysis avoid the specification of a genetic model of the inheritance of the disease trait and can, therefore, be more robust than parametric methods. A weakness of non-parametric analysis is that it is much less powerful than parametric methods, especially where the genetic model is well-characterized (Terwilliger and Ott, 1993b). Non-parametric methods have not been used in this thesis.

The general premise behind non-parametric methods of linkage analysis is that for a disease trait with a genetic basis, affected relatives are more likely to share alleles at a marker locus closely linked to a disease or susceptibility gene (Terwilliger and Ott, 1993b). Statistical tests are used to assess the evidence in favour of increased marker similarity between all combinations of pairs or sets of affected family members. With sufficient pedigree data, these methods are most powerful when the alleles that are shared are traceable as copies from the same parent, that is, the alleles are identical by descent. Marker informativeness is reduced when the analysis must rely upon sharing of alleles that are identical by state (Ott, 1991). In addition, allele sharing cannot be detected for markers with a low level of polymorphism (Sandkuijl, 1993).

In the sib-pair method, for example, pairs of siblings are expected to share 0, 1, or 2 alleles identical by descent in the ratio 1:2:1. In the absence of linkage between the marker and the disease trait, the same ratio is expected for a pair of affected siblings. Significant deviation from the expected proportions indicates an association between marker typing and disease status that may be due to linkage of the two loci.

Although information about the inheritance and frequency of the disease or susceptibility allele is not required for non-parametric methods, the mode of inheritance of the disease does affect the power of the analysis with, for example, affected sibling pairs. As shown by Sandkuijl (1993), it is easier to detect a significant deviation from expected allele sharing by affected siblings for a recessive disease. Non-parametric tests of linkage may frequently be falsely negative due to their limited power. As a result, large numbers of sibships are required for analysis (Morris, 1993). The power of affected sib-pair analysis decreases rapidly with increasing recombination distance between marker and disease loci (Morris, 1993), and is significantly lowered in the presence of phenocopies (Ott, 1991).

An extension of the sib-pair method to sets of other affected relatives has been developed, known as the affected pedigree member method of linkage analysis (Weeks et al., 1993). This method utilizes a weighting factor based on marker allele frequency to take into account sharing of rare alleles by distant relatives (Lange and Weeks, 1990). It is thus crucial for marker allele frequencies to be accurately estimated in an affected pedigree member analysis. Recent studies have shown that false positive evidence for linkage can arise with this method when allele frequencies are incorrectly specified (Wijsman, 1993). Pedigrees containing only one affected parent/child pair should be avoided because the probability of a false positive result can greatly increase when such pedigrees are analyzed (Weeks et al., 1993).

1.3.7 Selection of families

The number and size of affected families required for lod score analysis with a disease trait depends upon the penetrance of the disease and the linkage phase information available (Ott, 1991). The power of linkage analysis relates to the number of opportunities for recombination required to detect linkage if it exists. The following

calculations of sample size have been presented by Ott (1991). The calculations involve determination of the expected lod score (ELOD), which is a weighted average of $Z_i(\theta)$ for i phenotypic outcomes and a fixed θ, under different conditions; the results presented below are calculated for specific mating types.

Given a recombination fraction of 5% and a power of 80%, approximately 20 phase-known meioses are required to find evidence for linkage with a lod score of 3, using an approximate formula modified from Elandt-Johnson (1971). In the presence of unknown phase, one more offspring is required for each available sibship if linkage is tight. When the penetrance of the susceptible genotype is 50%, calculation of the ELODs in phase-known backcross families indicates that a three-fold increase in the number of meioses is required; the factor of increase is the inverse of the ratio of the ELODs for complete versus incomplete penetrance. With a penetrance of 80%, the required number of meioses is increased by a factor of about 1.6. A laboratory error rate of 5% for the genetic typing data requires a further increase by a factor of 1.3; the increase corresponds to the inverse of the relative ELODs for none versus 5% misclassification at a recombination frequency of 5%. With penetrance and error factors considered, the total number of phase-known meioses required is 78 for 50% penetrance and 42 for 80% penetrance. If both parents are informative, these requirements correspond to 39 and 21 offspring, respectively. When there is genetic heterogeneity such that only 80% of the study pedigrees are linked, the number of required phase-known meioses is increased by a factor of 3.2 for a recombination frequency of 5%. The increase in this case is equal to the inverse of the ratio of the variance in the recombination fraction with 100% versus 80% of the families being linked.

For a linkage analysis study, the type of family to be collected must also be considered. A small number of large pedigrees with many affected members can be selected or, alternatively, one might choose a larger number of small families with fewer cases in each pedigree. There are advantages and disadvantages associated with each of these strategies. Enough information can often be provided by one or two large, multiplex families to localize a disease locus (Greenberg, 1992). The increased informativeness of a large pedigree is the result of the additional phase information

available. However, large, dense pedigrees may be rare, particularly if the abnormal allele is very low in frequency, can be more difficult to collect, and are more likely to represent a chance combination of multiple genetic and/or environmental factors for complex traits (Greenberg, 1992). In addition, linkage results from large pedigrees are very sensitive to errors in typing or diagnosis (Greenberg, 1992). The large effect of a change in clinical status from unaffected to affected for two family members, as well as new marker typing information for two unaffected individuals, was demonstrated in a linkage analysis of affective disorder in a large Amish family. In a re-evaluation of the family, a significantly positive lod score of 4.08 at 0% recombination between the trait and a chromosome 11 locus decreased to a non-significant lod score of 1.03 at 17% recombination with the new information (Kelsoe et al., 1989).

Although more families would be required, the use of smaller families may mean greater ease in locating pedigrees and collecting complete data for the linkage analysis. Most importantly, these families may be more representative of the general population in which the disease occurs. Smaller pedigrees may be useful for other types of linkage analysis, such as the sib-pair method. When a complex disease is being studied, it may be possible to confirm positive results from parametric linkage analysis with a non-parametric method. Alternatively, non-parametric analysis may be able to indicate whether negative results from a lod score analysis may have been the result of parameter misspecification (Sandkuijl, 1993).

1.3.8 Effects of parameter misspecification

The specification of the required parameters for the lod score method of linkage analysis can be difficult if the genetic model is unknown or unclear. In this situation, one must use the most likely and reasonable parameter estimates that will not bias the results in a false positive direction. With respect to disease phenotype, the misclassification of an affected person as unaffected, known as a false negative, will tend to decrease the lod score and thus the power to detect linkage, but should allow detection of linkage if incomplete penetrance is specified in the analysis model, as pointed out by Greenberg (1992) and Baron et al. (1990). The designation of an unaffected individual as affected, known as a false positive, will force an incorrect recombination between the marker and

the disease locus if the two loci are truly linked, and close linkage will not be detectable (Greenberg, 1992; Martinez et al., 1989). Linkage may also be missed when an analysis is carried out on several families and unknown genetic heterogeneity exists. However, there can be enough power to observe linkage despite heterogeneity in a relatively large data set, if more than 30% of the available families show true linkage (Greenberg, 1992).

Experience has shown that the misspecification of penetrance, susceptibility allele frequency, and phenocopy rate does not tend to lead to false evidence in favour of linkage (Risch et al., 1989). It should still be possible to detect linkage if it exists, but the collection of more families may be required. Specification of 100% penetrance when penetrance is in fact incomplete will bias the recombination frequency upward, toward 50%, by causing obligate recombinations between the trait and marker (Ott, 1991). Misspecification of the mode of inheritance of the disease can significantly decrease the power of linkage detection, leading to a great reduction in the maximum lod score as described by Risch et al. (1989). These authors suggest using two robust models with particular abnormal allele frequencies and penetrances for dominantly and recessively inherited traits, respectively. The models should allow detection of linkage if it is present, but cannot be assumed to provide accurate estimates of the true recombination fraction.

The power to detect linkage is not greatly affected by inaccurate marker allele frequencies, as shown by Freimer et al. (1993). In contrast, the use of incorrect marker allele frequencies when marker typing data is not complete for all parents in a pedigree can increase the chances of a false positive linkage result (Ott, 1992). In the absence of reliable estimates of allele frequencies for the markers used, alleles have frequently been specified as equal in frequency in linkage analysis studies. Ott (1992) has demonstrated the bias of recombination fraction estimates in the false positive direction when allele frequencies are unequal but are assumed to be equal in the analysis. This likely occurs because a disease allele will tend to segregate in a family most often with the particular marker allele which is most common. If the common allele is taken to be more rare by using equal allele frequencies, an association could be induced so that apparent evidence for linkage is gained from the frequency of co-occurrence of the marker and disease

alleles in affected individuals (Ott, 1992). It is thus usually preferable to use reliably estimated marker allele frequencies rather than to assume equal frequencies.

1.3.9 Significance level of the lod score

As in an association analysis, the choice of a suitable marker locus for a linkage analysis study can involve the anonymous or the candidate marker method. With the anonymous approach, many previously mapped markers are usually selected at various chromosomal locations, whereas the candidate method can involve the formation of a hypothesis for the biological mechanism by which an abnormal allele may lead to the disease. There is some controversy concerning the value of the lod score required for statistically significant linkage when a candidate marker is used in a linkage analysis. It has been suggested that a lower threshold needs to be met by the lod score to accept linkage because of the higher prior probability that the candidate locus may be involved in disease etiology (Wijsman, 1990). In other words, a positive lod score of less than 3 may be considered statistically significant when a candidate marker is studied. Ott (1991) agrees with this reasoning but points out that localization of disease genes following lod score results of less than 3 has not been successful in practice thus far, suggesting that the gain in prior probability is only marginal. In his opinion, the candidate approach does not warrant a large departure from the classically required odds of 1000:1 in favour of linkage.

Testing multiple markers for linkage with a disease trait occurs particularly often when the anonymous approach of marker selection across the genome is taken. The testing of multiple, independent (i.e., unlinked) loci for linkage to a single gene does not increase the false positive rate of the analysis, as shown by Risch (1991; 1992). As the number of markers tested increases, the prior probability of linkage rises as does the significance level. As a result, the posterior type I error decreases and a higher critical lod score is not required for statistical significance (Ott, 1991), as demonstrated below:

	prior odds for linkage	×	*odds given data (critical lod score: 3)*	=	*posterior odds*
with 1 marker,	*0.02*	×	*1000*	=	*20:1 (95% probability of linkage); type I error 5%*
with 5 markers,	*0.10*	×	*1000*	=	*100:1 (99% probability of linkage); type I error 1%*

It is suggested by Ott (1991), however, that the above results apply only for the analysis of a single gene disorder. For the analysis of a complex trait, he advises that it may be preferable to adjust the critical lod score for statistical significance upward to $3 + log_{10}(g)$, where g is the number of unlinked markers. This correction is not necessary if numerous dependent (i.e., linked) markers are tested.

Because of the complexities associated with choosing parameter values, more than one genetic model is often tested in a parametric linkage analysis of a complex genetic disease. Multiple models are tested when different diagnostic categories are used; penetrance, disease allele frequency and/or phenocopy rates are varied; and when more than one mode of inheritance is specified for different analyses. As the number of models is increased the posterior type I error will also increase unless adjusted for statistical significance. The inflation of the posterior false positive rate is greater due to the maximization of lod scores over diagnostic definitions than as a result of varying penetrance (Ott, 1991). In general, the number of models should be limited and chosen *a priori*. It is inappropriate to report only the highest maximum lod score found when multiple models are analyzed. It has been suggested by Risch (1991) that a more conservative lod score should be calculated when multiple models are used: $Z_{max}(\theta) - log_{10}(t)$ where t is the number of models. It may also be prudent to estimate an empirical significance level for the maximized score. A method to determine this significance level, as well as to estimate the power of a given family for linkage analysis, involves computer simulation of pedigree data (see next section).

A less conservative approach to parametric linkage analysis with complex diseases is suggested by Greenberg (1992). He points out that the best description of family data is provided by the combination of parameter values that generate the highest lod score. He recommends that data exploration can be carried out by using different parameter values to find the maximum lod score that can be achieved over the models attempted. According to Greenberg (1992), a reasonable positive lod score generated in this manner is an acceptable indication of possible linkage and presents a hypothesis to be tested further, despite the increased likelihood that the linkage finding is spurious.

1.3.10 Computer simulation

An analysis program based on the LINKAGE programs, called SLINK, has been developed to randomly assign genotypes at two loci in a pedigree under a specified true or expected recombination frequency between the loci. To carry out its computations, the SLINK algorithm (Ott, 1989) implements the conditional probability distribution of the phased genotypes given the phenotypes, where

$$P(g \mid x) = P(g_1 \mid x)P(g_2 \mid g_1,x)P(g_3 \mid g_1,g_2,x) \ldots P(g_n \mid g_1,g_2, \ldots, g_{n-1},x)$$

For each $i = 1 \ldots n$ person, the probability of a specific multi-locus genotype, g_i, is calculated given the phenotypes in the family at the marker and disease loci, x, and the previously inferred genotypes. The first person in the pedigree data file is a parent and is randomly assigned one of his/her possible genotypes; the genotype probabilities are calculated for the next parent in the data file and, again, a genotype is randomly chosen. The genotypic probabilities for the offspring in the pedigree are conditional on the assigned parental genotypes. A genotype is randomly assigned to each offspring. The simulation can be carried out with any combination of known and unknown phenotype data as described by Weeks and Ott (1993). Using the same family, the simulation process is repeated many times, usually at least 1000, to generate many replicates of the pedigree with simulated data.

A simulation can be carried out on a pedigree with incomplete or absent marker typing and known phenotypes at the disease locus. Lod scores can be computed for large numbers of simulated replicates using a specified genetic model at various values of the recombination fraction, theta. At each value of theta, the average lod score arising from the replicated pedigrees is denoted as the expected lod score (ELOD), which should maximize at the theta value under which the simulation was executed. The expected lod score at this recombination fraction represents the informativeness of the given family (Ott, 1991). In this way, the relative usefulness of several available families can be compared prior to collection of blood samples or before genetic typing is initiated. Another valuable statistic is the distribution of the maximum lod score achievable across the replicates (denoted as $Z_{max}sim$); an analysis program will provide the mean of $Z_{max}sim$ and the expected proportion of maximum lod scores greater than a specified critical level.

The percentage of times Z_{max} *sim* is greater than 3 directly estimates of the power of the linkage analysis with the given family (Weeks and Ott, 1993). The simulated pedigrees can also be used to determine the gain in power that can be achieved by marker typing an individual who has not been typed before. This information can be useful in deciding who should be typed next if several family members are available to be typed for the first time.

The significance level for an observed lod score maximized over multiple models is the probability under absence of linkage that a lod score equal to or greater than the observed maximum will be computed. To approximate this probability the marker typing for a large number of pedigree replicates is simulated at a specified recombination frequency of 50% between the marker and the disease locus. When the two loci are unlinked, the generation of genotypes is carried out without using information about the inheritance of the disease trait (Ott, 1991). A more efficient program called SIMULATE (Terwilliger and Ott, 1993a) is available for simulation under the null hypothesis of absence of linkage. For each of the replicates with simulated marker data the same series of genetic models is run as in the non-simulated analysis. The maximum lod score over the various models (Z_{max} *sim*) is noted for each replicate and compared with the maximized score observed in the analysis of the observed data (denoted as Z_{max} *obs*). The proportion of replicates in which the maximum lod score from the simulated analysis is equal to or exceeds the maximum lod score observed approximates the p-value of the observed finding. If the p-value is small, the maximum observed lod score is a rare event in the absence of linkage and is unlikely to have arisen by chance (Ott, 1991).

1.3.11 Study of mycobacterial disease

With the preceding detailed introduction to the linkage analysis method in mind, the results of two linkage analysis studies of human mycobacterial disease will now be reviewed. Both studies used the lod score method and neither found significant evidence for linkage. Families from Desirade Island in the Caribbean, containing at least two members affected with leprosy, were analyzed in the first study (Abel et al., 1989). The study examined the evidence for linkage between the recessive traits of leprosy *per se* (any form) and non-lepromatous leprosy and five marker loci, including Rhesus factor

(Rh), Km and Gm immunoglobulin allotypes, HLA-A and -B haplotype, and ABO blood group.

Depending on which marker locus was analyzed, 7–9 pedigrees were used with a total of 18–32 informative matings. The marker allele frequencies for Rh, Km, Gm, and ABO were estimated from the Desiradian population. The five HLA-A and -B haplotypes were specified to be equally frequent. The genetic models for the disease loci were assigned their maximum likelihood estimates from a segregation analysis of leprosy in Desirade. The leprosy *per se* and non-lepromatous leprosy disease allele frequencies, q, were 0.312 and 0.305, respectively; frequencies one-tenth of these values were also considered. Ten age-specific penetrance classes were defined, with a maximum penetrance of 62% for leprosy *per se* and 50.2% for non-lepromatous leprosy by age 60 for the homozygous higher risk genotype, r/r. An age-specific penetrance reaching 2% by age 60 was assigned to the lower risk R/r and R/R genotypes for the leprosy *per se* trait.

The maximum lod scores achieved with each marker were close to zero for the trait of leprosy *per se*. Close linkage with the disease trait could be excluded for the Rh and Gm loci when q was equal to 0.312 and for the Rh, HLA, ABO, and Gm loci when q was equal to 0.0312 (Table 4). For the non-lepromatous leprosy trait, close linkage was excluded for Rh and HLA with the more rare disease allele frequency of 0.0305 (Table 4). The maximum lod score achieved with non-lepromatous leprosy was 1.02 with ABO at $\theta = 0.0$ and $q = 0.0305$. The positive direction of this result with non-lepromatous leprosy only suggests that genetic factors contributing to the two traits, if present, may exist at different loci. Further investigation of the ABO finding is warranted before any conclusion can be made about the result, which is not statistically significant. Analyses with the Km marker did not have sufficient power to support or reject linkage with either trait. This study, therefore, provided evidence against a single susceptibility gene linked to the Rh, Gm, HLA, or ABO loci for the leprosy *per se* trait in the Caribbean pedigrees.

Linkage analysis with one of the families studied in this thesis was carried out with the trait of primary tuberculosis and the HLA-A and -B haplotype (Miller, 1991).

Table 4. Two-point linkage analysis of leprosy pedigrees, Desirade Island (adapted from Abel et al., 1989)

			Lod score at specified recombination fraction		
Phenotype	Disease gene frequency	Marker	0.0	0.01	0.10
Leprosy *per se*	0.312	Rh	-2.80	-2.59	-1.38
	0.312	Gm	-2.02	-1.85	-0.94
	0.0312	Rh	-7.66	-5.84	-2.31
	0.0312	HLA	-4.80	-3.41	-0.69
	0.0312	ABO	-2.60	-2.04	-0.57
	0.0312	Gm	-9.43	-6.40	-1.96
Non-lepromatous leprosy	0.305	ABO	+0.71	+0.64	+0.44
	0.0305	Rh	-2.78	-1.92	-0.48
	0.0305	HLA	-2.34	-1.66	-0.25
	0.0305	ABO	+1.02	+0.96	+0.70

A 61-member Aboriginal Canadian pedigree with 22 affected family members in four generations was analyzed with four genetic models using the marker haplotype frequencies observed in founders of the pedigree. For all of the models, the disease trait was assumed to be autosomal recessive with a susceptibility allele frequency of 0.025, and the penetrance of the higher risk r/r genotype was set at 100%. Three of the four models were assigned a phenocopy rate of 0%.

The first model excluded linkage up to a recombination frequency of approximately 25%, for which the lod score was -1.96. The second analysis differed from the first only in the phenotype assignment of five affected individuals, for whom clinical status was made unknown. It was thought that these persons might have become affected for non-genetic reasons because three individuals were very young (under 2 years) and two persons developed disease in the 1950s, when they were children and perhaps

experiencing worse living conditions. Therefore, these persons may have been R/R or R/r at the susceptibility locus, because the penetrance or relative risk of disease for the R/R, R/r, and r/r genotypes were approximately the same. The coding of unknown status for these five individuals was maintained for the subsequent analyses. The second analysis excluded linkage up to a recombination frequency of approximately 15%; the lod score at 20% recombination was -1.51.

A non-zero risk of disease for the lower risk R/R and R/r genotypes was added to a third model, so that the proportion of unlinked cases among all cases was 6%. This resulted in a significant loss of power and a flattened lod score curve with a maximum of 0.28 at a recombination frequency of 15%. The fourth model had a phenocopy rate of 0% and assigned all unaffected individuals an unknown clinical status, in order to investigate the linkage information provided by the presumed genetically susceptible individuals who had expressed the disease phenotype. The "affecteds only" analysis yielded a minimum lod score of approximately -3 at a recombination frequency of 0%. The study, therefore, indicated that under these models there was evidence against close linkage of a tuberculosis susceptibility gene with the class I HLA-A and -B haplotype in the pedigree examined.

SECTION 2: STUDY OBJECTIVES

The main objective of the study was to determine whether the putative human homologue of the mouse *Bcg* gene has a role in susceptibility to the development of tuberculous disease following infection with *Mycobacterium tuberculosis*. Linkage analysis with appropriate genetic models and families from three continents was used to test for co-segregation of human chromosome 2q markers and the phenotype of tuberculous disease. The study sample of families represented distinct tuberculosis experiences, namely exposure to a short-term, high intensity epidemic (one Canadian pedigree), and lifetime exposure to high endemic levels of *M. tuberculosis* infection (five Colombian and 13 Hong Kong pedigrees).

This study was designed to demonstrate the contribution arising from combining epidemiological considerations with analytical methods originating from the field of

molecular biology. The complexity of the disease trait had important implications for the study design, materials, and methods. A secondary objective was to examine the advantages and limitations of the study approach, given the features of the disease, the study population, and the available data.

SECTION 3: METHODOLOGY

3.1 Study design

The research question was addressed with a linkage analysis study. This study design incorporated clinical and genetic marker data, as well as epidemiological considerations, to study genetic inheritance at a susceptibility locus as a risk factor for tuberculosis. Genetic inheritance was hypothesized to modify the risk of disease, such that genetically susceptible individuals would have an increased disease risk. The lod score method was used for data analysis. Phenotype data were collected and genotypes were inferred probabilistically using models relating phenotype and genotype. The sampling unit was the family so that persons of known biological relationship were studied, for whom the risk of disease depended on their relationship to others in the pedigree and the genotype inherited at the susceptibility locus. Specific inclusion criteria were used to select the study families, information about non-genetic risk factors for disease was collected for cases and non-cases in a uniform manner, and, in each region, the same diagnostic methods were used for cases and non-cases.

3.2 Family selection and sample collection

The study families originate from three regions: Canada, Colombia, and Hong Kong. Family enrollment, blood sampling, and collection of clinical and pedigree data were the responsibility of investigators in the three regions and, for the most part, were completed prior to this thesis project.

3.2.1 Colombian and Hong Kong families

A total of 18 families were identified by the collaborators in Colombia and Hong Kong; five families were selected in Colombia (coded for confidentiality as families GV, MAM, RAV, RES, and RJA), and 13 were enrolled in Hong Kong (numbered arbitrarily

as families 1 through 13). These families constitute the international study population. The methods of identification of the families were described in a questionnaire which I developed. The objectives of the questionnaire were:

(1) to confirm the family enrollment criteria, data collection methods, and the assignment of clinical phenotype to the family members

(2) to gather descriptive data on the families, including their location of residence and the tuberculosis incidence rate and trend in their region of residence

(3) to ascertain the perceived validity of the classification of the family members as affected or unaffected with tuberculosis

(4) to collect additional clinical and demographic information which was not available when the blood samples were received in Canada (e.g., date of birth, age at diagnosis, date of diagnosis, skin test and BCG vaccination data, chemoprophylaxis information, and details about previous experience with tuberculosis and general living conditions).

The questionnaire, written in English, was mailed to the investigators in Colombia and Hong Kong along with maps of the regions and pedigree diagrams. Appendix A contains a copy of the questionnaire sent to Colombia; the Hong Kong questionnaire is identical in format. The questionnaire was arranged in four sections referring to (A) Collection Methods, (B) Diagnostic Methods, (C) Living Conditions, and (D) Tuberculosis Situation.

Specific criteria were used by the collaborators to select the Colombian and Hong Kong families for inclusion in the linkage analysis study, although a few families were enrolled without meeting these criteria. The criteria were as follows: (1) two or more offspring in the family were affected with tuberculosis; (2) one parent was affected and the other parent was unaffected; and (3) additional relatives may have been affected. Criteria #1 and #3 were proposed to optimize the power of the analysis with the disease trait by obtaining families with several affected individuals. Criterion #2 was specified so that the study families would imitate the experimental mouse backcross study, given a recessive disease model. It was also thought that the presence of at least one affected family member in each of two generations would be an indication that there had been

many *M. tuberculosis* exposure opportunities in the study families. It would thus be more likely that all family members had been infected.

The selection criteria were met by four Colombian families and 10 families from Hong Kong. One of the Colombian families had two unaffected parents (family RJA). Three Hong Kong families had only one affected offspring (families #2, 4, and 12). It was not possible for the investigators in either region to indicate the prevalence of such families in their respective general populations, in order to estimate the proportion of the population represented by the study families.

The selection of the international study families was opportunistic, appropriate families being considered eligible as they came to the attention of the investigators. The Colombian families were ascertained through a search of a registry of tuberculosis patients. Four of the families were made known through affected children, and one was located through the affected parent (family RES). In Hong Kong, field workers' knowledge was used to find families #7 and 13, a search of a registry located families #1, 2, 3, and 10, and the remaining families were found during contact tracing of family members following a tuberculosis diagnosis in one or several relatives. Seven of the Hong Kong families were located through affected parents (#1, 2, 4, 7, 9, 10, and 12), four through affected children (#3, 6, 8, and 13), and two through the knowledge of both affected parents and affected children in the family (#5 and 11).

Collection of blood samples from families in Colombia took place at hospitals, local health clinics, or at home by two physicians and one technician, between February 1987 and July 1990. In Hong Kong, sample collection was carried out at local health clinics by a nurse and a technician, between February 1988 and January 1992. Pedigree structures for the Colombian and Hong Kong families were provided by the international investigators. Diagrams of these pedigrees are presented in Appendices B1 and B2.

3.2.2 Canadian family

In 1989, a large Aboriginal Canadian family was located for a study of diagnostic tests for tuberculosis and linkage analysis with HLA-A, -B, and -C typing (Miller, 1991). Most of the individuals were known to have been infected with *M. tuberculosis* because of a tuberculosis outbreak, and many were subsequently diagnosed with the disease.

Several characteristics of the family made it an ideal choice for inclusion in the previous study and the present thesis: the family had had uniform access to a single diagnostic and treatment facility; there were well-documented laboratory and clinical data; family members from at least two generations participated in the study; and, there had been at least a 3 month interval between exposure to *M. tuberculosis* and medical intervention in the family (Miller, 1991). For the present study, DNA was available on most of the individuals included in the HLA analysis by Miller (1991). The pedigree structure indicated by the contact family member during the fieldwork in 1989 is displayed in Appendix B3. A description of the tuberculosis experience of this family is presented below, based on a published report of the outbreak (Mah and Fanning, 1991) and information collected by Dr. Miller (Miller, 1991) and myself.

The Canadian family lived in a relatively isolated, rural Aboriginal community and was exposed to an epidemic of tuberculosis beginning in 1987 and ending in 1989. The population of the community was about 350 persons, including seven to eight large families related by marriage. Specific demographic data were not available; however, about 50% of the population were children. Cases of tuberculosis were common in the community in the 1950s. A few members of the family were diagnosed with tuberculosis during this time (see Appendix B3). The most recent smear-positive case of pulmonary tuberculosis diagnosed prior to the 1987 epidemic was identified in 1959. Routine BCG vaccination (likely in school) was not performed after 1963 and was non-uniform before that time. Prior to the epidemic, only one case of tuberculosis had been diagnosed in the population since 1969. Skin testing of adults and children ceased in 1981 and 1985, respectively.

The 1987–1989 outbreak was a "single point source epidemic" (Mah and Fanning, 1991). The 25 year-old mother of the infant index case was identified as the source case. From October to December 1986, the pregnant woman lived in the rural community with an elderly, ill grandparent. The young woman left the community in January 1987 to live with a sibling in Edmonton and delivered a child in February. Over the next five months the mother experienced symptoms of tuberculosis, including productive cough, fever, night sweats, occasional haemoptysis, and 22 kg of weight loss. From February

to August 1987 the mother frequently spent weekends in the family's rural community visiting with her 17 siblings and their children, who lived in several dwellings a short walking distance apart. The housing, which did not have indoor plumbing, was heated with wood stoves and ventilated in summer with open doors and windows. There was frequent close contact between members of the large family in the community during this time (Fanning, 1993, personal communication).

The child born in February 1987 was found to be tuberculin skin test-positive with an abnormal chest X-ray consistent with tuberculosis at the age of 4 months. The mother was found to be smear- and culture-positive on August 5, 1987 and was transferred with her child to Edmonton for treatment. Tuberculin skin testing of known family contacts and community members without previous positive results, as well as diagnostic testing (chest X-ray, specimen culture and smear) of all persons suspected as infected, was initiated in August 1987. The time-line in Figure 4 shows the week of diagnosis of the cases in the Canadian family, peaked in August 1987. Individuals diagnosed with tuberculosis were sent to Edmonton for treatment in groups; thus, exposure to infection varied (Fanning, 1993, personal communication). All members of the community with a skin test reaction of ≥ 10 mm of induration at 48–72 hours and close household contacts of cases were recommended isoniazid chemoprophylaxis in August 1987.

In the following description of the outbreak severity, persons considered infected during the epidemic were those diagnosed with tuberculosis and those with newly positive skin test (PPD) results. For the 228 assessed community persons and the 66 family members whose medical files were investigated by Dr. Miller, I estimated the infection rate during the epidemic as 21% (62/294), with 28 persons diagnosed with tuberculosis (24 in the study family and four in the community), and 25 family members and nine other community members with newly positive skin tests. Among the study family members I estimated the infection rate as 74% (49/66). The infected proportion did not include the following individuals despite their presence in the community during the epidemic: three previous cases and six previously positive PPD reactors, all of whom were not skin tested during the outbreak, and eight unaffected family members with negative PPD results in the outbreak investigation. The infected proportion also did not

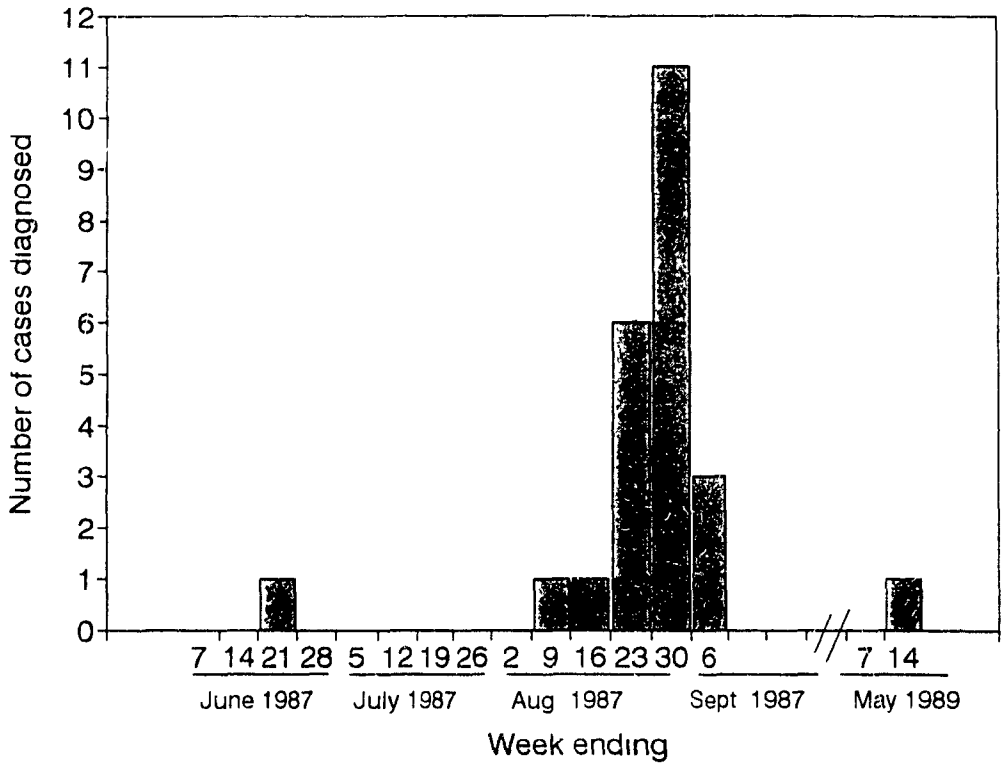


Figure 4. Time-line for diagnosis of cases in the Canadian family, 1987-1989

Notes.
The date of diagnosis was the date of a positive culture result or an abnormal chest X-ray result

source: Alberta Health Tuberculosis Services, Edmonton (compiled by Dr. Miller)

include the affected great-grandparent in the Canadian family who was diagnosed with tuberculosis both prior to and during the outbreak (see Appendix B3), since the disease during the outbreak may have been the result of reactivation of the previous infection.

I estimated the community attack rate, defined as the number of persons diagnosed with tuberculosis per number of persons investigated, as 9.5% when the study family was included in this calculation (28/294). Using the same rate definition, the family attack rate was 36% (24/66). There were four individuals diagnosed with tuberculosis in the community outside the study family, providing an attack rate among non-family community members of 1.8% (4/228). The differences in the infection and disease rates for the community when compared to the study family alone were consistent with the concentration of exposure to *M. tuberculosis* among the family members during the outbreak.

3.3 Laboratory analysis

The preparation of DNA and the genetic marker typing were completed by persons in the international laboratories and in the laboratories of Dr. Schurr and Dr. Morgan in Montreal. In Colombia, the time frame between blood collection to sample processing (setting up of cell lines and DNA extraction) in the laboratory was under 2 hours. All of the Hong Kong samples were taken in the morning at one of four central clinics and bloods arrived at the laboratory within 1–2 hours. In the Hong Kong laboratory, white blood cells were put in culture and Epstein Barr Virus- and Cyclosporin A- treated within 2 hours, so that an immortalized cell line was produced for each sample. The time range from sampling to end of culture set-up was 3–4 hours.

For the study by Miller (1991), each available member of the Canadian family provided 7–10 mL of heparinized blood and an amount of lymphocytes was set aside for the HLA-A, -B, and -C typing before transformation. Mononuclear cells, kept in culture media, were transported from Edmonton to Montreal. The Hong Kong samples were received in the laboratory in Montreal as frozen cell lines. A portion of the Colombian blood samples were processed and arrived as frozen cell lines or as extracted DNA; the rest of the Colombian samples were received as whole blood.

Figure 5 displays the process of sample preparation from blood to cell lines and DNA, carried out in the international laboratories and in the laboratory at the Montreal General Hospital Research Institute. The mononuclear cells are isolated from the available bloods and transformed with Epstein-Barr virus treatment so that an unlimited supply of DNA is, in principle, available for each family member bled. The immortalized cell lines are frozen, usually in several aliquots. When genetic analysis is required, cells are grown in an unfrozen aliquot, and DNA is extracted from the grown cells. The remaining cell line quantity is frozen for storage. The growing procedure can be repeated when more DNA is required for genetic testing.

Locus, chromosome location, probe, and alleles for each marker used are presented in Tables 5a and 5b (RFLP markers) and Table 6 (microsatellite markers). Typing was completed for the following chromosome 2q RFLP markers using the Southern blotting technique (reviewed by Gelehrter and Collins, 1990) (restriction enzyme used in parentheses): CRYG1-D (*Hind*III), CRYGP1-A (*Taq*I), CRYGP1-B (*Taq*I), CRYGP1-C (*Taq*I), DESMIN (*Eco*RV), FN61039-H (*Hae*III), FN61039-M (*Msp*I), INHB-A1 (*Taq*I), TNP-1 A (*Taq*I), TNP-1 B (*Taq*I), VIL E-62 (*Bgl*II), VIL E-84 (*Taq*I), VIL pEX2(a) (*Msp*I), and VIL 7 No.6 (*Bgl*II). At the time the typing was carried out for these markers, a human genetic map had not been published with these loci so that the distance between the markers along chromosome 2q was not known.

In addition to the markers listed above, an RFLP of the TNP-1 locus, designated as TNP-1 C and using the *Msp*I restriction enzyme, was used to type all the study families with the Southern blotting technique. A second RFLP marker of the same locus could be assayed using the Polymerase Chain Reaction and *Msp*I restriction enzyme digest. It gave the same results as TNP-1 C when the Canadian and Hong Kong families were typed. It was not clear from published sources how near to each other the polymorphic sites were. Because of the identical typing, locus, and enzyme used, the two markers will be considered the same and the name TNP-1 C will be used to denote them. However, for the purposes of Table 5b only, the polymorphism detected by Southern blotting is designated as TNP-1 C (a) while that analyzed by the Polymerase Chain Reaction is designated as TNP-1 C (b).

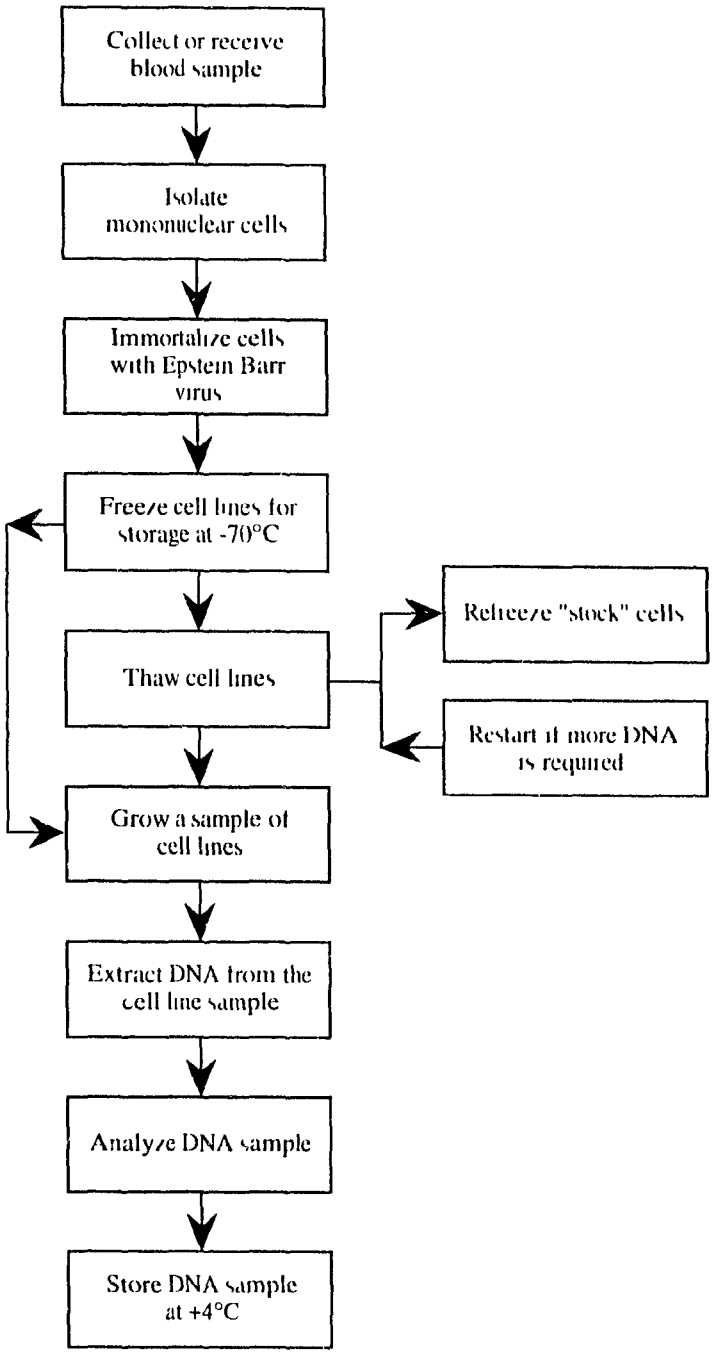


Figure 5 The process of sample manipulation from blood to cell lines, through to stored DNA

Table 5a. RFLP markers

Marker	Locus	Location	Probe	Restriction enzyme used to detect the polymorphism	Number of alleles	Size of allele diagnostic bands (kb*)
CRYG1-D[1]	CRYG1	2q33-q35	p16G3	*Hind*III	3	13; 11.5; 8.4
CRYGP1-A[2]	CRYGP1	2q33-q35	p5G1 (a)	*Taq*I	2	3.5; 3.3
CRYGP1-B[2]	CRYGP1	2q33-q35	p5G1 (b)	*Taq*I	2	2.2; 1.1
CRYGP1-C[2]	CRYGP1	2q33-q35	p5G1 (c)	*Taq*I	2	2.0; 1.4 & 0.6
DESMIN	DES	2q35	bluescript[3]	*Eco*RV**	2	12; 8
FN61039-H[4]	FN-1	2q34-q36	FN61039	*Hae*III	2	2.0; 1.5
FN61039-M[5]	FN-1	2q34-q36	FN61039	*Msp*I	2	2.7; 0.35

Notes.

*kb = kilobase pairs

[1]Willard, H.F., Meakin, S.O., Tsui, L.-C., and Breitman, M.L., 1985. Assignment of human gamma crystallin multigene family to chromosome 2. Somatic Cell and Molecular Genetics 11 (5): 511-516.

[2]O'Connell, P., Lathrop, G.M., Nakamura, Y., Leppert, M.L., Lalouel, J.-M., and White, R., 1989. Twenty loci form a continuous linkage map of markers for human chromosome 2. Genomics 5: 738-745.

[3]Li, Z , Lilienbaum, A., Butler-Browne, G., and Paulin, D., 1989. Human desmin-coding gene: complete nucleotide sequence, characterization and regulation of expression during myogenesis and development. Gene 78: 243-254.

[4]Colombi, M., Gardella, R., Barlati, S., Vaheri, A., 1987. A frequent *Hae*III RFLP of the human fibronectin gene. Nucleic Acids Research 15 (16): 6761.

[5]Gardella, R., Colombi, M., and Barlati, S., 1988. A common *Msp*I RFLP of the human fibronectin gene (FN1). Nucleic Acids Research 16 (4): 1651.

**The polymorphism was identified by collaborators in the laboratory at the Montreal General Hospital Research Institute (Liu, J. and Schurr, E.).

Table 5b. RFLP markers (continued)

Marker	Locus	Location	Probe	Restriction enzyme used to detect the polymorphism	Number of alleles	Size of allele diagnostic bands (kb*)
INHB-A1	INHB	2q33-qter	INHBA1†	*Taq*I†	2	2.6, 0.5
TNP-1 A[6]	TNP-1	2q34	TNP1.Clone 1	*Taq*I	2	15.25; 4.39
TNP-1 B[6]	TNP-1	2q34	TNP1.Clone 1	*Taq*I	2	8.5; 6.6
TNP-1 C (a)[6]	TNP-1	2q34	TNP1.Clone 1	*Msp*I	2	7.36; 6.43
TNP-1 C (b)[7]	TNP-1	2q34	TN1	*Msp*I	2	1.1; 0.4 & 0.7
VIL E-62	VIL	2q35-q36	E-62[8]	*Bgl*II†	2	18, 5
VIL E-84	VIL	2q35-q36	E-84[8]	*Taq*I†	2	1.9; 1.0
VIL pEX2(a)[9]	VIL	2q35-q36	pEX2(a)	*Msp*I	2	8; 6
VIL 7 No.6	VIL	2q35-q36	JL 7 No. 6†	*Bgl*II†	2	9.4; 6

Notes.

*kb = kilobase pairs

[6]Liu, J., Kim, H., Engel, W., and Schurr, E., unpublished observations.

[7]Hoth, C.F. and Engel, W., 1991. Two RFLPs at the TNP1 locus. Nucleic Acids Research 19 (24): 6979.

[8]Vidal, S.M., Malo, D., Vogan, K., Skamene, E., and Gros, P., 1993. Natural resistance to infection with intracellular parasites: isolation of a candidate for *Bcg*. Cell 73: 469-485.

[9]Pringault, E., Arpin, M., Garcia, A., Finidor, J., and Louvard, D., 1986. A human villin cDNA clone to investigate the differentiation of intestinal and kidney cells *in vivo* and in cell culture. EMBO Journal 5: 3119-1324.

†The polymorphism was identified or the probe was cloned by collaborators in the laboratory at the Montreal General Hospital Research Institute (Liu, J. and Schurr, E., 1991 and 1992).

Table 6. Microsatellite markers

Marker name and locus	Location	Number of alleles	Allele sizes (bp*)
CRYG1-A[1]	2q33-q35	3	176-185
CRYG1-B[2]	2q34-q35	3	118-127
D2S102[3]	2	6	140-162
D2S104[3]	2	5	114-132
D2S120[4]	2	3	159-179
D2S125[4]	2	6	87-101
D2S126[3]	2	4	140-154
D2S128[4]	2	6	150-166
D2S137[4]	2	7	138-158
D2S140[4]	2	5	155-163
D2S157[4]	2	7	106-120
D2S172[4]	2	10	254-296
D2S173[4]	2	4	113-121
D2S206[4]	2	4	113-125
D2S211[5]	2q34-q37	7	141-161
PAX-3[6]	2q35-q37	8	306-336

Notes.

*bp = base pairs

[1]Hearne, C.M. and Todd, J.A., 1993. Trinucleotide repeat polymorphism at the CRYG1 locus. Nucleic Acids Research 19 (19): 5450.

[2]Polymeropoulos, M.H., Xiao, H., Rath, D.S., and Merril, C.R., 1993. Trinucleotide repeat polymorphism at the human gamma-B-crystallin gene. Nucleic Acids Research 19 (16): 4571.

[3]NIH/CEPH Collaborative Mapping Group, 1992. A comprehensive genetic linkage map of the human genome. Science 258: 67-86.

[4]Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morisette, J., Millasseau, P., Vaysseix, G., and Lathrop, M., 1992. A second generation linkage map of the human genome. Nature 359: 794-801.

[5]Barber, T.D., Morell, R., Johnson, D.L., Asher, J.H.Jr., and Friedman, T.B., 1993. A highly informative dinucleotide repeat polymorphism at the D2S211 locus linked to ALPP, FN1 and TNP1. Human Molecular Genetics 2 (1): 88.

[6]Wilcox, E.R., Rivolta, M.N., Ploplis, B., Potterf, S.B., and Fex, J., 1992. The PAX3 gene is mapped to human chromosome 2 together with a highly informative CA dinucleotide repeat. Human Molecular Genetics 1 (3): 215.

In addition, the Canadian family was typed for the following chromosome 2 microsatellite markers using the Polymerase Chain Reaction (Hearne et al., 1992): CRYG1-A, CRYG1-B, D2S102, D2S104, D2S120, D2S125, D2S126, D2S128, D2S137, D2S140, D2S157, D2S172, D2S173, D2S206, D2S211, and a microsatellite closely linked to the PAX-3 gene (denoted simply as PAX-3) (Table 6). Results for the D2S211 marker were difficult to read and were thus considered unreliable and will not be presented here. The CRYG1-A and CRYG1-B markers typed for the same polymorphism in the human gamma-B crystallin gene and thus yielded the same pattern of typing. Therefore, only the CRYG1-A marker was used for linkage analysis.

Most of the microsatellite markers were anonymous, that is, they did not detect polymorphisms in genes of known encoding function, with the exception of CRYG1-A. As shown in Table 6, the location for most of the microsatellite markers was not yet well defined. The relative order of these markers was not completely known, because different research groups had published linkage maps for different subsets of the markers. There was likewise a lack of published information which would allow an integration of the microsatellite markers with the RFLP markers tested in this study; only markers of the CRYG1 locus appear in both Tables 5a and 6.

3.4 Clinical phenotype

Assignment of clinical status to the family members was carried out by specialists in the three regions, blinded to the marker data because diagnoses were completed prior to genetic analysis. The phenotype definitions below are presented separately for the international and the Canadian families. More information was available for the Canadian family members who were assigned phenotypes in Alberta and the clinical definitions related to a well-defined outbreak situation, with a short interval between exposure and disease development, and a few previous cases. The "affected" status in the Canadian family referred to a diagnosis of primary tuberculosis. Cases of tuberculosis in the international families were diagnosed over a period of thirty years with most diagnoses in the 1980s, some individuals were diagnosed with the disease more than once, and there was an indeterminate interval between infection and development of disease.

The clinical definitions used for the Colombian families were determined from the completed questionnaire. Individuals were either classified as affected due to a diagnosis of tuberculosis in the past (based on clinical history), or were defined as affected on the basis of a positive culture and/or sputum (or other fluid) smear, with or without tuberculosis symptoms (see Appendix B5). Culture of biological specimens was rarely performed; if an individual was culture positive, however, he/she was classified as affected despite negative smears and/or normal chest X-ray results. When a chest X-ray was performed in the absence of a specimen culture, an abnormal result was used to define the individual as affected, even if the results of a smear were negative. In two cases the clinical records of individuals classified as affected were lost so that the method of diagnosing these persons could not be confirmed. Unaffected persons had negative results on all tests, no tuberculosis symptoms, and no history of a diagnosis of the disease.

Based on the completed questionnaires, affected status for a Hong Kong family member was similarly defined as any diagnosis of tuberculosis in the past or at least one positive result on culture, smear, or chest X-ray, with or without tuberculosis symptoms (see Appendix B5). In the absence of culture, smear, or chest X-ray tests, an individual with tuberculosis symptoms was also considered affected. When no diagnostic tests were done, individuals were classified as unaffected if symptoms were absent and they had no history of a diagnosis of the disease.

I reviewed the assignments of infection and disease status for the Canadian family members, which were completed by Dr. Miller in 1989, with the assistance of Dr. Fanning (Miller, 1991). Dr. Miller recorded and interpreted all laboratory and clinical data collected during the 24-month outbreak period for each study person. The criteria used for assignment of infection and disease status were consistent with standards set by the American Thoracic Society. A case met the following criteria: isolation of *M. tuberculosis* from a respiratory or gastric specimen and presence of at least one symptom indicative of active tuberculosis. The symptoms included weight loss, fatigue, night sweats, cough, or a pulmonary infiltrate which was non-responsive to anti-bacterial drugs. The criteria for a culture-negative case were the following: a negative culture result, presence of at least one tuberculosis symptom, and positive clinical response to

anti-tuberculosis therapy. An old case was someone with a tuberculosis diagnosis in the past or a previously positive clinical response to anti-tuberculosis treatment, prior to the 1987–1989 epidemic (see Appendix B5). Non-cases did not meet any of the three definitions above and were, therefore, considered free of disease. Infected individuals were those with a skin response of ≥ 5 mm of induration at 48 or 72 hours following administration of a 5 TU tuberculin skin test (Tubersol®, Connaught Laboratories Ltd.) or previous or current diagnosis of tuberculosis. PPD-negative persons were defined as those with a < 5 mm reaction at 48 or 72 hours following administration of the tuberculin skin test. For the present study, assignments of disease status for the Canadian family members were accepted as determined by Miller (1991).

3.5 Data entry and verification

3.5.1 Genetic marker data

All RFLP data was verified by Dr. Schurr or Ms. Fujiwara. The data was re-checked by Ms. Liu or Ms. Buu and myself, with the exception of the typing for markers at the DES and VIL loci in the Colombian families and Hong Kong families #6-13, DESMIN and VIL pEX2(a) marker typing in the Canadian family, and typing at the CRYG1-D and the CRYGP1-A,B and C loci in all families. All microsatellite marker results were checked by Ms. Paradis, Ms. Fujiwara, and myself. All typing results on films and data sheets were checked against the data entered in the database, from which text files were exported for linkage analysis. Any data added directly to text files was meticulously compared to data sheets for accuracy. All microsatellite marker typing had to be re-coded for use with the analysis programs so that alleles were numbered from 1 ... n. In most cases, the largest allele observed was numbered 1, the next largest 2, and so on until the nth allele. The re-coding was done to the data files for linkage analysis.

On occasion, a typing result for an offspring was incompatible with that of his/her indicated parent(s) (e.g., the offspring was typed as a 1/2 and both parents were typed as 1/1 individuals). In these cases the films concerned were re-read; if any typing was not clear enough for definitive reading the individuals in question were re-typed. Repeatedly inconsistent typing that was clear-cut on the films was taken to indicate either sample mix-up, non-parentage, or mutation of microsatellites (Weber and Wong, 1993).

3.5.2 Affection status

The phenotype at the susceptibility locus was designated as an affection status in the analysis programs. No changes were made to the phenotype assignment of the Canadian family members as indicated in the project database by Dr. Miller. Information presented in the questionnaires was used to confirm the affected or unaffected status of members of the Colombian and Hong Kong families. The Colombian questionnaire confirmed the clinical status of family members in the original pedigree diagrams for all but two individuals. One family member was reported as smear positive with an abnormal chest X-ray, and his/her initially unaffected status was changed to affected in the analysis files. Although the clinical records for the second family member were not available, he/she was extensively investigated after the family was enrolled in the study and determined to be unaffected (Garcia, 1994, personal communication), and the necessary status change was made in the analysis files.

Based on the medical record information in the Hong Kong questionnaires, changes were made to the clinical status originally recorded for three individuals. In one case, an individual originally indicated as affected was described in the questionnaire in the following manner: smear and culture negative, no tuberculosis symptoms shown, and chest X-ray not determinable or ambiguous. The clinical status of this individual was changed to unaffected in the analysis files. In two cases, individuals originally indicated as unaffected clearly met the criteria for affected status on the basis of the questionnaire documentation. The two persons were described as smear and culture negative with abnormal X-rays and display of tuberculosis symptoms. Clinical status for both individuals was changed to affected in the analysis files.

3.5.3 Pedigree structure and sample identification

To validate the pedigree information in Appendices B1-B3 and to ensure quality control of the DNA samples, genetic typing was used to indicate possible parentage misidentification or sample mix-up. The typing for offspring in the study families was examined for genetic consistency with parental typing, given the pedigree diagrams. Two or more marker inconsistencies in a particular offspring, for which at least one inconsistency involved an RFLP marker, were taken to indicate either (1) misidentification

of the parentage of the offspring or, (2) mix-up of the offspring sample. In these cases the validity of the parental typing was accepted, since other results were consistent with the identified parentage. If inconsistencies were found for an individual who was both a parent and an offspring in the pedigree diagrams, then the typing for the person and his/her children was considered unreliable.

Inconsistent typing was found for several RFLP markers in offspring in two Colombian families. Four offspring were thus excluded from the linkage analysis because re-sampling could not be done. One Colombian family (RJA) was completely excluded from the analysis because of problems with sample identification. Several polymorphic chromosome 12 markers were used to confirm parent-offspring relationships and sample identification in the Hong Kong families, including variable number of tandem repeat (VNTR) markers with a four-base pair (Goltsov et al., 1993) and 30 base-pair tandem repeat (Goltsov et al., 1992) in the phenylalanine hydroxylase gene, and a four-base pair tandem repeat in the parathyroid hormone-related peptide gene (Pausova et al., 1993). The Hong Kong samples were also typed with the D2S128 chromosome 2 microsatellite marker and were tested with a marker located on the X and Y chromosomes for consistency with stated gender (Bailey et al., 1992). On the basis of the typing with several markers including an RFLP marker, inconsistencies were found in two samples from Hong Kong family #2. Because of the stated relationships of the two persons with the other family members, the entire family was excluded from the linkage analysis.

The study contact for the Canadian family was sent a set of questions and a pedigree diagram in August 1993 to re-confirm the pedigree structure. The pedigree information and sample identification for members of the Canadian family were verified with the HLA serological typing (Terasaki et al., 1973) and the polymorphic microsatellite marker typing, given the indicated parent-offspring relationships. A modification of the pedigree was required for analysis because of a sample inconsistency identified by marker typing. All of the Canadian samples used for genetic analysis were also tested with the X and Y chromosome marker (Bailey et al., 1992) for consistency with indicated gender.

Figure 6 presents a summary of the number of families and persons enrolled in the study; exclusions from analysis because of detectable inconsistencies or additions to

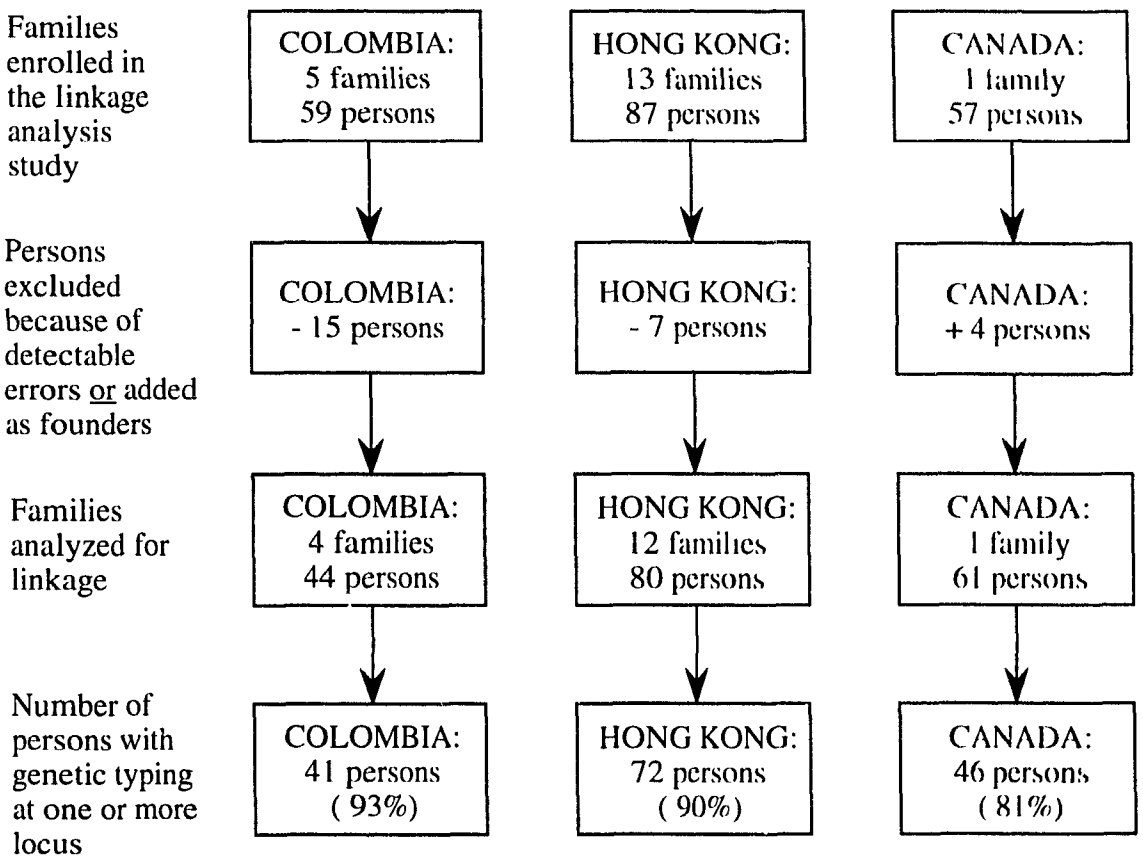


Figure 6. The number of families and persons in the study

pedigree as founders; those analyzed for linkage; and, the number of persons with typing at one or more marker loci. A total of 19 families were enrolled by the international and Canadian investigators. For the Canadian family, 57 of the 66 potential participants were included in the present linkage analysis study. Because of the marker typing inconsistencies described above, some persons were excluded so that, of the international data set, four Colombian families (44 persons) and 12 Hong Kong families (80 persons) were analyzed for linkage. Four founder individuals of unknown clinical status were added to the Canadian linkage analysis pedigree to complete its structure, so that this pedigree contained a total of 61 individuals.

Figure 6 also presents the number of family members with genetic typing at one or more marker loci from each geographic region. Not all persons were typed because some individuals were deceased or otherwise unavailable; in addition, some samples were depleted prior to the present study or DNA could not be isolated from the blood. Therefore, for some markers there are more typing observations--more people typed--than for others. The persons for whom there is no genetic typing data are indicated in the pedigree diagrams in Appendices B1-B3 with an asterix. In the linkage analysis pedigrees, the proportion of persons with known clinical status who were typed at one or more marker loci was 91% (41/44) for the Colombian families, 90% (72/80) for the Hong Kong families, and 81% (46/57) for the Canadian family.

3.6 Computer programs and statistical tests

All pedigree, clinical status, and genetic marker data were entered into a Double Helix® database, version 3.5r9, on an Apple Macintosh computer. Pedigree diagrams were drawn by myself with MacDraw™ Pro 1.5v1 and Pedigree/Draw version 4.0 in the Macintosh operating system and by Dr. Morgan with Pedpack version 3.0 (Alun, 1991) for the UNIX operating system. Text files in the form required for the LINKAGE programs were generated from the database and translated with Maclink Plus version 7.5 for use on the DOS machine. Linkage analysis was carried out using the LINKAGE and SLINK programs, DOS version 5.1. Selected linkage analysis with the Canadian pedigree was carried out by Dr. Morgan with the LINKAGE and MENDEL programs (Lange et al., 1988) for the UNIX operating system.

All members of the Canadian family with known clinical status were investigated for tuberculosis during the same well-defined period of time (approximately 24 months), and their clinical records were checked for prior diagnoses of primary tuberculosis. Therefore, all of those classified as non-cases were investigated for disease during the same period of time, in contrast to the situation with the non-cases in the international families. The student t-test was used to test the significance of a difference in mean age for cases (at the time of diagnosis of primary disease) and non-cases (at the time of investigation for disease) in the Canadian linkage analysis pedigree. The z test was used to test the significance of a difference in proportion of females among the same cases and non-cases. A p-value of 0.05 or less was accepted as evidence for a statistically significant difference.

3.7 Linkage analysis

3.7.1 Preliminary models

3.7.1.1 Mode of inheritance of the disease trait

For the linkage analysis in this study, the genetic model of the inheritance of the disease trait was autosomal recessive, consistent with the mouse *Bcg* model, where "r" was the susceptibility allele conferring an increased disease risk. Recessive inheritance was supported for the non-lepromatous leprosy trait following *M. leprae* infection in humans on the basis of segregation analysis performed by Abel and Demenais (1988) on the Desirade leprosy population. In the mouse, recessive susceptibility to early growth of *M. leprae* is also under the control of the *Bcg* gene. Because non-genetic cases of tuberculosis were incorporated into analysis described later in section 3.7, the r/r genotype at "BCG" is referred to as the higher risk genotype while the R/r and R/R genotypes are the lower risk genotypes.

3.7.1.2 Susceptibility allele frequency

Much consideration was involved in the estimation of a susceptibility allele frequency, q. The availability of applicable data was limited. The prevalence of tuberculosis in seemingly suitable populations, such as Aboriginal Canadian, South American, or Asian peoples, could not be used for estimation of the frequency. Because of the poor living conditions experienced by many people in these regions, including

overcrowding and inadequate nutrition, non-genetic cases likely account for a large, unknown proportion of the prevalent cases. The estimation of the total disease prevalence alone would be difficult.

Of the linkage analysis literature studied, the robust recessive model proposed in an article by Risch et al. (1989) implemented a disease allele frequency of 0.3, which translated to a higher risk (r/r) genotype frequency of 9% in the population, under the assumptions of Hardy-Weinberg equilibrium, complete penetrance of the r/r genotype, and no phenocopies. Given the history of tuberculosis experience in the populations from which the study families originated, natural selection may have decreased the susceptibility allele frequency to a relatively low level (that is, less than 0.50). Therefore, a frequency of 0.3 was considered the upper bound of the susceptibility allele frequency for this study.

All the study families were chosen because of having multiple affected family members. This ascertainment method increased the apparent frequency of the higher risk (r/r) genotype in the study population, under a recessive model with no phenocopies. Information about community members outside the Canadian family was used in the estimation of the susceptibility allele frequency. In the community, four individuals were diagnosed with tuberculosis out of 228 persons assessed, resulting in a disease prevalence of 4/228. If the disease trait was assumed to be due only to a single, completely penetrant, recessive gene, then $q^2 = 4/228$ or 0.0175, and an estimate of the frequency of q was 0.13. Given this data and the considerations above, a susceptibility allele frequency of 0.2 was chosen, which implied a higher risk (r/r) genotype frequency of 4%, under the Hardy-Weinberg equilibrium assumption. This susceptibility allele frequency was used in the analysis of all the study families, and implied a disease prevalence, under complete penetrance, of 4 affected persons per 100 individuals.

3.7.1.3 Penetrance

For a genetic disorder with no phenocopies, the penetrance for the susceptible genotype can be estimated with the disease concordance in monozygous twins. For the tuberculosis trait, the monozygous twin concordance is a less appropriate estimate because an unknown proportion of cases may be non-genetic in origin, particularly if

environmental conditions favour high levels of bacterial exposure and inadequate immune status. Of the twin studies available, the Kallmann and Reisner (1943) investigation conducted in New York State provided data to estimate penetrance, assuming only genetic cases occurred. The monozygous concordance rate of 62%, not adjusted for differences in age distribution, was used as the estimate of the probability of disease given the higher risk (r/r) genotype for initial analysis of all the study families. In the preliminary model, the phenocopy rate was assumed to be zero. Penetrances were specified as in the table below, where the abnormal susceptibility allele is designated as r.

Penetrance specifications in the preliminary model (no phenocopies)

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.0	0.0	0.62
unaffected	1.0	1.0	0.38

The phenotype at the susceptibility locus was coded as an affection status for the linkage analysis. Affected individuals were assigned the code "2", unaffected persons were coded as "1", and those of unknown affection status were given the code "0". The three persons in the Canadian family diagnosed with disease prior to the outbreak were considered affected with primary disease. Lod scores at specified recombination fractions of 0.0, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4 were computed with the MLINK program for all analyses, unless otherwise stated.

The penetrances specified in the preliminary model were modified to allow for phenocopies in a second model, which is termed the preliminary phenocopy model. A 50% phenocopy rate in the population was incorporated in order to see the effect of an appreciable proportion of non-genetic cases on the lod scores for linkage between disease and each marker. The following derivation is specific to the particular rate chosen and shows the relation between the penetrances of the less susceptible genotypes (R/R and R/r) and the most susceptible genotype (r/r), and the susceptibility allele frequency.

With a 50% phenocopy rate, the number of affected individuals who have the lower risk genotypes is equal to the number who have the higher risk genotype at the susceptibility locus. Assuming that the risks of disease for the R/R and R/r genotypes are the same then

$$f_x p^2 + f_x 2pq = f_y q^2$$

where f_x is the penetrance of the lower risk R/R and R/r genotypes, f_y is the penetrance of the r/r genotype, and p and q are the frequencies of the normal and abnormal alleles, respectively. Under the assumption of Hardy-Weinberg equilibrium in the population, where $p^2 + 2pq + q^2 = 1$ or $p^2 + 2pq = 1 - q^2$ the above equation becomes

$$f_x (1 - q^2) = f_y q^2 \text{ and}$$

$$\frac{f_y}{f_x} = \frac{1 - q^2}{q^2}$$

where q^2 is the expected proportion of the higher risk genotype and $(1 - q^2)$ is the expected proportion of the lower risk genotypes. The disease risk ratio of the higher risk relative to the lower risk genotypes, f_y / f_x, is the relative risk of a susceptibility gene which is neither necessary nor sufficient for disease, and is equal to the probability of disease given the inheritance of an abnormal susceptibility allele versus the probability of disease without the inheritance of a susceptibility allele (Greenberg, 1993). This relative risk is expressed as a function of the frequency of the abnormal allele. For $q = 0.20$ and $f_y = 0.62$ as before and solving for the penetrance of the lower risk genotypes

$$f_x = \frac{f_y (0.2^2)}{1 - (0.2^2)}$$

or $$f_x = \frac{f_y}{24}$$

The resulting penetrances for the preliminary phenocopy model are given in the following table.

Penetrance specifications in the preliminary phenocopy model

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.026	0.026	0.62
unaffected	0.974	0.974	0.38

3.7.1.4 Assumptions about the disease phenotype

The preliminary models made the assumption that all family members in the linkage study were infected with *M. tuberculosis*. Despite the availability of very little tuberculin testing information, this assumption was likely appropriate for the Colombian and Hong Kong families who were exposed from birth to high endemic rates of tuberculosis. In fact, tuberculin skin testing is not performed in Hong Kong because there is a policy to vaccinate the entire population, and environmental exposure to *M. tb* is likely high (Dr. Higgins, 1993, personal communication).

It was indicated in the Colombian questionnaire that all unaffected family members were living with infectious relatives at the time the tuberculosis diagnoses were made. At the same time, the incidence of tuberculosis was reported to be fairly high in the areas where the families lived, ranging from 22/100,000 to 82/100,000 population. Based on visits to the homes, investigators indicated that there was high potential for airborne transmission in the five Colombian households, which contained 8–15 persons each.

According to the Hong Kong questionnaires, all but three unaffected persons in the 13 study families were living with infectious family members at the time of the diagnoses. For five families, ventilation of housing, which contained 1–2 sleeping rooms for 6–9 occupants, was judged by a home visit to be "poor" to "fair". For seven families, a high potential for airborne transmission was reported. In the regions of residence of the Hong Kong families at the time of the diagnoses the reported tuberculosis incidence was very high, ranging from 116/100,000 to 223/100,000 population.

For the Canadian linkage analysis pedigree, the assumption that all persons were infected during the outbreak was likely acceptable for the purposes of a preliminary model, given the large number of infectious cases and the frequent contact between members. All pedigree members were either present in the community during the outbreak or lived with the source case elsewhere. Positive tuberculin skin tests were obtained during the outbreak for all but five of the members who had not previously been tested positive. The five PPD-negative non-cases, aged 32, 12, 9, 4, and 3 years, presented with 0 mm of induration following skin testing during the outbreak. A child diagnosed as a culture-negative case, aged 1 year, was also PPD-negative in August 1987 and January 1988, possibly due to severe tuberculous disease. The negative results in the non-cases could have been the outcome of inadequate testing procedure, perhaps most likely to have occurred with the children, or could be due to poor general health status (e.g., viral infection or malnourishment). In addition, the skin tests may have been done too early, before the development of cell-mediated immunity as a result of infection. The probability that the PPD-negative tests were falsely negative was incorporated in a subsequent analysis model.

The preliminary models also assumed that the clinical status assigned to each family member on the basis of certain diagnostic criteria was correct. Estimates of the uncertainty of the diagnoses was incorporated in a subsequent analysis model to analyze the effects of possible misclassification.

3.7.1.5 Marker allele frequencies

For all families, published allele frequencies in Caucasians (n = 643–648, depending on the marker) were used for analysis with the RFLP markers at the CRYG1 and CRYGP1 locus (O'Connell et al., 1989). All other RFLP marker allele frequencies were estimated from data on 24 unrelated individuals in 12 Caucasian families collected to study diabetes susceptibility (48 chromosomes total). DNA for the diabetes families was purchased from the Human Biological Data Interchange (U.S.). For the microsatellite markers, allele frequencies were estimated using the observed and inferred data in 15 founder individuals assumed to be unrelated in the Canadian pedigree. In this case,

between 21 and 24 chromosomes were informative, depending on the marker; in some instances, typing for only one of two chromosomes of a founder could be inferred.

3.7.2 Model of diagnostic uncertainty

In the LINKAGE programs, individuals with different levels of disease risk can be assigned to different liability classes for an affection status locus; for each class there is a set of penetrance values. A level of diagnostic uncertainty was incorporated into the recessive preliminary model without phenocopies, by modifying the penetrance for the three genotypes at the susceptibility locus in several liability classes. The purpose of these modifications was to demonstrate how diagnostic uncertainty could be used in a linkage analysis and to examine the impact of incorporating the perceived accuracy of the diagnoses on the lod scores of the preliminary model. For each liability class, the penetrances were weighted by the positive predictive value (PPV) or negative predictive values (NPV) of the diagnoses, defined respectively as the probability of being truly affected given an affected diagnosis, and the probability of being truly unaffected given an unaffected diagnosis. When the certainty of an affected diagnosis and an unaffected diagnosis were the same for a group of individuals, each person was assigned to the same liability class, using the MLINK program's phenotype coding system.

PPV and NPV estimates for the diagnoses of the Colombian, Hong Kong, and Canadian pedigree members were obtained from the medical investigators in each region. Each investigator was asked to indicate approximate PPVs and NPVs for the diagnostic situations encountered with the study family members. Seven diagnostic liability classes were constructed using these estimates, listed as follows: (1) 100% PPV or definitely affected (for cases in Hong Kong families #1–13); (2) 100% NPV or definitely unaffected (for non-cases in Hong Kong families #1–9 and #11–13); (3) 80% PPV or affected with 80% certainty (for cases in the Colombian families); (4) 80% NPV or unaffected with 80% certainty (for non-cases in Hong Kong family #10 and for non-cases under 13 years of age in the Canadian family); (5) 85% NPV or unaffected with 85% certainty (for non-cases in the Colombian families); (6) 90% PPV or affected with 90% certainty (for cases in the Canadian family); and (7) 50% NPV or unaffected with 50% certainty (for non-cases over 13 years of age in the Canadian family). Because diagnoses for classes 1 and

2 and for classes 3 and 4 were equal in certainty, five separate liability classes were required. The penetrance specifications were calculated as follows:

(1) For liability class 1 (definitely affected or definitely unaffected):

In the absence of phenocopies, and using a base penetrance level of 62% for the higher risk r/r genotype, the penetrances required no adjustment; that is, the following table applied.

Penetrance specifications for liability class 1 (100% diagnostic certainty)

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.0 = P(truly affected) × P(affected|R/R) = 1 × 0	0.0 = P(truly affected) × P(affected|R/r) = 1 × 0	0.62 = P(truly affected) × P(affected|r/r) = 1 × 0.62
unaffected	1.0	1.0	0.38

(2) For liability class 2 (affected or unaffected with 80% certainty):

In the absence of phenocopies, and using a 62% base penetrance level for the higher risk r/r genotype, the following table applied.

Penetrance specifications for liability class 2 (80% diagnostic certainty)

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.2 = P(truly affected) × P(affected|R/R) + P(truly unaffected) × P(unaffected|R/R) = (0.8 × 0) + (0.2 × 1)	0.2 = P(truly affected) × P(affected|R/r) + P(truly unaffected) × P(unaffected|R/r) = (0.8 × 0) + (0.2 × 1)	0.572 = P(truly affected) × P(affected|r/r) + P(truly unaffected) × P(unaffected|r/r) = (0.8 × 0.62) + (0.2 × 0.38)
unaffected	0.8	0.8	0.428

(3) For liability classes 3, 4, and 5 with diagnostic certainty levels of 85%, 90%, and 50%, respectively, the resulting penetrances were as given in the following tables.

Penetrance specifications for liability class 3 (85% diagnostic certainty)

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.15	0.15	0.584
unaffected	0 85	0.85	0.416

Penetrance specifications for liability class 4 (90% diagnostic certainty)

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.1	0.1	0.596
unaffected	0.9	0.9	0.404

Penetrance specifications for liability class 5 (50% diagnostic certainty)

	penetrance of genotype		
phenotype	R/R	R/r	r/r
affected	0.5	0.5	0 5
unaffected	0.5	0.5	0.5

For the Canadian linkage analysis pedigree, non-cases who were also PPD-negative were not assigned a diagnostic NPV because of the possibility that they were actually not exposed to *M. tuberculosis*. Therefore, the five PPD-negative persons in the Canadian pedigree were given an unknown clinical status for the model of diagnostic uncertainty. All other Canadian family members were assumed infected. The unknown status code was also assigned to the three persons in the Hong Kong families who did not

live with infectious relatives. The effect of diagnostic uncertainty on the lod scores for linkage between disease and chromosome 2q markers was examined in the Canadian, Colombian, and Hong Kong families with the RFLP markers only.

3.7.3 Sensitivity of lod score results for disease and TNP-1 C, Canadian family

In linkage analysis with the Canadian family under the preliminary model, a significantly positive lod score was computed with the TNP-1 C marker. The lod scores for linkage between the disease and the TNP-1 C locus were then evaluated for various values of penetrance, susceptibility allele frequency, marker allele frequency, and consanguinity in the pedigree, using the MLINK program at specified recombination fractions.

3.7.3.1 Penetrance

The 62% penetrance level in the preliminary model was derived from the Kallmann and Reisner twin study (1943) and may not have accurately represented the risk of disease for those of higher risk genotype in the Canadian family. The relationship between the lod scores for linkage between disease and TNP-1 C and penetrance was examined by varying the penetrance in a single liability class recessive model, with the susceptibility allele frequency held constant at 0.2. The penetrance for the higher risk (r/r) genotype in the Canadian family was expected to be substantial given the infection intensity but likely not complete due to non-uniformity of exposure and the intervention with chemoprophylaxis and treatment in August 1987. The penetrance of the r/r genotype was thus varied from 50% to 95% in increments of 5% and the maximum lod score at each penetrance value was determined. Each penetrance model was analyzed a second time with a 50% phenocopy rate in the population to examine the effect of varying penetrance when phenocopies were included. For the phenocopy models, the penetrance assigned to the lower risk R/R and R/r genotypes was equal to the penetrance of the higher risk r/r genotype divided by 24, as determined in section 3.7.1.3.

3.7.3.2 Susceptibility allele frequency

Despite much consideration about an appropriate estimate of the susceptibility allele frequency in the Canadian family, the choice of a frequency of 0.2 was fairly

arbitrary. To examine the relationship between the lod scores for linkage between disease and TNP-1 C and susceptibility allele frequency, *q*, the recessive, preliminary model with 62% penetrance and no phenocopies was analyzed with the *q* value varied from a low frequency of 0.05 (0.25% of the population would be expected to be homozygous r/r), through 0.1 to 0.9 in increments of 0.1, to a high frequency of 0.95 (90% of the population would be expected to be homozygous r/r). The same susceptibility allele frequencies were then used for analysis with a 50% phenocopy rate in the population, which meant a penetrance of 0.62/24 was assigned to the lower risk R/R and R/r genotypes.

3.7.3.3 Marker allele frequency

The lod score method of linkage analysis can yield false positive evidence for linkage if marker allele frequencies are inappropriately specified and there are untyped parents in the pedigree. Several parents were untyped in the Canadian pedigree. The frequency of the TNP-1 C allele 1 used in the preliminary model was estimated in the diabetes families as 0.33. The relationship between the lod scores for linkage between disease and TNP-1 C and marker allele frequency was examined by varying the frequency of the TNP-1 C allele 1 from very rare values of 0.01 and 0.05, through the range 0.1 to 0.9 in increments of 0.1, to extremely common values of 0.95 and 0.99. Each time, the recessive, preliminary model was analyzed with $q = 0.2$, 62% penetrance, and no phenocopies. The models were also analyzed with the addition of a 50% phenocopy rate in the population while the frequency of the TNP-1 C allele 1 was varied in the same manner as above.

3.7.3.4 Consanguinity

In the process of estimating the susceptibility allele frequency, it was questioned whether some of the founders of the Canadian pedigree were closely related as a consequence of endogamy. Specifically, it was thought that possibly (1) descendants of the great-grandparents were closely related to their mates, resulting in consanguinity loops and/or, (2) the mates of the descendants of the great-grandparents were closely related to other in-marrying individuals, resulting in marriage loops (see Appendix B3). Pedigree loops, if they involve very close relationships, should be incorporated into the linkage

analysis as they may provide additional information about phase, as long as typing at marker loci is available for proximal ancestors in the loops. If none or few of the persons connecting the two closely-related family members in the pedigree are typed, the lod scores would not be expected to change significantly, if at all, when the loops are incorporated in the analysis.

Further investigation of possible pedigree loops in the Canadian family was undertaken for close relationships between two parents (such as first cousin, first cousin once removed, and second cousin relationships), or between in-marrying individuals (such as parent-child, sibling, half-sibling, aunt/uncle-nephew/niece relationships, in addition to first or second cousins). The contact family member was sent a questionnaire and pedigree diagram in August 1993 and was interviewed for more detailed information. This information was incorporated by modifying a pedigree file for linkage analysis with the loops. The correct specification of the loops in the pedigree file was verified by inputting the file to Pedigree/Draw, which produced a diagram that was checked by hand with the loop information. The pedigree including the loops identified in 1993 is given in Appendix B4. The effect of the previously unknown consanguinity in the pedigree on the lod score results for disease and TNP-1 C was evaluated by running the preliminary model with and without a 50% phenocopy rate on the looped and the non-looped pedigrees. Selected linkage analyses were carried out by Dr. Morgan with the looped pedigree using the MENDEL program to verify MLINK results.

3.7.4 Adjustment to the preliminary models

The original estimate of penetrance for the r/r genotype in the preliminary models (62%) was directly estimated as the monozygous twin tuberculosis concordance rate (unadjusted for age) in the Kallmann and Reisner twin study (1943). This concordance rate was actually an underestimate of the penetrance because the sample of index twins was ascertained from a population of affected individuals instead of twins, so that a proportion of monozygous twins concordant for non-disease were not sampled. Although the effect of the initial, inadequate estimate on the linkage analysis results was expected to be small, an appropriate estimate of penetrance was subsequently made to take into account the sample ascertainment method in the source study. The properly estimated

penetrance was 76%. Appendix C describes the justification for the required estimate and shows linkage results for selected analysis with the appropriate estimate.

3.7.5 Two-point linkage analysis of markers

A detailed genetic map is crucial for the localization of a disease or susceptibility locus to a specific chromosomal region. Linkage analysis of markers enables the construction of such a map, and will often involve both two-point and multi-point analysis in order to assign markers to particular chromosomes and then order the markers. The key to genetic maps is the occurrence of recombination, which is necessary for the ordering of loci (Ott, 1991). It is preferable that the marker mapping data is not drawn exclusively from families studied for a disease trait and, in general, a fairly large number of independent and informative families are needed for an accurate map. Map construction has been facilitated by the occurrence of highly polymorphic loci and the use of reference pedigrees, such as the Centre d'Etude du Polymorphisme Humaine (CEPH) families, which have a particular structure (usually three generations). With a known map of markers in a region of interest, the power of linkage analysis with a disease trait is increased.

A genetic map containing all of the markers used in this study had not been made. Partial maps with either RFLP markers of genes of known function or anonymous microsatellite markers were available from published sources. In addition, the Cooperative Human Linkage Center (CHLC), provided marker maps by electronic mail for the microsatellite markers.

In an attempt to assemble a consensus map from various sources, two-point linkage analysis of the marker data was carried out in order to estimate recombination fractions between the markers, using the MLINK and ILINK programs. Four individuals and two families were excluded as in the linkage analysis of the disease trait. The total number of families for analysis of the RFLP markers was 17, whereas only the Canadian pedigree was studied for the microsatellite markers. Marker allele frequencies were those used in the preliminary linkage analysis of the disease trait, and the Canadian family was analyzed without pedigree loops.

In the Canadian pedigree, haplotypes were constructed to determine a likely order of the microsatellite markers. Using the published map information, the method involved inspection of the segregation of the marker alleles in the family to infer haplotypes for 14 markers, and to identify the intervals between markers in which a crossover event likely occurred. The review of published and CHLC sources, along with the two-point linkage analysis and haplotype construction, were used to develop a tentative order for the chromosome 2 markers used in this study. This map was used in the presentation of the lod score results for the linkage analysis between the markers and the disease trait in section 4.

3.7.6 Epidemiological models, Canadian family

3.7.6.1 Model specifications

With the Canadian linkage analysis pedigree, more detailed penetrance specifications were used that incorporated epidemiological data from the community and the clinical information available for 66 family members. A total of three epidemiological models were developed and are described as follows.

Model 1: No phenocopies model

In this epidemiological model, which assumed the absence of phenocopies, there were three liability classes of disease risk; the different classes had different probabilities of disease for the r/r genotype. For the susceptibility locus, liability class 1 was assigned to those persons most likely to develop disease, who were pedigree members probably exposed to *M. tuberculosis* for the first time. If, under the conditions of high intensity of exposure in the Canadian family, genotype at the putative "BCG" susceptibility locus was crucial for containment or development of disease, the expected penetrance for the susceptible r/r genotype would be very high. However, the disease trait was likely incompletely penetrant, primarily because exposure in the family was probably not uniform in intensity for all members. Dose and frequency of exposure would differ according to (1) the nature of contact with the source case, (2) the nature of contact with other affected persons in the family, and (3) the infectiousness of each person's contacts. In fact, the presence of a total of eight PPD-negative non-cases in the 66-member

Canadian family, if indicating insufficient mycobacterial exposure to mount delayed-type hypersensitivity, suggested variability in exposure level within the pedigree.

In order to take into account such variability in exposure, the penetrance assigned to the r/r genotype in class 1 was (1 – an index of insufficient exposure). The index used was the number of PPD-negative persons among the non-cases in the family divided by the number of family members skin tested, which was 8/55 or 0.145. The penetrance in the first liability class was thus estimated as *(1 – 0.145)* or 0.85. Class 1 liability was assigned to the 41 infected persons who developed a newly positive skin test with ≥ 5 mm of induration during the outbreak, and to the three persons with disease in the 1950s who were not known to be vaccinated or PPD-positive prior to being diagnosed as cases. The previous cases were 31, 5, and 3 years old at the time of diagnosis. Of course, under this multiple liability model without phenocopies, the assignment of diseased individuals to a particular class made no difference to the inference of their genotype; all diseased individuals must necessarily be inferred as r/r.

Class 2 liability was assigned to persons with a lower risk of tuberculous disease because of a protective effect provided by previous mycobacterial infection or BCG vaccination, perhaps influenced by cell-mediated aspects of the immune response (Sutherland, 1976; Stead, 1992). For this class, the assumption was made that previous exposure was similar in extent of protection as BCG vaccination and that the effect would persist for many years. A matched case-control study (n = 100 matched triplets; 2 controls for each case) of the effectiveness of BCG vaccination in preventing later tuberculous disease in a comparable Aboriginal population was used to estimate penetrance in class 2 (Houston et al., 1990). The protective effect of BCG vaccination was estimated as 53% when the vaccination history of tuberculosis cases and controls was studied. The Mantel-Haenszel odds ratio of 0.47 for disease was used as the estimate of penetrance for the susceptible r/r genotype in this class. The odds ratio reflected a variety of *M. tuberculosis* exposure experiences, time periods since vaccination, and presumably genetic variation. The assumption was made that either vaccination failed predominantly among those of "BCG" susceptible genotype or that failure was independent of genotype at the "BCG" locus.

The odds ratio was similar to the failure rate of BCG vaccination in a case-control study of Manitoba Aboriginals in which the proportion of vaccinated individuals that were later diagnosed with disease was 40% (Young and Hershfield, 1986). Class 2 liability was assigned to eight pedigree members with evidence of prior *M. tuberculosis* exposure, as indicated by a positive skin test usually in the 1970s or early 1980s, or a record of BCG vaccination and a positive skin test in the 1980s. One person in the class, however, had not been skin tested since a positive result in 1958.

The third liability class corresponded to those at the lowest risk for disease during the outbreak because of absence of infection with *M. tuberculosis*. Class 3 liability was assigned to the five pedigree members with negative skin test results who may not have been infected if the test was truly negative. The false negative rate of PPD skin testing among tuberculosis cases has been estimated as 25% during initial evaluation of the patients (Holden et al., 1971). The risk of disease for the r/r genotype in class 3 was set to 25%.

The penetrances for the first epidemiological model (no phenocopies) were as presented in the following table, where the susceptibility allele is denoted as r.

Penetrance specifications for epidemiological model 1 (no phenocopies)

phenotype	liability class	penetrance of genotype		
		R/R	R/r	r/r
affected	1	0.0	0.0	0.85
	2	0.0	0.0	0.47
	3	0.0	0.0	0.25
unaffected	1	1.0	1.0	0.15
	2	1.0	1.0	0.53
	3	1.0	1.0	0.75

Unaffected individuals could be any of three possible genotypes, because the disease trait was incompletely penetrant. The relative probabilities of the three genotypes

differed depending on the liability class; while the penetrances of the R/R and R/r genotypes were always assigned to be the same, the probability that an r/r genotype would be inferred for an unaffected person increased as one moved from liability class 1 to 3.

Under the assumption that the frequency of the susceptibility allele was the same in all the generations of the Canadian family, the expected prevalence of disease, K, for the epidemiological models was

$$K = \Sigma_i \Psi_i [f_{xi} (1 - q^2) + f_{yi} q^2]$$

where Ψ_i is the prior probability of being in liability class i (or the proportion of at-risk persons in the i'th liability class), f_{xi} is the average penetrance of the lower risk genotypes R/R and R/r in liability class i, f_{yi} is the average penetrance of the higher risk genotype r/r in liability class i, and q is the frequency of the susceptibility allele.

The prior probability of being in liability class 3 was estimated from the proportion of PPD-negative individuals in the Canadian family, regardless of disease status, which was 9/55 family members tested or 16%. A community-derived estimate of PPD negativity would not have been a reliable estimate of this prior probability because a lower level of exposure was likely experienced by other persons (outside the family) during the outbreak A community estimate was available for the prior probability of being in class 2 because 14% of the community was known to be skin test positive or BCG vaccinated prior to the outbreak (Dr. Fanning, 1993, personal communication). The prior probability of being in class 1 was estimated as $1 - (0.14 + 0.16)$ or 70% because the three liability classes were mutually exclusive[4]. With these prior probabilities the expected prevalence of disease in the population under the first epidemiological model could be estimated as

$$\begin{aligned} K = \; & 0.70 [f_{x1} (1 - q^2) + f_{y1} q^2] + \\ & 0.14 [f_{x2} (1 - q^2) + f_{y2} q^2] + \\ & 0.16 [f_{x3} (1 - q^2) + f_{y3} q^2] \end{aligned}$$

[4]An estimate of this prior probability from the family would be 44/55 or 80%, calculated as the proportion of newly PPD-positive persons among those skin tested.

where $f_{x1} = f_{x2} = f_{x3} = 0$ and $f_{y1} = 0.85$, $f_{y2} = 0.47$, $f_{y3} = 0.25$
which simplified to

$$K = [0.7(0.85) + 0.14(0.47) + 0.16(0.25)]\, q^2$$
$$K \approx 0.7\, q^2$$

Model 2: Phenocopy model with three liability classes

The second epidemiological model differed from the first in that susceptibility was added to the R/R and R/r genotypes in each liability class. Under this model, it was assumed that the relative risk for disease for genotypes at higher risk (r/r genotype) was ten times the risk for genotypes at lower risk (R/r and R/R genotypes), regardless of liability class, so that $f_{yi} / f_{xi} = 10/1$ for $i = 1 \ldots 3$. This relative risk assigned a relatively strong protective effect to the R allele. As shown by Greenberg (1993), linkage may be difficult to detect for a susceptibility locus unless the probability of disease is ten times more likely given a higher risk genotype. The penetrances required under the second epidemiological model are presented in the following table.

Penetrance specifications for epidemiological model 2 (with phenocopies)

phenotype	liability class	penetrance of genotype		
		R/R	R/r	r/r
affected	1	0.085	0.085	0.85
	2	0.047	0.047	0.47
	3	0.025	0.025	0.25
unaffected	1	0.915	0.915	0.15
	2	0.953	0.953	0.53
	3	0.975	0.975	0.75

The expected prevalence of disease under the second epidemiological model was estimated with the same prior probabilities of belonging to each liability class as in the first model. This gave

$$K = 0.70\,[0.085(1 - q^2) + 0.85q^2] +$$

$$0.14\,[0.047(1 - q^2) + 0.47q^2] +$$

$$0.16\,[0.025(1 - q^2) + 0.25q^2]$$

$$K \approx 0.07 + 1.5\,q^2$$

Model 3: Phenocopy model with four liability classes

The third epidemiological model differed from the previous two in that the effect of age on disease risk was included in a fourth liability class. Infants younger than 2 years may have increased susceptibility in general to tuberculosis due to the immaturity of the immune system. The capability of particular host responses to infection, perhaps including macrophage function, likely increases in normal infants as the immune system develops. Children under 2 years in the Canadian family were likely at different stages of immune development so that the protective effect of the resistant genotype was less strong. At the same time, older persons greater than 65 years of age experience a deterioration of immune function which may increase the risk of tuberculosis. The protective effect of the resistant genotype might be expected to wane for elderly persons exposed to infection.

The average risk of disease assigned to the R/R and R/r genotypes in the fourth liability class was set to one-half the risk in the r/r genotype. In other words, the R/R and R/r genotypes were assumed to have had a weaker protective effect in this class than in liability classes 1, 2, and 3, and a larger proportion of phenocopies could have occurred among affected individuals in the fourth liability class. Class 4 liability was assigned to the four persons in the pedigree under 2 years of age. As it happened, the only person over 65 years in the pedigree was a previous case; for this person, the affected status referred to his/her first diagnosis of tuberculosis which occurred at the age of 31 years, so that he/she remained in class 1 under the third epidemiological model.

The penetrance of the r/r genotype in class 4 should have been at least as high as that in class 1 since persons in class 4 were assumed to have been newly infected. In the absence of any additional information for disease risk in infants under the conditions experienced by the Canadian family, the penetrance for the r/r genotype in class 4 was

set to the same level as that in class 1, or 85%. The penetrances for the third epidemiological model are presented in the following table.

Penetrance specifications for epidemiological model 3 (with phenocopies)

phenotype	liability class	penetrance of genotype		
		R/R	R/r	r/r
affected	1	0.085	0.085	0.85
	2	0.047	0.047	0.47
	3	0.025	0.025	0.25
	4	0.425	0.425	0.85
unaffected	1	0.915	0.915	0.15
	2	0.953	0.953	0.53
	3	0.975	0.975	0.75
	4	0.575	0.575	0.15

In order to estimate the prior probability of belonging to liability class 4, family data was used because community age statistics were not available. In the 66 member family, seven persons were under 2 or over 65 years, that is, about 10%. The probability of membership in class 1 was thus modified to $[1 - (0.14 + 0.16 + 0.1)]$ or 0.6 under this model. The expected prevalence of disease under the model was estimated as

$$\begin{aligned} K = \; & 0.70\,[0.085(1 - q^2) + 0.85q^2] + \\ & 0.14\,[0.047(1 - q^2) + 0.47q^2] + \\ & 0.16\,[0.025(1 - q^2) + 0.25q^2] + \\ & 0.10\,[0.425(1 - q^2) + 0.85q^2] \end{aligned}$$

$$K \approx \; 0.11 + 2.3\,q^2$$

Table 7 demonstrates the predicted disease prevalence and phenocopy rate for each of the epidemiological models, for values of $q = 0.2$ (4% of population homozygous r/r) and $q = 0.4$ (16% of population homozygous r/r).

Table 7. Predicted prevalence of disease and phenocopy rates under three epidemiological models (for two selected susceptibility allele frequencies, q)

Model	Liability class	Number of persons assigned to class	Prior probability	q	Predicted disease prevalence	Predicted phenocopy rate
1*	1	44	0.70	0.2	2.8%	0
	2	8	0.14	0.4	11%	0
	3	5	0.16			
2†	1	44	0.70	0.2	9.5%	70%
	2	8	0.14	0.4	17%	34%
	3	5	0.16			
3†	1	40	0.60	0.2	13%	78%
	2	8	0.14	0.4	20%	44%
	3	5	0.16			
	4	4	0.10			

Notes.

* no phenocopies
† with phenocopies

3.7.6.2 Two-point linkage analysis with disease

With the Canadian pedigree, linkage analysis was carried out under the three epidemiological models with the disease trait and TNP-1 C, the marker which provided the highest lod score in the preliminary analysis with the disease. The non-looped and looped pedigrees were used under each of the three models, with susceptibility allele frequencies of 0.2 and 0.4, using the MLINK program at selected recombination fractions

(the results are displayed in Appendix D). The lod score results under the three epidemiological models were fairly similar, producing parallel curves. The models differed only in allowance for phenocopies and the re-assignment of a few individuals to a fourth liability class of greater liability for the lower risk genotypes, which modified the lod scores only slightly. Linkage analysis of the disease and the full set of the RFLP and microsatellite markers was carried out under the most comprehensive epidemiological model (four liability classes) only, as this model was designed to be the most "realistic". The model was run with the larger susceptibility allele frequency of 0.4 to allow for a higher prevalence of disease in the population and a lower proportion of phenocopies; without evidence of concomitant disease in the family, the phenocopy rate of 78% predicted with an allele frequency of 0.2 may have been too high.

Section 3.7.3 concentrated on the lod score results achieved with the disease trait and TNP-1 C, the marker providing the highest lod score with the Canadian pedigree. However, there were two other markers of the TNP-1 locus which showed a lack of recombination with TNP-1 C in two-point analysis of the markers. Because more matings may have become informative when the three markers were combined, it was more appropriate to use the markers simultaneously and form a TNP-1 haplotype to analyze with the disease trait. There was a total of 14 sets of matings in the Canadian pedigree. Under no phenocopy models for which affected parents were inferred to be r/r at the susceptibility locus, the three markers were informative for different matings in the pedigree: TNP-1 B and C were informative for more matings than TNP-1 A (seven and eight matings for B and C, respectively); TNP-1 A was informative for only two matings, one of which was not informative for TNP-1 B or C.

The TNP-1 haplotype and the other RFLP markers were analyzed under the third epidemiological model (four liability classes) with the non-looped and looped Canadian pedigree using the MLINK program, at recombination fractions of 0.0, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, and 0.35. The microsatellite markers were analyzed at the same recombination fractions with the looped and the non-looped pedigree, using the marker allele frequencies estimated on the founders who were originally presumed to be unrelated. Linkage analysis of the complete set of the microsatellite markers was not

practical because of excessive demands on computing time for markers with more than three alleles in a pedigree with five loops. However, for the markers which were analyzed with both pedigree structures, the results showed only slight changes in the lod scores with the addition of the loops.

SECTION 4: RESULTS

4.1 Description of the linkage analysis pedigrees

4.1.1 Colombian and Hong Kong families

The Colombian and Hong Kong families lived in regions of elevated tuberculosis incidence. Incidence rates for years in which cases were being diagnosed, as estimated in the questionnaires, are presented in Table 8. All families except Hong Kong #10 lived in regions where the tuberculosis incidence was decreasing at the time the first case was diagnosed in each family. The location of residence of Hong Kong family #10 had a stable incidence of disease.

With respect to the structure of the pedigrees studied with linkage analysis (recalling the exclusion of several family members), the total number of persons in the four Colombian (GV, MAM, RAV, and RES) and 12 Hong Kong families (#1, 3–13) was 44 and 80, respectively. All of these international pedigrees contained one affected and one unaffected parent in the first generation, except family MAM. Three of the Colombian families (GV, MAM, and RAV) consisted of two generations with 1–6 affected children. Family RES consisted of three generations and contained two affected offspring. The Hong Kong pedigrees contained two generations of individuals and the number of affected children ranged from 1 to 4.

According to the clinical information in the Colombian questionnaire, the ages at diagnosis for the affected persons in the analyzed pedigrees ranged from 2 to 58 years. Included in this group were three persons who had had two tuberculosis diagnoses, 5–30 years apart. The total number of persons affected at least once was 15, including 10 males and five females with dates of diagnosis ranging from 1959–1990. Of the cases and non-cases analyzed, 87% and 76% were known to have been BCG vaccinated, respectively. For families GV, MAM, and RAV, all children but no parents were

Table 8. Tuberculosis incidence for Colombia and Hong Kong

Family	Region of residence	Tuberculosis cases per 100,000 (year)	Family	Region of residence	Tuberculosis cases per 100,000 (year)
GV	Itagui, Antioquia	25.8 (1987)	HK 5	Wong Tai Sin, New Kowloon	138.3 (1985)
MAM	Medellin, Antioquia	71.2 (1989)	HK 6	Tai Kok Tsui, Kowloon	223.2 (1971) 136.6 (1983)
RAV	Medellin, Antioquia	82.4 (1987)	HK 7	Ngau Tau Kok	123.6 (1988)
RES	Itagui, Antioquia	22.0 (1989)	HK 8	Temple Hill, New Kowloon	175.5 (1976)
RJA	Bello, Antioquia	22.6 (1989)	HK 9	Diamond Hill, New Kowloon	183.6 (1975)
HK* 1	Peng Chau Island	129.5 (1987)	HK 10	Chai Wan	134.3 (1986)
HK 2	Kwai Chung	129.5 (1987)	HK 11	Shek Kip Mei, Kowloon	138.3 (1985)
HK 3	Tsuen Wan	129.5 (1987)	HK 12	Kwun Tong	134.3 (1986) 109.2 (1991)
HK 4	Wan Chai	129.5 (1987)	HK 13	Tai Po	116.4 (1989)

source: International questionnaire data

* The abbreviation HK denotes Hong Kong

vaccinated. In family RES, all offspring who had no children of their own were vaccinated, as was the affected founder of the family (Appendix B1).

The ages at diagnosis for the Hong Kong family members ranged from 5 to 72 years in the questionnaires, the cases having been identified between 1971 and 1990. Six persons were indicated to have had a prior tuberculosis diagnosis but for only one individual was the year of the earlier diagnosis available. The total number of affected persons in the analyzed pedigrees was 39, including 27 males and 12 females. BCG vaccination status was unknown for 36% of the cases and 20% of the non-cases (22 individuals in total). Of the cases and non-cases whose vaccination status was known, 72% and 85% were indicated to have been vaccinated, respectively. Only in one family was every member known to have been vaccinated (Hong Kong #10) (Appendix B2).

4.1.2 Canadian family

The Canadian linkage analysis pedigree contained 61 persons, with clinical status information for 57 individuals. Twenty members of the pedigree were diagnosed with tuberculosis during the 1987–1989 outbreak. One of these individuals and two other pedigree members were previously affected with primary disease. The total number of affected persons in the linkage analysis pedigree was thus 22. For the person with two diagnoses of disease, the age at diagnosis used for the following descriptions was that at the time of the primary diagnosis. The mean age of the affected persons in the pedigree was 13 years with 36% of these being female. For the 37 members who remained non-cases during the outbreak and did not have a previous diagnosis of disease, their mean age at investigation was just under 20 years with 49% of these persons being female. A frequency distribution for the cases and non-cases in the pedigree by age and gender is presented in Table 9. The mean age of the cases and non-cases did not differ significantly although the non-cases tended to be older, perhaps because of a greater chance of being protected by previous exposure or BCG vaccination (t_{ss} = 1.91; $0.05 < p < 0.10$, two-tailed). Three persons, all non-cases, were BCG vaccinated as school-aged children in the late 1950s and early 1960s. The total proportion of females in the two groups did not differ significantly ($z = 0.9$; $p = 0.37$, two-tailed) (Armitage and

Table 9. Frequency distributions for cases and non-cases in the Canadian linkage analysis pedigree, by age and gender

Age category	Cases		Non-cases	
	Number	Number of females	Number	Number of females
≤ 2 years	3	0	1	1
3–13 years	8	3	16	7
14–25 years	9	4	7	3
26–60 years	2	1	11	6
> 60 years	0	0	0	0
Totals	22	8 (36%)	35	17 (49%)

Berry, 1987). The age and gender comparisons of the cases and non-cases made theassumption, as did many of the linkage analysis models, that all of the non-cases were infected during the outbreak and thus had the opportunity to develop disease.

4.2 Lod scores for linkage between disease and markers under the preliminary model

4.2.1 Colombian and Hong Kong families

Because of the similarity of the clinical phenotype definitions in the Colombian and Hong Kong families, the lod scores at each recombination fraction were summed across these pedigrees. Under the preliminary recessive model with 62% penetrance, a susceptibility allele frequency of 0.2 and no phenocopies, total lod scores summed across the international pedigrees were generally negative because recombinations were observed (Table 10). Significant evidence against close linkage between markers and the disease trait was achieved, with at least 100:1 odds (lod score of -2), at 0% recombination for TNP-1 A, to 1% recombination for VIL E-84, DESMIN, and INHB-A1, to 5% recombination for VIL 7 No.6, VIL E-62, and VIL pEX2(a), and to 10% recombination for TNP-1 C. The maximum lod score was not significant, reaching 0.3 for CRYG1-D at 20% recombination.

Table 10. Preliminary model lod score results, Colombian and Hong Kong families (no phenocopies)

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	-0.42	-0.45	-0.30	0.01	0.30	0.25	0.09
CRYGP1-A	-1.82	-1.71	-1.29	-0.85	-0.34	-0.11	-0.02
CRYGP1-B	-0.45	-0.36	-0.16	-0.05	0.02	0.02	0.01
CRYGP1-C	-1.88	-1.69	-1.19	-0.81	-0.39	-0.16	-0.04
FN61039-H	-0.81	-0.71	-0.46	-0.30	-0.13	-0.05	-0.01
FN61039-M	-1.18	-1.05	-0.66	-0.36	-0.08	0.00	0.01
TNP-1 A	**-2.06**	-1.88	-1.38	-0.95	-0.45	-0.17	-0.04
TNP-1 B	0.03	0.03	0.03	0.02	0.01	0.01	0.00
TNP-1 C	**-5.39**	**-4.64**	**-3.09**	**-2.07**	-0.97	-0.39	-0.09
VIL 7 No.6	**-3.43**	**-3.03**	**-2.20**	-1.57	-0.79	-0.34	-0.09
VIL E-62	**-3.75**	**-3.41**	**-2.53**	-1.71	-0.71	-0.24	-0.05
VIL E-84	**-3.32**	**-2.89**	-1.95	-1.30	-0.60	-0.23	-0.05
VIL pEX2(a)	**-3.18**	**-2.82**	**-2.16**	-1.64	-0.82	-0.32	-0.07
DESMIN	**-3.19**	**-2.72**	-1.83	-1.25	-0.60	-0.25	-0.06
INHB-A1	**-3.12**	**-2.70**	-1.88	-1.32	-0.66	-0.28	-0.07

Note. Lod scores were summed across all families at each recombination fraction. Indicated in bold print are lod scores which significantly exclude linkage at the recombination fraction specified. Marker loci are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

When the lod scores for linkage of disease and each marker were examined by family, the scores at 0% recombination were predominantly negative, with values from close to zero to about -2 (data not shown). Most often, negative values of between -0.1 and -1 were computed for a single family. Similar evidence against linkage was obtained by both Colombian and Hong Kong pedigrees despite their differences in ethnicity. Of the positive lod scores computed at 0% recombination, most were less than 0.5. A lod score of approximately 0.6 was calculated for two markers (CRYG1-D and VIL E-62) for Hong Kong families #6 and 10, respectively. A maximum lod score of 0.82 was computed at 0% recombination for CRYG1-D in the MAM family; in contrast, the same family provided a lod score of -2.04 at 0% recombination for DESMIN. For each marker, only 1–6 families were informative of the 16 international pedigrees analyzed.

Table 11 presents the total lod scores summed across the Colombian and Hong Kong families for analysis under the preliminary phenocopy model, with a 50% phenocopy rate in the population. Less evidence against linkage was obtained for markers which had provided negative minimum lod scores under the no phenocopy model (Table 10); despite the general increase in lod scores, however, recombinations were still observed because the majority of the scores were negative. The maximum lod score computed was 0.23 for CRYG1-D at 20% recombination. Linkage with disease was significantly excluded at 0% recombination for VIL pEX2(a), and to 1% recombination for TNP-1 C, VIL 7 No.6, and VIL E-62. Between one and 10 families, depending on the marker, were informative for analysis under the preliminary phenocopy model, an increase compared to analysis without phenocopies because a heterozygous R/r genotype could now be inferred for an affected person.

Based on conserved linkage between mouse chromosome 1 and human chromosome 2q, several of the RFLP markers were in the vicinity of the target region for a human "BCG" susceptibility gene. Under the preliminary models these markers, in particular those at the villin locus, provided significant evidence against linkage with the disease trait when the lod scores were added across the Colombian and Hong Kong families.

Table 11. Preliminary phenocopy model lod score results, Colombian and Hong Kong families

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	-0.41	-0.37	-0.15	0.06	0.23	0.18	0.06
CRYGP1-A	-1.13	-1.06	-0.80	-0.54	-0.23	-0.08	-0.02
CRYGP1-B	-0.13	-0.11	-0.05	-0.01	0.02	0.02	0.01
CRYGP1-C	-0.89	-0.84	-0.67	-0.50	-0.26	-0.11	-0.03
FN61039-H	-0.38	-0.35	-0.26	-0.18	-0.09	-0.03	-0.01
FN61039-M	-0.16	-0.12	-0.01	0.07	0.10	0.07	0.02
TNP-1 A	-1.36	-1.28	-0.98	-0.71	-0.34	-0.14	-0.03
TNP-1 B	0.03	0.03	0.02	0.02	0.01	0.00	0.00
TNP-1 C	**-2.73**	**-2.55**	-1.92	-1.38	-0.68	-0.29	-0.07
VIL 7 No.6	**-2.14**	**-2.02**	-1.61	-1.20	-0.63	-0.28	-0.06
VIL E-62	**-2.57**	**-2.41**	-1.82	-1.25	-0.53	-0.18	-0.04
VIL E-84	-1.78	-1.64	-1.23	-0.87	-0.42	-0.17	-0.04
VIL pEX2(a)	**-1.97**	-1.89	-1.58	-1.21	-0.62	-0.25	-0.06
DESMIN	-1.73	-1.61	-1.24	-0.89	-0.45	-0.19	-0.04
INHB-A1	-1.68	-1.58	-1.24	-0.93	-0.48	-0.21	-0.05

Note. Lod scores were summed across all families at each recombination fraction. Indicated in bold print are lod scores which significantly exclude linkage at the recombination fraction specified. Marker loci are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

4.2.2 Canadian family

Under the preliminary recessive model (no phenocopies) with the Canadian family, maximum lod scores approaching or exceeding a value of 2 (100:1 odds in favour of linkage) were computed at 0% recombination with TNP-1 B, TNP-1 C, and D2S128, and at 5% recombination with CRYG1-A (Tables 12 and 13). A maximum lod score approaching or exceeding a value of 1 (10:1 odds in support of linkage) was calculated at 20% recombination for CRYG1-D, D2S157, D2S137, D2S173, and D2S126. Close linkage was excluded with a lod score or -2 or less at 0% recombination for D2S157, D2S137, D2S173, and D2S172, to 1% recombination for D2S120 and PAX-3, and to 5% recombination for D2S206. Other loci either yielded minimum lod scores between -1 and -2 or lod scores within one unit of zero for all recombination fractions specified. The four VIL markers, which were expected to be close to the target region for a susceptibility gene on the basis of human-mouse homology, were most often homozygous, providing no information in favour of or rejecting linkage with the disease trait. The exception was the VIL E-84 marker, for which weak evidence against close linkage was observed.

Most significantly, the maximum lod scores with D2S128, TNP-1 B, and TNP-1 C were 1.96, 2.55, and 3.31, respectively, all at 0% recombination, indicating close linkage between the disease trait and the marker loci in the pedigree. The TNP-1 C result exceeded the critical lod score value of 3 for statistically significant support of linkage with disease. The approximate 95% confidence interval for the TNP-1 C result, defined as the range in recombination fraction values from Z_{max} to Z_{max} - *1* for a $Z_{max} > 3$ (the 1-lod unit support interval; Ott, 1991), was large and ranged from a lower bound of 0% recombination to an upper bound of 15% recombination.

Figure 7 shows lod scores for selected markers under the preliminary model (no phenocopies) for the Canadian and the international pedigrees. There is a large difference between the TNP-1 C lod scores of the Canadian family versus the Colombian and Hong Kong families, suggesting that the disease trait, as defined, is closely linked in the Canadian pedigree and not linked in the international pedigrees when their lod scores are summed. The Canadian pedigree provided weakly negative lod scores for the VIL E-84 marker, whereas the total lod scores for the same marker were significantly negative in

Table 12. Preliminary model lod score results with RFLP markers, Canadian family (no phenocopies)

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	-0.16	0.01	0.46	0.73	**0.86**	0.68	0.31
CRYGP1-A	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CRYGP1-B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CRYGP1-C	-1.22	-1.14	-0.88	-0.65	-0.34	-0.16	-0.05
FN61039-H	-1.83	-1.42	-0.64	-0.22	0.10	0.13	0.05
FN61039-M	-0.77	-0.56	-0.23	-0.08	0.03	0.06	0.05
TNP-1 A	-0.13	-0.12	-0.10	-0.08	-0.04	-0.02	0.00
TNP-1 B	**2.55**	2.52	2.39	2.20	1.73	1.15	0.49
TNP-1 C	**3.31**	3.25	3.00	2.68	2.00	1.27	0.53
VIL 7 No.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VIL E-62	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VIL E-84	-0.19	-0.18	-0.15	-0.12	-0.06	-0.03	-0.01
VIL pEX2(a)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DESMIN	-0.98	-0.86	-0.57	-0.38	-0.17	-0.07	-0.02
INHB-A1	-1.13	-0.91	-0.50	-0.28	-0.07	0.02	0.04

Note. Indicated in bold print are maximum lod scores in favour of linkage between the marker and the disease trait. Marker loci are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

Table 13. Preliminary model lod score results with microsatellite markers, Canadian family (no phenocopies)

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-A	1.25	1.49	**1.75**	1.74	1.45	1.00	0.44
D2S157	**-3.07**	-0.16	0.80	1.10	**1.11**	0.82	0.37
D2S128	**1.96**	1.94	1.84	1.69	1.33	0.89	0.38
D2S137	**-3.13**	-0.72	0.30	0.69	**0.87**	0.69	0.31
D2S104	-0.73	-0.61	-0.35	-0.20	-0.06	0.00	0.02
D2S173	**-2.98**	-0.92	0.15	0.56	**0.74**	0.57	0.22
D2S120	**-3.94**	**-3.06**	-1.25	-0.47	0.03	0.08	-0.01
D2S126	-1.59	-1.33	-0.14	0.42	**0.69**	0.55	0.22
D2S102	-1.25	-1.24	-1.11	-0.90	-0.51	-0.25	-0.09
PAX-3	**-2.67**	**-2.41**	-1.33	-0.68	-0.12	0.03	0.00
D2S172	**-2.67**	-1.38	-0.67	-0.27	0.06	0.10	0.02
D2S206	**-2.62**	**-2.61**	**-2.32**	-1.69	-0.85	-0.39	-0.12
D2S125	-0.94	-0.95	-1.01	-1.06	-0.93	-0.55	-0.22
D2S140	-0.89	-0.87	-0.77	-0.63	-0.39	-0.19	-0.06

Note. Indicated in bold print are maximum lod scores in favour of linkage and minimum lod scores that significantly exclude linkage at the specified recombination fraction. Markers are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the microsatellite map in section 4.5.

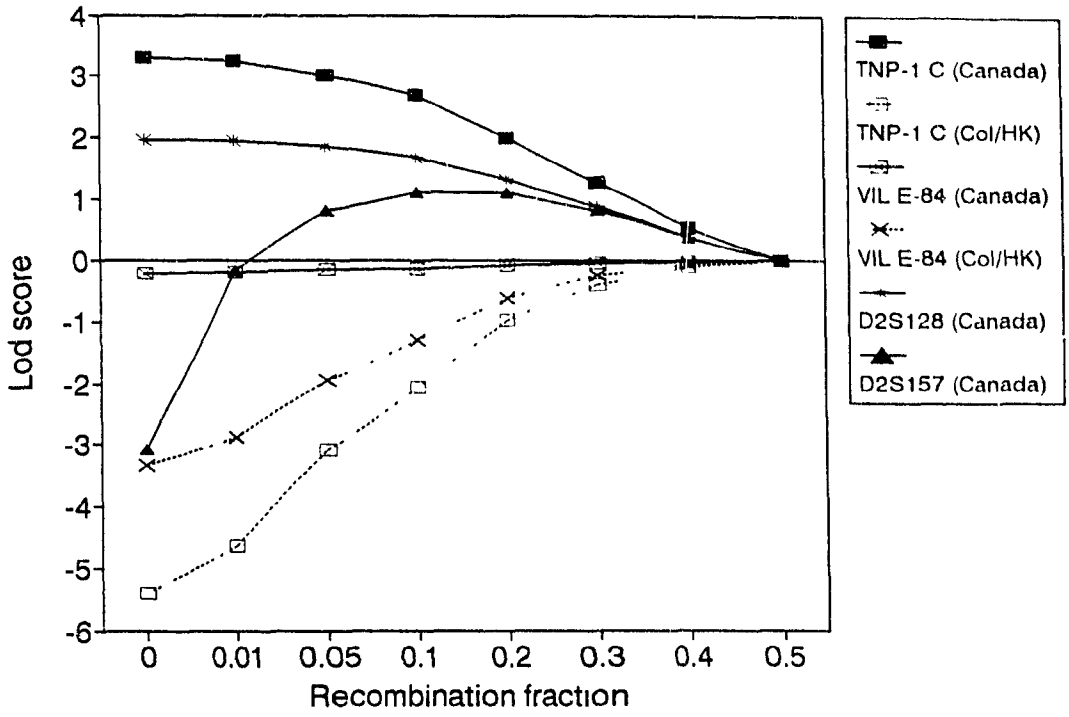


Figure 7. Lod score curve for selected markers under the preliminary model (no phenocopies)

Notes.
Canada refers to results with the Canadian pedigree
Col/HK refers to results summed over the Colombian and Hong Kong pedigrees

the international families, producing a similar curve as that for TNP-1 C. The three informative curves for the Canadian pedigree indicate that there is evidence for a susceptibility gene closely linked to TNP-1 C and D2S128 (with at least 100:1 odds), and at 10–20% recombination from D2S157 (with 10:1 odds). The D2S157 lod score curve for linkage with disease has a sharp bend because at least one recombination was observed (negative lod scores were computed at low recombination fractions); linkage with this marker is supported at higher recombination fractions, although not significantly.

The lod score results for the Canadian pedigree under the preliminary phenocopy model, with a 50% phenocopy rate in the population, are presented in Tables 14 (RFLP markers) and 15 (microsatellite markers). In general, the addition of phenocopies resulted in elevation of the minimum lod scores to more positive values, when compared to the results under the preliminary, no phenocopy model (see Tables 12 and 13 for comparison). For several markers, recombinations were no longer observed so that minimum lod scores greater than zero were calculated (CRYG1-D, D2S157, D2S137, D2S173, D2S126, and D2S206). The maximum lod scores for TNP-1 B and C decreased to 2.41 and 3.02, respectively, at 0% recombination, while maximum lod scores for other markers either remained at about the same level (CRYG1-D, CRYG1-A, D2S157, D2S173, and D2S126) or increased by about 0.1–0.2 (D2S128 and D2S137). With the addition of phenocopies, the recombination fraction at which the maximum lod scores occurred tended to be shifted closer to zero since recombinations were less likely under this model. Weak evidence against close linkage with disease was still observed for VIL E-84. Significant exclusion of linkage was provided by the D2S206 marker to 1% recombination. The linkage of the disease trait to the TNP-1 and D2S128 region continued to be supported in the Canadian pedigree.

4.3 Lod scores for linkage between disease and markers under the model of diagnostic uncertainty

4.3.1 Colombian and Hong Kong families

The effect of diagnostic uncertainty on the lod score results with the Colombian and Hong Kong families depended on the marker used, because estimates of diagnostic

Table 14. Preliminary phenocopy model lod score results with RFLP markers, Canadian family

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	0.45	0.53	0.75	**0.87**	0.87	0.64	0.27
CRYGP1-A	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
CRYGP1-B	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
CRYGP1-C	-0.82	-0.77	-0.61	-0.45	-0.24	-0.11	-0.03
FN61039-H	-0.36	-0.27	-0.05	0.11	0.21	0.17	0.05
FN61039-M	-0.08	-0.07	-0.04	-0.02	0.01	0.03	0.03
TNP-1 A	-0.06	-0.06	-0.05	-0.04	-0.02	-0.01	0.00
TNP-1 B	**2.41**	2.38	2.23	2.02	1.55	1.01	0.41
TNP-1 C	**3.02**	2.96	2.72	2.41	1.77	1.11	0.44
VIL 7 No.6	-0.01	-0.01	-0.00	0.00	0.00	0.00	0.00
VIL E-62	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00
VIL E-84	-0.16	-0.15	-0.12	-0.10	-0.05	-0.02	-0.01
VIL pEX2(a)	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
DESMIN	-0.57	-0.53	-0.40	-0.29	-0.14	-0.06	-0.01
INHB-A1	-0.36	-0.33	-0.25	-0.17	-0.06	0.00	0.02

Note. Indicated in bold print are maximum lod scores in favour of linkage between the marker and the disease trait. Marker loci are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

Table 15. Preliminary phenocopy model lod score results with microsatellite markers, Canadian family

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-A	1.60	1.63	**1.66**	1.60	1.31	0.88	0.37
D2S157	0.82	0.90	1.08	**1.16**	1.04	0.74	0.31
D2S128	**2.07**	2.04	1.88	1.68	1.26	0.81	0.32
D2S137	0.91	0.95	1.05	**1.09**	0.97	0.69	0.29
D2S104	-0.30	-0.27	-0.17	-0.10	-0.01	0.02	0.03
D2S173	0.35	0.45	0.68	**0.81**	0.78	0.54	0.20
D2S120	-1.10	-0.91	-0.41	-0.10	0.11	0.09	-0.02
D2S126	0.07	0.22	0.56	**0.74**	0.74	0.52	0.20
D2S102	-0.98	-0.94	-0.79	-0.61	-0.35	-0.17	-0.06
PAX-3	-1.41	-1.23	-0.73	-0.37	-0.04	0.03	-0.01
D2S172	-0.63	-0.52	-0.22	-0.02	0.13	0.10	0.01
D2S206	**-2.27**	**-2.18**	-1.77	-1.33	-0.74	-0.37	-0.12
D2S125	-0.84	-0.84	-0.86	-0.86	-0.71	-0.44	-0.19
D2S140	-0.78	-0.75	-0.65	-0.53	-0.34	-0.18	-0.06

Note. Indicated in bold print are maximum lod scores in favour of linkage and minimum lod scores that significantly exclude linkage at the specified recombination fraction. Markers are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the microsatellite map in section 4.5.

uncertainty differed for different families, and the particular families who were informative for linkage analysis depended on the marker. No change in results from the preliminary model was observed if the only informative pedigrees were among Hong Kong families 1–9 and 11–13, for whom diagnostic certainty (both PPV and NPV) was estimated as 100%.

The addition of diagnostic uncertainty to the preliminary recessive model (no phenocopies) for the Colombian and Hong Kong families generally resulted in a loss of power to exclude linkage (Table 16; see Table 10 for comparison). Significant evidence against close linkage was no longer provided by the VIL pEX2(a) and DESMIN markers. The FN61039-H marker locus became essentially non-informative with the addition of diagnostic uncertainty. The lod score at 0% recombination with the CRYGP1-A marker was modified from -1.01 to a non-significant maximum lod score of 0.30. Close linkage with the susceptibility locus was still excluded at 0% recombination for TNP-1 A, VIL E-84, and INHB-A1, to 1% recombination for VIL 7 No.6 and VIL E-62, and to 10% recombination for TNP-1 C.

4.3.2 Canadian family

For the Canadian family, the addition of diagnostic uncertainty to the preliminary recessive model (no phenocopies) had a large effect on the lod scores for linkage between the disease trait and the RFLP markers (Table 17; see Table 12 for comparison). The highest maximum lod scores were still provided by the TNP-1 B and TNP-1 C markers, although the scores at 0% recombination were non-significant, at 0.69 and 0.83, respectively. The incorporation of diagnostic uncertainty, particularly high for the non-affected diagnosis among adults, thus greatly affected the positive lod scores at the TNP-1 B and C loci, decreasing the scores to about one-quarter of their previous values. Markers for which lod scores approximately equal to or less than -1 had been computed at 0% recombination became less informative, with scores ranging from -0.07 to -0.30 (that is, the scores decreased in value by 75–95%).

Table 16. Lod score results with the diagnostic uncertainty model, Colombian and Hong Kong families

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	-0.10	-0.12	-0.12	-0.04	0.06	0.07	0.02
CRYGP1-A	0.30	0.29	0.26	0.22	0.13	0.07	0.02
CRYGP1-B	0.36	0.35	0.30	0.25	0.16	0.07	0.02
CRYGP1-C	-0.50	-0.47	-0.39	-0.30	-0.17	-0.07	-0.02
FN61039-H	-0 01	-0.01	-0.01	0.00	0.00	0.00	0.00
FN61039-M	-0.26	-0.23	-0.13	-0.05	0.01	0.02	0.01
TNP-1 A	**-2.12**	-1.94	-1.43	-1.00	-0.48	-0.19	-0.04
TNP-1 B	0.03	0.03	0.03	0.02	0.01	0.01	0.00
TNP-1 C	**-5.39**	**-4.64**	**-3.10**	**-2.07**	-0.98	-0.39	-0.09
VIL 7 No.6	**-2.94**	**-2.58**	-1.83	-1.28	-0.61	-0.25	-0.06
VIL E-62	**-2.69**	**-2.41**	-1.79	-1.25	-0.55	-0.20	-0.04
VIL E-84	**-2.01**	-1.84	-1.37	-0.98	-0.48	-0.19	-0.05
VIL pEX2(a)	-1.55	-1.25	-0.77	-0.50	-0.23	-0.09	-0.02
DESMIN	-0.77	-0.72	-0.55	-0.40	-0.20	-0.08	-0.02
INHB-A1	**-2.09**	-1.75	-1.13	-0.75	-0.35	-0.14	-0.03

Note. Lod scores are summed across all families at each recombination fraction. Indicated in bold print are lod scores which significantly exclude linkage at the recombination fraction specified. Marker loci are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

Table 17. Lod score results with the diagnostic uncertainty model for RFLP markers, Canadian family

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	0.06	0.07	0.08	0.08	0.06	0.03	0.01
CRYGP1-A	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CRYGP1-B	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CRYGP1-C	-0.30	-0.29	-0.24	-0.19	-0.11	-0.05	-0.02
FN61039-H	-0.09	-0.08	-0.06	-0.04	-0.02	-0.01	-0.01
FN61039-M	-0.03	-0.03	-0.03	-0.03	-0.02	-0.01	0.00
TNP-1 A	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TNP-1 B	**0.69**	0.67	0.58	0.47	0.27	0.12	0.03
TNP-1 C	**0.83**	0.80	0.68	0.54	0.30	0.13	0.00
VIL 7 No.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VIL E-62	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VIL E-84	-0.07	-0.07	-0.06	-0.05	-0.03	-0.01	0.00
VIL pEX2(a)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DESMIN	-0.17	-0.17	-0.13	-0.10	-0.05	-0.02	-0.01
INHB-A1	-0.07	-0.07	-0.06	-0.04	-0.02	-0.01	0.00

Note. Indicated in bold print are maximum lod scores in favour of linkage. Marker loci are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

4.4 Sensitivity of lod score results for linkage between disease and TNP-1 C, Canadian family

4.4.1 Penetrance

Figure 8 displays the relationship between lod scores for linkage between disease and TNP-1 C and penetrance level, under recessive models with a susceptibility allele frequency of 0.2, with and without phenocopies. As penetrance varied from 50-95%, the maximum lod scores under models without phenocopies ranged within less than one lod unit, from 2.91 at 0% recombination and 50% penetrance to the maximum score of 3.46 at 0% recombination and 75% penetrance, and down to 2.93 at 5% recombination and 95% penetrance. The empirical significance level of the maximum score of 3.46 was estimated using computer simulation with the assistance of Dr. Morgan; the methods and results of this procedure are presented in Appendix E. The 62% penetrance level specified for the higher risk r/r genotype under the preliminary model (no phenocopies) had generated a maximum lod score of 3.31 that was about mid-way along the top curve of maximum lod scores in Figure 8 as penetrance increased from 50–75%.

The lod score curve for the models with phenocopies is very similar in shape but is consistently lower than the curve for the models without phenocopies, so that there is slightly decreased evidence for linkage with TNP-1 C. The maximum lod scores under the phenocopy models also ranged in value within less than one lod unit, from 2.67 at 0% recombination and 50% penetrance to 3.18 at 0% recombination and 75% penetrance, and down to 2.6 at 5% recombination and 95% penetrance.

4.4.2 Susceptibility allele frequency

Figure 9 displays the maximum lod scores for linkage between disease and TNP-1 C under the recessive models with and without a 50% phenocopy rate, as the susceptibility allele frequency was varied from 0.05 to 0.95. The penetrance of the higher risk r/r genotype was set to 62% for all analyses. The lod score results under the models without phenocopies were quite robust to the variation of susceptibility allele frequency, ranging from a lod score of 3.34 at 0% recombination when $q = 0.05$ to a score of 2.74 at 0% recombination when $q = 0.95$. In between the two extremes in allele frequency, the curve without phenocopies decreases gradually, but is nearly flat in the susceptibility

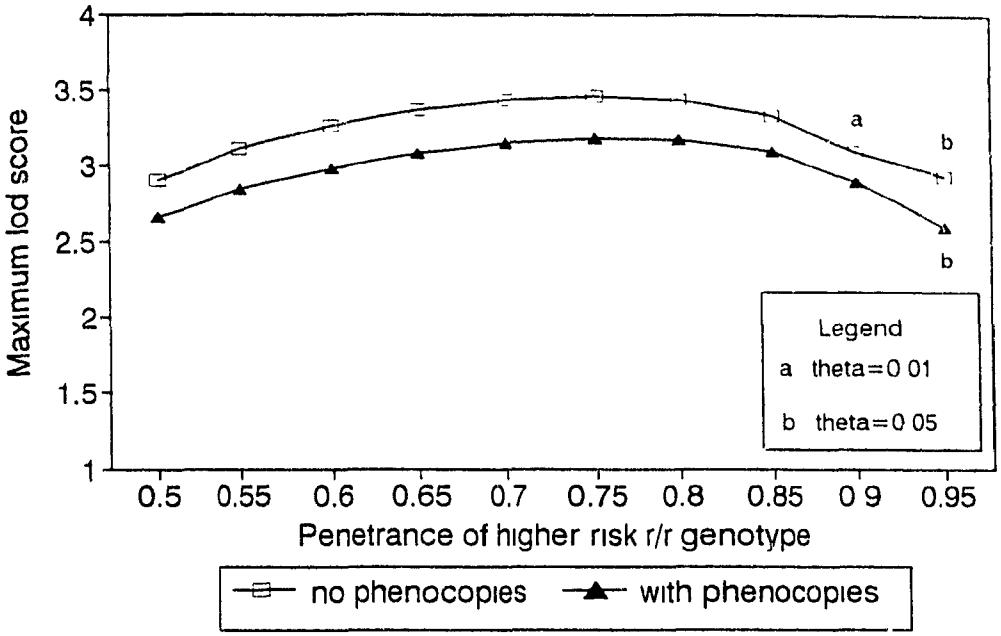


Figure 8. Lod scores for linkage between disease and TNP-1 C as a function of penetrance level

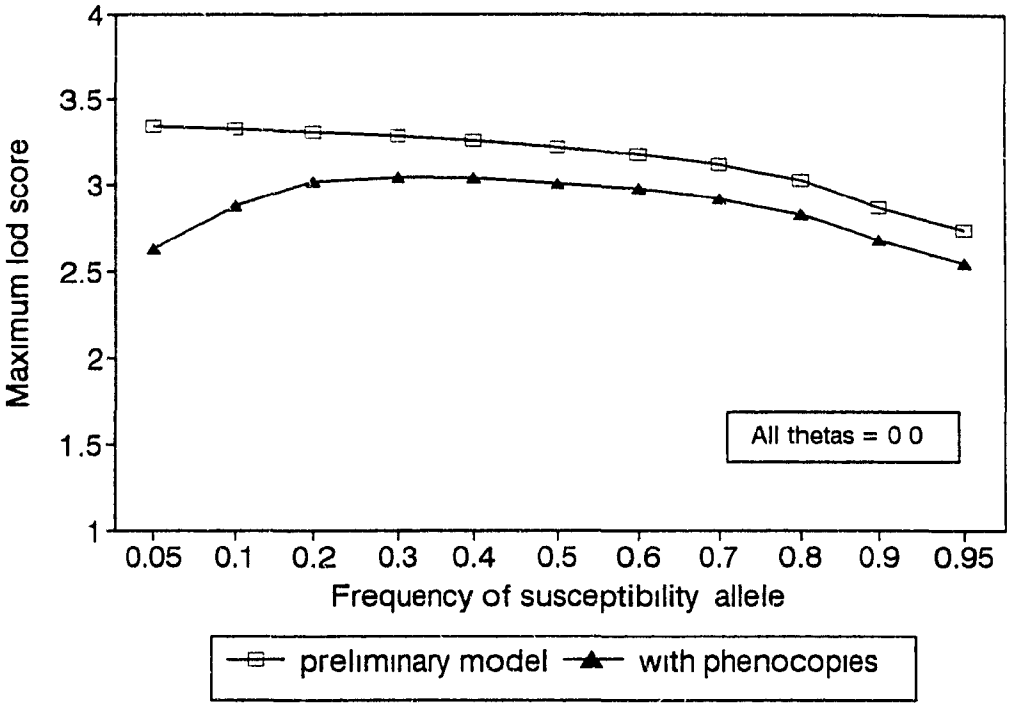


Figure 9. Lod scores for linkage between disease and TNP-1 C as a function of susceptibility allele frequency

allele frequency range of 0.05 to 0.4, which includes the allele frequency of 0.2 used for the preliminary model.

The addition of phenocopies to the models resulted in a lower curve of maximum lod scores at all susceptibility allele frequencies. The phenocopy curve is very similar in shape to the upper curve for susceptibility allele frequencies of 0.3–0.95. In this range of allele frequencies, the maximum lod scores under the phenocopy models were about 0.2 less than those under models without phenocopies. The two curves differ the most when the susceptibility allele frequency is set to 0.05 and 0.1. For these low allele frequencies the maximum lod scores under the phenocopy models decreased to 2.63 and 2.88, respectively, differing on average by about 0.6 from the scores under the corresponding models without phenocopies.

4.4.3 Marker allele frequency

Figure 10 displays the relationship between lod scores for linkage between disease and TNP-1 C and the frequency of marker allele 1, under the recessive models with and without phenocopies. The lod scores in both curves were computed with 62% penetrance for the higher risk r/r genotype and a susceptibility allele frequency of 0.2. As the marker allele frequency was varied from 0.01 to 0.99 the maximum lod scores decreased substantially: under the model without phenocopies the lod score decreased from 3.48 at 0% recombination to 1.15 at 20% recombination, while under the model with phenocopies the score decreased from 3.19 at 0% recombination to 1.25 at 10% recombination. However, within a wide range of marker allele 1 frequencies, from 0.01 to 0.5, all maximum lod scores under the model without phenocopies were greater than 3. For the same region, maximum lod scores under the model with phenocopies were consistently lower by about 0.3. Maximum lod scores of 3 and greater were maintained under the phenocopies model for marker allele 1 frequencies of approximately 0.4 and less.

The two curves in Figure 10 are very similar in shape. The phenocopy curve has less of a downward inflection at a marker allele 1 frequency of 0.8 so that the maximum lod scores for allele 1 frequencies larger than 0.8 are greater than those calculated under the models without phenocopies, by an average of 0.15. The frequency of marker allele 1

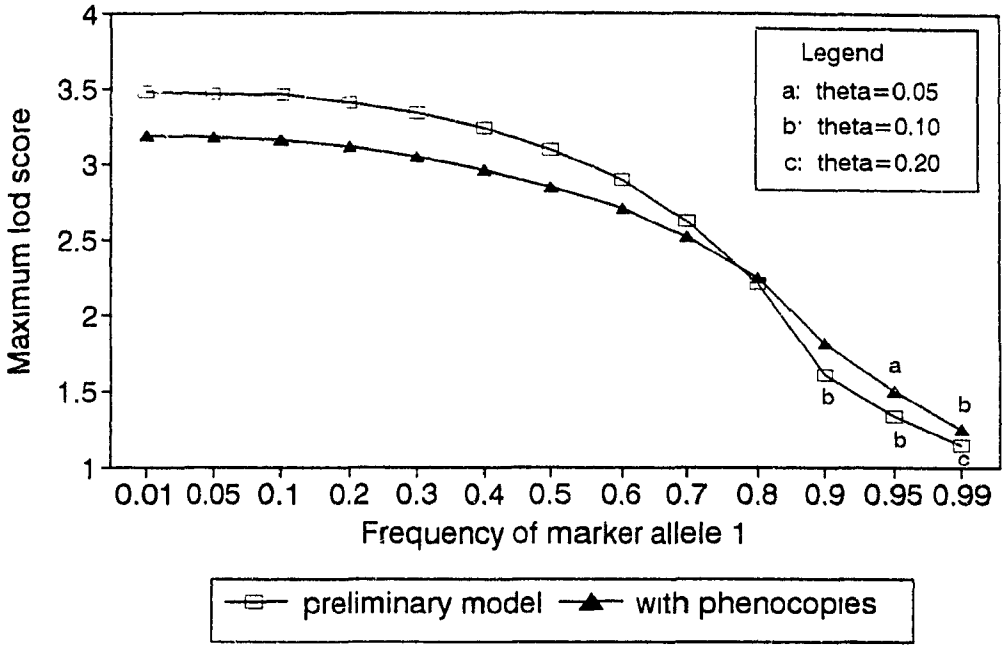


Figure 10. Lod scores for linkage between disease and TNP-1 C as a function of marker allele frequency

that was used in the preliminary model analysis was 0.33, which provided a maximum lod score in a region where the score is decreasing slightly as allele 1 frequency increases.

4.4.4 Consanguinity

Several marriage and consanguinity loops were indicated in the Canadian pedigree (the looped pedigree is displayed in Appendix B4). With respect to the marriage loops, two siblings and two half-siblings entered the pedigree by marriage to offspring of the great-grandparents. Another in-marrying individual was the nephew/niece of two siblings that entered the pedigree. One pair of siblings entered the pedigree by marriage to an offspring of the great-grandparents and his/her nephew/niece. In terms of consanguinity between marriage partners, there were two second cousin marriages.

Figure 11 presents the lod scores for linkage between disease and TNP-1 C with the non-looped and the looped pedigree, under the recessive preliminary model with and without phenocopies. Under the no phenocopies model (62% penetrance for the r/r genotype and a susceptibility allele frequency of 0.2) the maximum lod score decreased from 3.31 at 0% recombination with the non-looped pedigree to 3.14 at 0% recombination when the looped pedigree was analyzed. As the recombination fraction increased, the difference in the lod scores achieved with the two pedigrees decreased so that once the recombination fraction reached 0.1, there was essentially no difference between the two curves.

With the addition of a 50% phenocopy rate in the population, both the non-looped and the looped pedigree provided maximum lod scores of 3.02 at 0% recombination. The curves under the phenocopy models were essentially the same with the two pedigrees; the greatest difference in lod score occurred at recombination fractions of 0.05–0.2 for which scores with the non-looped pedigree were, on average, about 0.17 less than those with the looped pedigree.

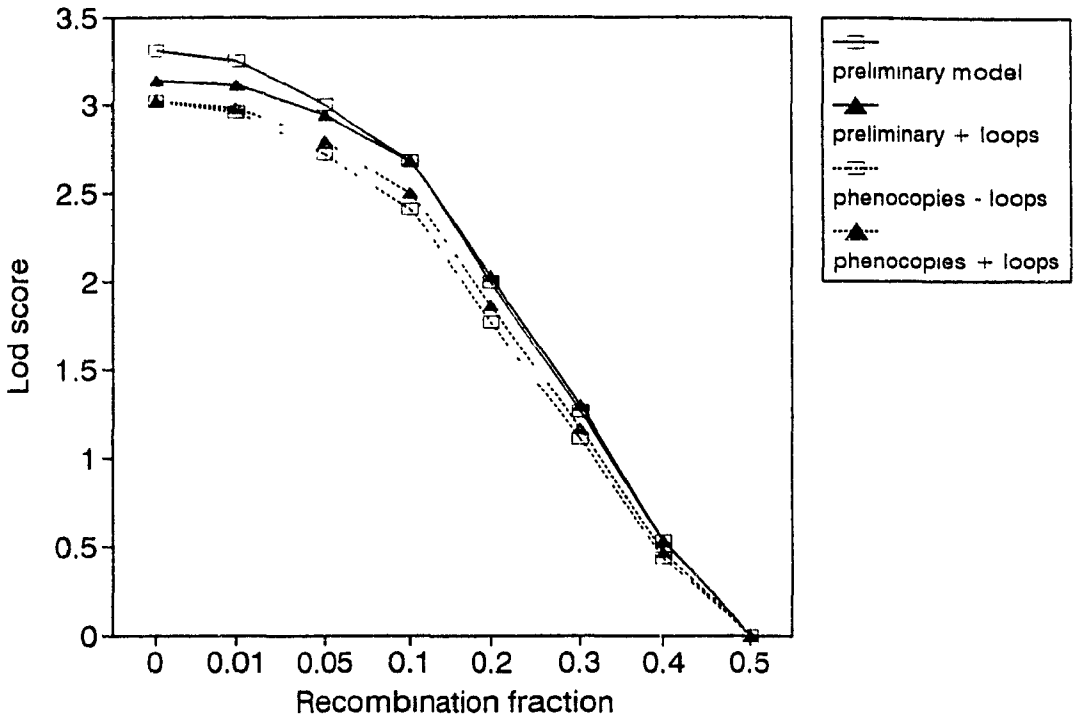


Figure 11. Effect of marriage and consanguinity pedigree loops on lod scores for linkage between disease and TNP-1 C

4.5 Composite marker maps of the chromosome 2q region

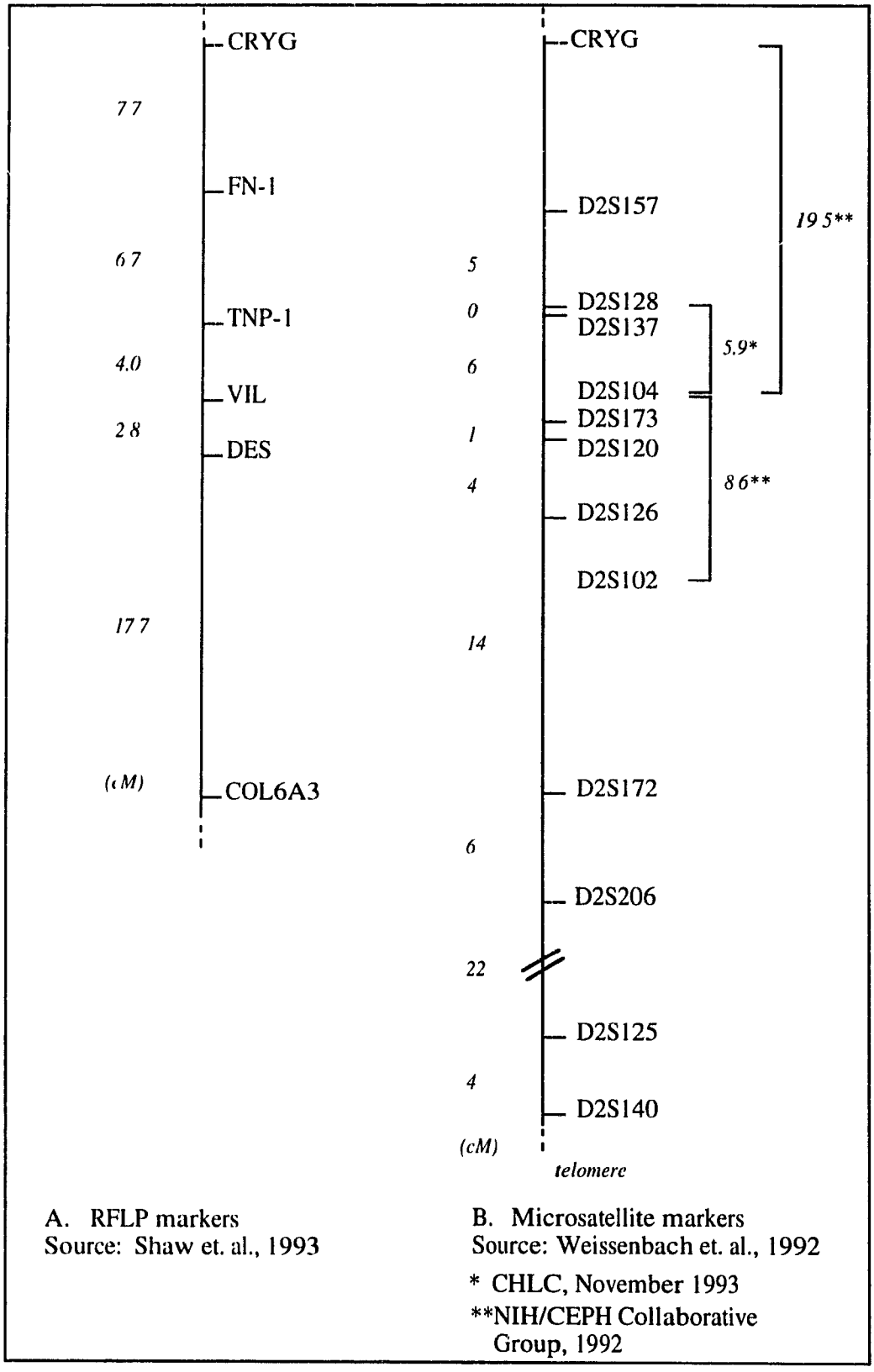
Figure 12 presents sex-averaged[5] linkage maps for the (A) RFLP and (B) microsatellite markers in the chromosome 2q33–qter region, including many of the markers analyzed in this study. The RFLP map generated recently by Shaw et al. (1993) is the most detailed map available for genes encoding products of known function in this region. The gene order and genetic distances were calculated from pairwise linkage analysis of the markers in 35 multiplex families with leprosy, tuberculosis, or visceral leishmaniasis from Pakistan and Brazil (n = 310 persons); recombination frequencies were converted to genetic distances with the Kosambi mapping function. Pairwise lod scores greater than 3 were calculated between adjacent loci in the proximal region of the map (CRYG, FN-1, TNP-1, VIL, and DES). Statistical support for the overall gene order presented is weak, but is consistent with the order found on mouse chromosome 1 (Malo et al., 1993).

In this thesis, results from pairwise linkage analysis of the RFLP markers, carried out with the four Colombian families, 12 Hong Kong families, and the single Canadian family, did not allow a confirmation of the order generated by Shaw et al. (1993) because of a lack of recombination observed between FN-1/TNP-1, TNP-1/VIL, VIL/DES, and DES/VIL loci (data not shown). The only informative result was a recombination fraction of 0.05 between the CRYG and FN-1 loci (a genetic distance of 5 cM, using the Kosambi mapping function; see notes to Table 18 for the formula).

The second map (B) in Figure 12 displays the order and genetic distance between microsatellite markers on the basis of two published genome mapping studies and information provided by the Cooperative Human Linkage Center (CHLC, 1993). The location of the CRYG locus is presented on this map to serve as a link between maps A and B; otherwise, there is little overlap in the maps of these two types of marker loci. Map B was generated with linkage analysis in families in the Centre d'Etudes du

---

[5]The recombination frequency between loci can differ depending on the gender of the informative parent(s) (Ott, 1991); the total length of the female autosomal genetic map is greater than the male map. If potential gender differences are not considered in linkage analysis a sex-averaged recombination fraction is estimated, rather than gender-specific fractions.



A. RFLP markers
Source: Shaw et. al., 1993

B. Microsatellite markers
Source: Weissenbach et. al., 1992

* CHLC, November 1993

**NIH/CEPH Collaborative Group, 1992

Figure 12. Marker maps of the chromsome 2q33-qter region from published sources

Polymorphisme Humain (CEPH) reference panel. The order of these markers with respect to known genes was not published, so that maps A and B could not be integrated from independent publications.

Tables 18 and 19 present pairwise, sex-averaged linkage analysis results of microsatellite markers and RFLP versus microsatellite markers, respectively, analyzed with the Canadian pedigree. Table 18 displays significant results with maximum lod scores > 3 for adjacent markers (except for D2S172 with D2S125, which had a greater score than that between D2S206 and D2S125), in accordance with the order presented in map B of Figure 12. The results were informative for the ordering of some groups of markers in the haplotype analysis of the pedigree, which assumed the minimum number of cross-overs between loci. The microsatellite results are fairly consistent with those obtained in the CEPH families (Figure 12), with the most notable exceptions being the pairwise analysis of CRYG1A/D2S157 and D2S173/D2S120 in the Canadian pedigree, which estimated much smaller and larger genetic distances than the published source, respectively.

Of the significant lod scores (> 3) calculated between RFLP and microsatellite markers in the Canadian family, the most informative results are presented in Table 19. On the basis of the pairwise linkage results found in this study, haplotype construction with microsatellite marker data in the Canadian pedigree, the order of loci on mouse chromosome 1 (Malo et al., 1993), and the human maps presented in Figure 12, a tentative order of the marker loci used in this study was proposed, listed as follows proximal to distal along chromosome 2q:

CRYG–FN-1–D2S157–TNP-1–D2S128–D2S137–D2S104–D2S173–VIL–DES–INHBA1

–D2S120–D2S126–D2S102–PAX-3–D2S172–D2S206–D2S125–D2S140

Late in the progress of this thesis, a human homologue of a candidate mouse gene for *Bcg*, *Nramp*, was isolated (Gros, Cellier et al., unpublished observations). At the time of writing, sequencing of the human gene (NRAMP) and identification of DNA sequence variants had resulted in six markers which were informative for linkage analysis of the Canadian pedigree. Although the results of linkage analysis with these markers are not

Table 18. Pairwise linkage analysis results for microsatellite markers, Canadian family

Markers	Maximum lod score	Recombination fraction (genetic distance in cM*)
CRYG1A and D2S157	12.51	0.0
D2S157 and D2S128	6.85	0.055 (5.5)
D2S128 and D2S137	7.15	0.0
D2S137 and D2S104	5.04	0.039 (3.9)
D2S104 and D2S173	4.36	0.0
D2S173 and D2S120	3.76	0.141 (14.5)
D2S120 and D2S126	8.85	0.043 (4.3)
D2S126 and D2S102	6.94	0.0
D2S102 and PAX-3	8.49	0.0
PAX-3 and D2S172	6.45	0.073 (7.4)
D2S172 and D2S206	8.93	0.041 (4.1)
D2S172 and D2S125	2.01	0.227 (24.5)
D2S125 and D2S140	6.62	0.0

Notes.

*Genetic distance *(x)* in cM was calculated from the maximum likelihood estimate of the recombination fraction *(θ)* between pairs of loci using the Kosambi mapping function, (Ott, 1991):

$$x = 100 \times \tfrac{1}{4} \ln \frac{1 + 2\theta}{1 - 2\theta}$$

Table 19. Pairwise linkage analysis results for RFLP versus microsatellite markers, Canadian family

Markers	Maximum lod score	Recombination fraction (genetic distance in cM*)
FN61039-H and CRYG1A	7.04	0.025 (2.5)
FN61039-H and D2S157	6.67	0.031 (3.1)
TNP-1 B and CRYG1A	5.33	0.0
TNP-1 C and D2S157	4.19	0.047 (4.7)
TNP-1 C and D2S128	5.82	0.0
TNP-1 C and D2S137	5.17	0.0
INHBA1 and D2S173	3.07	0.09 (9.1)

Notes.
*Recombination fraction converted to genetic distance with the Kosambi mapping function (see notes to Table 18 for formula)

included in this thesis, haplotypes were constructed with the marker data to infer the location of the NRAMP gene and closely-linked markers on the chromosome 2q marker map above. The VIL, DES, and INHB loci were placed in the interval between D2S173 and D2S120 on the basis of ongoing physical mapping with markers of NRAMP, which VIL and DES are expected to flank, in yeast artificial chromosome (YAC) clones (Liu, unpublished observations). The results of the haplotype analysis in the Canadian pedigree were consistent with NRAMP localization between D2S173 and D2S120.

4.6 Lod scores for linkage between disease and markers under the epidemiological model, Canadian family

The lod score results for linkage between disease and the RFLP or microsatellite markers under the comprehensive epidemiological model (four liability classes, $q = 0.4$) are presented in Tables 20a and 20b. Non-informative analyses, for which lod scores at all recombination fractions were essentially equal to zero, are not displayed in these tables. With the non-looped pedigree, evidence in favour of close linkage was obtained

Table 20a. Lod score results under the comprehensive epidemiological model (four liability classes), Canadian family

	Non-looped pedigree		Looped pedigree	
Marker	Z max (theta)	Z min (theta)	Z max (theta)	Z min (theta)
CRYG1-A	1.11 (0.15)	*	1.29 (0.1)	*
CRYG1-D	0.98 (0.15)	*	1.25 (0.1)	*
CRYGP1-C	*	-0.39 (0)	*	-0.44 (0)
FN61039-H	0.34 (0.2)	-0.18 (0)	0.31 (0.2)	-0.31 (0)
FN61039-M	*	-0.11 (0)	*	-0.20 (0)
D2S157	0.88 (0.15)	*	n.d.	n.d.
TNP-1 B	1.94 (0)	*	1.80 (0.01)	*
TNP-1 C	2.43 (0)	*	2.33 (0)	*
TNP haplotype	2.32 (0)	*	n.d.	n.d.
D2S128	1.58 (0)	*	n.d.	n.d.
D2S137	0.76 (0.15)	*	n.d.	n.d.
D2S104	*	-0.55 (0)	0.02 (0.35)	-0.32 (0)
D2S173	0.36 (0.2)	-0.16 (0)	n.d.	n.d.

Notes.

* The relevant lod score was 0 at a recombination fraction of 0.5.

n.d. Analysis not done

n/a Not applicable

Lod score results are presented by marker in the tentative order suggested in section 4.5 proximal to distal along chromosome 2q. Markers of the same locus are displayed in alphabetical order. Results for non-informative markers with the non-looped pedigree are not shown.

Table 20b. Lod score results under the comprehensive epidemiological model (four liability classes), Canadian family (continued)

	Non-looped pedigree		Looped pedigree	
Marker	Z max (theta)	Z min (theta)	Z max (theta)	Z min (theta)
DESMIN	*	-0.31 (0)	*	-0.40 (0)
INHB-A1	*	-0.84 (0)	*	-0.74 (0)
D2S120	0.35 (0.2)	-0.05 (0)	n.d.	n.d.
D2S126	1.14 (0.05)	*	n.d.	n.d.
D2S102	*	-0.74 (0)	*	-0.88 (0)
PAX-3	0.13 (0.25)	-0.48 (0)	n.d.	n.d.
D2S172	*	-0.75 (0)	n.d.	n.d.
D2S206	*	-1.64 (0)	n.d.	n.d.
D2S125	*	-1.29 (0)	n.d.	n.d.
D2S140	*	-1.12 (0)	*	-1.33 (0)

Notes.

* The relevant lod score was 0 at a recombination fraction of 0.5.

n.d. Analysis not done

n/a Not applicable

Lod score results are presented by marker in the tentative order suggested in section 4.5 proximal to distal along chromosome 2q. Markers at the same locus are displayed in alphabetical order. Results for non-informative markers with the non-looped pedigree are not shown.

with the TNP-1 C marker and the TNP-1 haplotype (composed of the TNP-1 A, B, and C markers), with odds for linkage at 0% recombination of approximately 210:1 ($10^{2.32}$) and 270:1 ($10^{2.43}$), respectively. The microsatellite markers which provided the most support for linkage were CRYG1-A, D2S128, and D2S126, with respective lod scores of 1.11 at 15% recombination, 1.58 at 0% recombination, and 1.14 at 5% recombination. Linkage analysis of disease and other markers computed weakly positive maximum lod scores (range, 0.13–0.98) for CRYG1-D, FN61039-H, D2S157, D2S137, D2S173, D2S120, PAX-3, and D2S206. The rest of the markers provided weak evidence against linkage with minimum lod scores at 0% recombination ranging from -0.11 to -1.64. The chromosomal region which was the most supportive for linkage, then, was closer to the TNP-1 locus than to the DES and INHB loci.

With the looped pedigree, the lod scores for the markers analyzed showed minimal change when compared to the results with the non-looped pedigree. In general, slightly more evidence was provided against linkage for markers with negative lod scores at 0% recombination, with the exception of D2S104 and INHB-A1. For the markers with strongly positive lod scores in favour of linkage in the non-looped pedigree, the maximum lod scores increased slightly for CRYG1-A and CRYG1-D and decreased marginally for TNP-1 B and C. This indicated that the additional phase information, although minor, provided further support for linkage of disease and the CRYG locus and was less supportive of linkage to the TNP-1 locus.

SECTION 5: DISCUSSION

5.1 Colombian and Hong Kong Families

When the lod scores for linkage between markers and disease in the four Colombian and 12 Hong Kong families were summed, the results under the preliminary model (no phenocopies) suggested that a tuberculosis susceptibility gene is not located in the close proximity of the TNP-1, ViL, DES, or INHB loci. These loci define the region in which a homologous human "BCG" gene might be expected to be located on the basis of the experimental mouse *Bcg* model. Thus, when the affected phenotype corresponded to at least one diagnosis of tuberculosis in family members from regions of

endemic tuberculosis experience, there was evidence against linkage of the trait to chromosome 2q, under the simple Mendelian models considered (recessive single gene; $q = 0.2$; marker allele frequencies estimated in Caucasian populations; penetrance of the higher risk r/r genotype 62% or 76%; with or without a 50% phenocopy rate in the population). The higher penetrance level (76%), based on monozygous twin tuberculosis concordance adjusted for ascertainment of diseased index twins in the Kallmann and Reisner (1943) study (Appendix C), increased the evidence against linkage to markers on chromosome 2q.

Linkage analysis of the Colombian and Hong Kong pedigrees was also carried out with the addition of diagnostic uncertainty (as indicated by the international collaborators) to the preliminary, no phenocopy model (62% penetrance for the r/r genotype) in order to evaluate the linkage information more conservatively. Some misclassification of phenotype may have occurred because of a lack of microbiologic confirmation of disease (Appendix B5). Linkage to the disease phenotype was significantly excluded to within 10% recombination of TNP-1 C and 1% recombination of two of the VIL markers, which were expected to be the closest to the putative "BCG" susceptibility gene on the basis of human-mouse homology.

The genetic models of tuberculosis susceptibility used in the linkage analysis of the international families may not be correct; linkage to the chromosome 2q region may exist for a different trait, that is, for a different definition of susceptibility. The affected phenotype of at least one tuberculosis diagnosis may not have been appropriate because many other factors, besides a susceptibility gene, may have contributed to the outcome of exposure to *M. tuberculosis*, such as intensity of infection, nutritional status, concomitant disease, age, and vaccination or chemoprophylactic status. As a result of these other factors, many of the affected individuals in the Colombian and Hong Kong families may represent non-genetic cases and may be interpreted as recombination under models with no allowance for phenocopies. With the preliminary phenocopy model, lod scores became less negative but did not reach substantial positive values, and linkage to markers of the TNP-1 and VIL loci was still excluded. These results indicate that the evidence against linkage was not solely due to the occurrence of non-genetic cases. On

the other hand, the risk of disease for some family members may have been reduced significantly, if BCG vaccination has protected individuals from disease progression regardless of genetic susceptibility at the putative chromosome 2q locus. It is reasonable, therefore, for more negative lod scores to be calculated under a model with higher penetrance (76%).

A major difficulty in analyzing tuberculosis families, then, is to characterize separately the early control of mycobacterial proliferation by macrophage function, thought to be regulated by a human chromosome 2q gene which is homologous to *Bcg*, from other, later determinants of the host immune response to *M. tuberculosis* infection. In particular, cell-mediated immunity and delayed-type hypersensitivity, which follow the early response and are likely affected by BCG vaccination and prior exposure, may have a more significant role in determining the outcome of infection in a population with endemic tuberculosis experience. Alternatively, the effect of the putative "BCG" gene may still be important, but is harder to detect in the presence of other factors in the international families: those cases that resulted from susceptibility at the chromosome 2q locus may represent only a small proportion of all affected individuals.

A different measure of susceptibility in such families may allow detection of the macrophage response to *M. tuberculosis* infection. For example, an immunological quantification of the early response, rather than the disease/non-disease outcome, might be more powerful for detection of linkage to the chromosome 2q region. In this way, a biological assay of suboptimal macrophage activity or extensive bacillary growth could be used to indicate a susceptible phenotype, or measures of such factors could be used as quantitative traits. Such assays, more similar to those in the mouse *Bcg* model, may be less likely to be affected by other host characteristics such as vaccination status. A difficulty with this approach would be to control for variation in intensity of infection and to make measurements at the appropriate time after exposure in order to reflect the initial response to mycobacteria.

In addition to the use of an assay measure, the lod score analysis of families similar to the international pedigrees may require the development of a more accurate model in order to detect linkage, assuming that the assumption of a single, major

susceptibility gene can be made. A more appropriate scheme would likely involve the modelling of penetrance to take into account the operation of non-genetic factors more precisely. For instance, the interval between exposure and disease could be taken into account using liability classes. In general, there is a 5% risk of disease in the first two years following infection and a 10% lifetime risk (Styblo, 1989). Other risk factors include age at exposure, vaccination and chemoprophylactic status, as well as concomitant disease. Without (1) extensive information on dates of preventive procedures and personal histories such as HIV status and (2) quantification of the combined effects of these factors on disease risk, it is unlikely that any arbitrary model will accurately reflect the relative contributions of genetics and other host and environmental factors to the risk of tuberculosis.

The mouse *Bcg* locus on chromosome 1 regulates macrophage function and the early phase of resistance/susceptibility to various mycobacterial infections, including *M. bovis* and *M. leprae*. Exclusion of linkage between human susceptibility to leprosy and RFLP markers in the chromosome 2q33–q37 region has recently been reported in a study of 17 multiplex families from regions with high leprosy prevalence, including Pakistan (10 families) and Brazil (seven families) (Shaw et al., 1993). Under a recessive model for susceptibility to leprosy *per se*, with a susceptibility allele frequency of 0.1, marker allele frequencies estimated in the study families, no phenocopies, and 100% penetrance for the r/r genotype, linkage with disease was significantly excluded with the lod score method to 10–20 cM of the CRYGP1, TNP-1, VIL, and DES loci and to 5–10 cM of the FN-1 locus. Significant evidence against linkage to the TNP-1, VIL, and DES loci was also obtained under recessive models of incomplete penetrance (80% and 60%) and by using the affected pedigree member method of linkage analysis. The negative lod score results were due to several families for each locus: five families in the case of TNP-1, seven for VIL, and 10 families for DES. Close linkage to these loci was also excluded for the following models with the lod score method: dominant susceptibility to leprosy *per se* (TNP-1), and dominant or recessive susceptibility to the tuberculoid form of leprosy (TNP-1 and TNP-1, VIL, and DES, respectively).

Shaw et al. (1993) suggest that the evidence against linkage in their study is consistent with the fact that the affected phenotypes considered were based on cutaneous manifestations of leprosy, and such manifestations are not influenced by genotype at the *Bcg* locus in mouse studies. Therefore, even the phenotype of "leprosy *per se*" may represent a specific type of disease, which is not expected to be linked to the "BCG" susceptibility locus, as suggested by human association and segregation studies (Bothamley et al., 1989; Abel and Demenais, 1988). It was stated by Shaw et al. (1993) that more variable lod scores have been generated in multiplex tuberculosis families in ongoing research, and a proportion of these families may provide evidence for linkage to chromosome 2q; however, no specific results were given in this paper. I would speculate at this point that the general phenotype of tuberculous "disease" versus "no disease", defined by traditional clinical tests, is not an appropriate phenotype for linkage analysis of families from endemic regions. Because of the likely modification of the "BCG" effect by host cell-mediated immunity, age, and vaccination, treatment, and exposure history, positive results in such families may be chance findings or may only be observed infrequently, when the levels of these other factors have less variation and/or less effect in a particular pedigree.

The results with the Hong Kong and Colombian families are based on a maximum of six and 10 informative families per marker analyzed, under models with and without phenocopies, respectively. For some markers, neither parent in a particular family was inferred to be doubly heterozygous for the trait and the marker. The lack of informativeness related in part to the type of family ascertained: because one parent was most often affected, he/she was homozygous for the susceptibility allele under a recessive model with no phenocopies. The results thus apply to a small number of families of particular structure, and may not be relevant to other families in the same or other populations. A greater number of families and a greater variety of family types (structurally as well as racially) should be collected to replicate the negative findings of the Colombian and Hong Kong families in this study.

The HOMOG programs carry out statistical tests for the presence of genetic heterogeneity in a group of families analyzed for linkage (Ott, 1991). The programs are

intended for this purpose when the same disease phenotype is analyzed in all the families. There were essentially two groups of families in this study, with disparate tuberculosis experience and possibly disease outcomes of a different nature: in the Canadian family, primary disease developed after a short period of time, whereas the exposure-disease intervals for the international cases (many of which may represent reinfection or reactivation disease) are indeterminate. For these reasons, a statistical test of genetic heterogeneity was not carried out on the linkage analysis results. The large differences in the lod scores provided at the same loci, under the preliminary models, by the international families versus the Canadian family may reflect the different nature of the tuberculosis experience (including exposure and vaccination status) in the two groups. It is possible that a second, single, non-"BCG" gene is linked to disease susceptibility in the Colombian and Hong Kong families. More realistically, however, there may be a number of major genes contributing to the disease outcome in these pedigrees.

5.2 Canadian family

Under the preliminary models (with or without phenocopies), significant evidence for linkage with disease--with a lod score greater than the traditional critical value of 3--was obtained with the TNP-1 C marker at a recombination frequency of 0%, suggesting that a susceptibility gene might exist in the vicinity of the TNP-1 locus. The approximate 95% confidence interval for the recombination frequency, under the preliminary model (no phenocopies) and based on 1-lod unit support (Ott, 1991), was 0–15% or 0–15.5 cM from the TNP-1 locus, using the Kosambi mapping function. Several markers close to TNP-1 C also provided positive lod scores of about 2 under the preliminary model (with and without phenocopies), including TNP-1 B, D2S128, and CRYG1-A. Evidence against linkage of disease and the VIL E-84 marker suggested that the susceptibility gene was located closer to the TNP-1 locus than to the VIL locus.

The linkage results in the Canadian family were, not surprisingly, greatly affected by substantial uncertainty of the clinical diagnosis. When estimates of the positive and negative predictive values of the diagnoses were considered, the latter particularly low for 28% of the pedigree members (i.e., the non-cases over 13 years), the lod score for linkage between disease and TNP-1 C was markedly decreased from 3.31 to 0.83, at

recombination fractions of 0%. On the other hand, the examination of the relationships of penetrance, susceptibility allele frequency, and marker allele frequency with the lod score results for disease and TNP-1 C revealed that the scores were quite robust to alteration of the values of these parameters within a reasonable range. These observations, however, were made by modifying the value of one parameter at a time and do not address the simultaneous effect of alterations in more than one parameter value. The addition of consanguinity and marriage loops to the pedigree modified the lod score results only slightly, because most of the more distant relatives in the loops had unknown clinical status and unknown marker typing. With the loops, the evidence for linkage of disease and TNP-1 C decreased marginally under the preliminary model (with and without phenocopies) and under the comprehensive epidemiological model (four liability classes). The use of microsatellite marker allele frequencies estimated in the pedigree founders was not the most appropriate when the looped pedigree was analyzed; however, this procedure likely did not significantly bias the results because the relative order of the alleles by frequency would not be expected to greatly change.

The single liability class, preliminary models (with and without phenocopies) and the diagnostic uncertainty model made the assumption that all members of the Canadian linkage analysis pedigree were infected during the epidemic or before, and thus had the opportunity to express disease. When analyzed under the most comprehensive epidemiological model that included phenocopies and four liability classes (with allowance for absence of exposure), the maximum lod scores with disease were 2.43 for TNP-1 C and 2.32 for the TNP-1 haplotype, both at 0% recombination. These may be more "realistic" results, given the conditions of protection and liability provided by the exposure variation and young age in the pedigree. Information arising from the infant cases in the family was greatly de-weighted under this model. The model, then, did not make the assumption that all non-cases had been infected during the 1987–1989 epidemic but rather incorporated clinical information about PPD-negative skin tests, as well as previous mycobacterial exposure and known BCG vaccination.

The linkage analysis in this study was targeted at a candidate chromosomal region, based on the mouse *Bcg* model and evolutionary homology (that is, the demonstration of

conserved homology between a region on mouse chromosome 1 and human chromosome 2q). If the critical value for a significant lod score is relaxed because a candidate genetic region with higher prior odds for linkage was used (Wijsman, 1990), then the lod scores provided under the comprehensive epidemiological model in the Canadian pedigree are likely statistically significant. Recalling that the posterior odds for linkage or the significance level of a result (1 – type I error) is equal to the prior odds for linkage multiplied by the odds provided by the data, the lod score of 2.32 for the TNP-1 haplotype ($10^{2.32}$ odds for linkage) has a posterior odds of 20:1, a 95% significance level, and a 5% type I error, if the prior odds for linkage to the region are equal to about 0.1, or 5 times the prior odds for a randomly chosen locus (which is about 0.02; see section 1.3.9).

The positive results in the Canadian family were obtained for the phenotype of primary tuberculosis following heavy exposure to *M. tuberculosis*. Most members of the family were exposed to tuberculosis for the first time during an epidemic, and developed disease within a short time. Three individuals were assigned the status of "affected" because of a diagnosis of primary tuberculosis prior to the 1987–1989 outbreak, at a time when the disease was very prevalent in the community and exposure to infection may have been intense (Mah and Fanning, 1991). For a host infected with mycobacteria, the early immune response, which may determine the extent of the bacterial load that cell-mediated immunity must manage, could be crucial: if deficient, delayed-type hypersensitivity may be detrimental in the presence of high levels of antigen, leading to tissue destruction and a greater risk of disease progression (Dannenberg, 1989). The early immune response may be most important when exposure to *M. tuberculosis* is intense and the host does not have pre-existing protective immunity against mycobacteria.

The outbreak situation and the diagnoses of primary disease in the Canadian family may have allowed the independent characterization of the effects of the early immune phase from the later cell-mediated stages of the host response to infection. It has been possible, therefore, to use an experimental model developed in the mouse to form, address, and support a hypothesis for the human response to mycobacterial infection, using homology mapping between the murine and human genome. This method of

research will be supported even further if evidence is found for linkage of tuberculosis susceptibility and a human candidate "BCG" gene. At that point, researchers may be able to gain more insight into the mechanism by which the early immune response can control mycobacterial growth and, conversely, the role of host genotype and immune function in permitting chronic infection by *M. tuberculosis* (Marrack and Kappler, 1994).

The lod score results with the Canadian family were obtained in a pedigree that contained a subset of the family members, because nine persons were not enrolled in the previous linkage analysis and diagnostic study for which the family was selected. In addition, DNA was not available for several members in this study. In a large pedigree, lod score results can be particularly sensitive to changes in information (e.g., from unknown to known marker typing) (Greenberg, 1992). Despite the robustness of the lod scores in this study with respect to variation in values of penetrance and allele frequencies, marker typing was missing for a few family members who were critical persons in the linkage analysis, based on their clinical status and positions in the pedigree. The results, therefore, could be modified in either a positive or negative direction if these key individuals were typed.

The positive findings of this study are at best generalizable to Aboriginal Canadian populations of similar ethnicity and with a similar history of contact with tuberculosis as the Canadian study family. Of course, the type I error possibility exists. Confirmation of the results in other, independently ascertained families will be important to assess the validity and applicability of the findings. A starting point for collection and linkage analysis of additional families could be other communities with experience of a tuberculosis outbreak. Because of the nature of many public health investigations during and following an outbreak, there may be greater opportunity, in a study of epidemic experience, to assess time of infection, exposure intensity during the outbreak, latency of disease development, and possibly status (vaccination, chemoprophylactic, and general health), age, and exposure history prior to the outbreak. Genetic study of such experience is likely facilitated by many of the factors which may have contributed to the occurrence of the outbreak in the first place, particularly lack of previous exposure or BCG vaccination.

5.3 Linkage analysis of tuberculosis

This thesis contained a combination of epidemiological and genetic data and utilized genetical statistical methods to address the research question. When the study families were categorized into two groups on the basis of the nature of their tuberculosis experience, both negative and positive evidence with respect to linkage between disease and chromosome 2q markers was obtained. The results of this study should be interpreted with the specific materials and genetic models used in mind. The linkage analysis was complicated by the lack of knowledge about a genetic model for inheritance of tuberculosis susceptibility in humans and the role of other non-genetic factors in disease risk. In addition, the genetic models used in the study made the assumption of a single gene effect, although tuberculous disease susceptibility may be inherited in a more complex manner, perhaps involving several interacting genes.

Past studies of linkage analysis have demonstrated that the misspecification of the required parameter values, particularly mode of inheritance, can decrease the power to detect true linkage (Risch et al., 1989). It was shown by Ott (1992) that the use of inaccurate marker allele frequencies can lead to false positive results when families with incompletely typed parents are analyzed. Results from a recent workshop (Genetic Analysis Workshop 8, 1992), in which various linkage analysis methods were used to study Alzheimer's disease, indicate that false positive evidence for linkage can arise when the value of one or several parameters, at the disease or marker locus, is unknown and misspecified (Wijsman, 1993). The significant lod scores in the Canadian pedigree, if indeed false, may be the result of a combined parameter misspecification effect, since the scores were fairly robust to alteration of one parameter at a time.

Clearly, an appropriate genetic model of susceptibility is crucial for valid linkage analysis of a complex disease with the lod score method. If incorrect models are used, false negative or false positive lod score results can arise. The estimation of the significance level of an observed positive lod score (maximized-over-models) with computer simulation in unlinked pedigrees, as carried out in this thesis (Appendix E), will by no means guarantee a valid evaluation of the finding if the models are incorrectly specified. A significant p-value may be obtained for a false positive result if lod scores

tend to be underestimated under most of the models considered, due to parameter misspecification.

Linkage analysis of tuberculosis and the study of genetic susceptibility to the disease in general could benefit from segregation analysis, which was not available for this study. Segregation analysis may be able to indicate whether there is evidence for a major gene contributing to disease risk, and may help establish a genetic model for inheritance of susceptibility. In a similar manner to the segregation analysis of leprosy in Desirade Island by Abel and Demenais (1988), the study population could be one in which tuberculosis has been diagnosed for a number of family members and the study participants are known or can be assumed to have been infected with *M. tuberculosis*. The likelihood of the data under various models (such as random environmental, multifactorial with a transmissible component, Mendelian single gene, or a mixed model incorporating all three effects) can be compared to determine which models are rejected on statistical grounds. If, for instance, evidence is found for a particular major gene model, linkage analysis with the lod score, single-locus method can then be carried out using parameter estimates from the segregation study. The parametric analysis of linkage could thus incorporate a more secure model of inherited susceptibility.

To confirm or refute the findings of this study, it would also be useful to collect numerous families for non-parametric linkage analysis, such as small families with affected sib-pairs. A method of analysis which may indicate whether a particular marker locus is a linked determinant of disease has been developed by Risch (1987). He demonstrates that for a single disease or susceptibility locus closely linked to a marker locus, the posterior probability that two relatives share zero marker alleles identical by descent, given that both are affected with the disease, is equal to the prior probability that the pair share no alleles identical by descent, divided by the increased disease risk for the relatives over population disease prevalence. This relationship holds for any mode of inheritance, and any number, frequency, population prevalence, and penetrance of alleles at the disease locus. The prior probabilities of zero allele sharing by full siblings is, for example, 0.25. Given an estimate of the increased risk of disease for two relatives from, for example, the Kallmann and Reisner (1943) twin family study (keeping in mind the

possible biases in this kind of study), the expected proportion of affected relatives sharing zero alleles at a particular marker can be calculated. With marker data on affected pairs (and information on parents), one can determine whether the observed sharing of zero alleles identical by descent is likely, given the expected proportion under the single, linked gene assumption. If the observed data is very unlikely, the implication is that there is an unlinked determinant of disease.

Using the sib-pair example and disease risk estimates from Kallmann and Reisner (1943), the increased sibling risk for tuberculosis can be estimated as 18.9/1.1 or 25.5/1.4, depending on whether crude risks or risks adjusted for age differences are used, respectively. The posterior probability of two affected siblings sharing zero alleles is thus 0.25/17.2 or 0.25/18.2, respectively, or about 0.014. Knowing this expected proportion, marker data could be collected for affected sibling pairs and used to determine whether the single locus model of susceptibility can be rejected. At the same time, the effect of a particular marker locus on disease risk can be measured using the observed sharing of zero alleles and the simple prior probabilities of such sharing by two relatives. Risch (1987) also developed a method to assess the evidence for an undetected linked allele using a model of two susceptibility loci with a multiplicative effect (unlinked to each other), haplotype discordance data from affected relatives, and genotype penetrances from association studies.

For further parametric and non-parametric linkage analysis of tuberculosis, markers in the close vicinity of the TNP-1 locus and in the human homologue of the mouse *Nramp* gene should be a priority. The same markers could be used for an association study. In order to study susceptibility to tuberculosis with linkage or association analysis, it is clear that the study sample should ideally be selected in such a way as to avoid differences between cases and non-cases with respect to the following: exposure to *M. tuberculosis*; non-genetic risk factors for tuberculosis; prior BCG vaccination; and chemoprophylaxis. If non-cases were not infected, were vaccinated, or received chemoprophylaxis, or if cases had other risk factors for tuberculosis not present among the non-cases, linkage or association could be missed (increased type II error) or false positive evidence could arise (increased type I error).

5.4 Conclusions

1. Linkage analysis of a large, multiplex, Aboriginal Canadian family provided evidence for linkage of a tuberculosis susceptibility gene to the TNP-1 locus on chromosome 2q, using a single-locus, incompletely penetrant susceptibility trait. This result was robust to variation in susceptibility allele and marker allele frequencies and penetrance, within reasonable ranges of these parameters. Linkage to the same region was significantly excluded for the combined results of 16 smaller families from Colombia and Hong Kong.
2. The effect of a tuberculosis susceptibility gene may be more important or easier to detect in a predominantly newly-infected family with minimal BCG vaccination and exposure to an intense, short-term epidemic. Conversely, the study of families from regions with endemic tuberculosis and vaccination coverage may not be as useful to investigate the early host response to infection with *M. tuberculosis*.
3. The results of the lod score method of linkage analysis are sensitive to errors in diagnosis and marker typing. Finally, for the study of a complex disease such as tuberculosis, careful choice of families and appropriate genetic models are crucial.

References

Abel, L., Demenais, F., Baule, M-S., Blanc, M., Muller, A., Raffoux, C., Millan, J., Bois, E., Babron, M-C., and Feingold, N., 1989. Genetic susceptibility to leprosy on a Caribbean island: linkage analysis with five markers. International Journal of Leprosy 57 (2): 465-471.

Abel, L. and Demenais, F., 1988. Detection of major genes for susceptibility to leprosy and its subtypes in a Caribbean island: Desirade Island. American Journal of Human Genetics 42: 256-266.

Alun, T., 1991. Pedpack 3.0. School of Mathematical Sciences, University of Bath, Bath, England.

Armitage, P. and Berry, G., 1987. Statistical Methods in Medical Research. 2nd ed. Oxford, England: Blackwell Scientific Publications.

Ayvazian, L.F., 1993. "History of tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 1. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 1-20.

Bailey, D.M.D., Affara, N.A., and Ferguson-Smith, M.A., 1992. The X-Y homologous gene amelogenin maps to the short arms of both the X and Y chromosomes and is highly conserved in primates. Genomics 14: 203-205.

Baird, P.A., 1990. Genetics and health care: a paradigm shift. Perspectives in Biology and Medicine 33 (2): 203-213.

Barnes, P.F., Le, H.Q., and Davidson, P.T., 1993. Tuberculosis in patients with HIV infection. Medical Clinics of North America 77 (6): 1369-1389.

Baron, M., Endicott, J., and Ott, J., 1990. Genetic linkage in mental illness - limitations and prospects. British Journal of Psychiatry 46 (4): 917-940.

Bass, J.B.Jr., 1993. "The tuberculin test." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 7. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 139-148.

Bates, J.H. and Stead, W.W., 1993. The history of tuberculosis as a global epidemic. Medical Clinics of North America 77 (6): 1205-1217.

Belknap, H.R. and Hayes, W.G., 1961. A genetic analysis of families in which leprosy occurs. Abstract. Leprosy in India 29: 375.

Black, F.L., 1992. Why did they die? Science 258: 1739-1740.

Bloom, B.R., 1993. The critical importance of research for global TB control. Abstract presented at the World Congress on Tuberculosis, Bethesda, Maryland. 16-19 Nov. 1992.

Bloom, B.R. and Murray, C.J.L., 1992. Tuberculosis: commentary on a reemergent killer. Science 257: 1055-1064.

Borch-Johnsen, K. and Sorensen, T.I.A., 1993. Genes and environment in the inheritance of morbidity and mortality. Acta Psychiatrica Scandinavica Supplement 370: 73-78.

Bothamley, G.H., Beck, J.S., Schreuder, G.M.Th., D'Amaro, J., de Vries, R.R.P., Kardjito, T., and Ivanyi, J., 1989. Association of tuberculosis and *M. tuberculosis*-specific antibody levels with HLA. The Journal of Infectious Diseases 159 (3): 549-555.

Brahmajothi, V., Pitchappan, R.M., Kakkanaiah, V.N., Sashidhar, M., Rajaram, K., Ramu, S., Palanimurugan, K., Paramasivan, C.N., and Prabhakar, R., 1991. Association of pulmonary tuberculosis and HLA in South India. Tubercle 72: 123-132.

Breitner, J.C.S., Gatz, M., Bergem, A.L.M., Christian, J.C., Mortimer, J.A., McClearn, G.E., Heston, L.L., Welsh, K.A., Anthony, J.C., Folstein, M.F., and Radebaugh, T.S., 1993. Use of twin cohorts for research in Alzheimer's disease. Neurology 43: 261-267.

Bueckert, D., 1992. Deadly TB strain hits Canada. 15 Aug. 1992. The Toronto Star: A1.

Centers for Disease Control, 1991. Tuberculosis morbidity in the United States: final data, 1990. Morbidity and Mortality Weekly Report 40 (SS-3): 23-27.

Centers for Disease Control, 1990. Tuberculosis in developing countries. Morbidity and Mortality Weekly Report 39 (33): 561-569.

Chakravarti, A. and Lander, E.S., 1990. "Genetic approaches to the dissection of complex diseases." Genetics and Biology of Alcoholism. Cloninger, C.R. and Begleiter, H., eds. Banbury Report 33. U.S.A.: Cold Spring Harbor Laboratory Press. 307-312.

Christie, A.B., 1987. "Tuberculosis: non-tuberculous mycobacteriosis." Infectious Diseases. Chapter 17. 4th ed. Vol. 1. New York: Churchill Livingstone. 492-540.

Collins, F.M., 1993. Tuberculosis: the return of an old enemy. Critical Reviews in Microbiology 19 (1): 1-16.

Comstock, G.W., 1978. Tuberculosis in twins: a re-analysis of the Prophit survey. American Review of Respiratory Disease 117: 621-624.

Comstock, G.W., 1975. Frost revisited: the modern epidemiology of tuberculosis. American Journal of Epidemiology 101 (5): 363-382.

Comstock, G.W. and Cauthen, G.M., 1993. "Epidemiology of tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 2. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 23-48.

Comstock, G.W. and O'Brien, R.J., 1991. "Tuberculosis." Bacterial Infections of Humans: Epidemiology and Control. Evans, A.S. and Brachman, P.S., eds. Chapter 36. 2nd ed. New York: Plenum Medical Book Company. 745-771.

Connor, J.M. and Ferguson-Smith, M.A., 1987. Essential Medical Genetics. 2nd ed. Oxford: Blackwell Scientific Publications.

Crowe, R.R., 1993. Candidate genes in psychiatry: an epidemiological perspective. American Journal of Medical Genetics (Neuropsychiatric Genetics) 48: 74-77.

Dannenberg, A.M., 1989. Immune mechanisms in the pathogenesis of pulmonary tuberculosis. Reviews of Infectious Diseases 11 (Supplement 2): S369-S378.

De Cock, K.M., Soro, B., Coulibaly, I.M., and Lucas, S.B., 1992. Tuberculosis and HIV infection in sub-Saharan Africa. Journal of the American Medical Association 268 (12): 1581-1587.

Denis, M., Forget, A., Pelletier, M., Gervais, F., and Skamene, E., 1990. Killing of *Mycobacterium smegmatis* by macrophages from genetically susceptible and resistant mice. Journal of Leukocyte Biology 47: 25-30.

Des Prez, R.M. and Heim, C.R., 1990. "Mycobacterium tuberculosis." Principles and Practice of Infectious Diseases. Mandell, G.L., Douglas, R.G.Jr., and Bennett, J.E., eds. Chapter 229. New York: Churchill Livingstone. 1877-1906.

DOE Human Genome Program, 1992. Primer on Molecular Genetics. Washington: United States Department of Energy.

Dowling, P.T., 1991. Return of tuberculosis. American Family Physician 43: 457-467.

Dunlap, N.E., and Briles, D.E., 1993. Immunology of tuberculosis. Medical Clinics of North America 77 (6): 1235-1251.

Elandt-Johnson, R.C., 1971. Probability models and statistical methods in genetics. New York: John Wiley.

Elston, R.C. and Lange, K., 1975. The prior probability of autosomal linkage. Annals of Human Genetics 38: 341-350.

Enarson, D.A. and Grzybowski, S., 1986. Incidence of active tuberculosis in the native population of Canada. Canadian Medical Association Journal 134: 1149-1152.

Epstein, D.J., Vekemans, M., and Gros, P., 1991. Splotch (Sp^{2H}), a mutation affecting development of the mouse neural tube, shows a deletion within the paired homeodomain of *Pax-3*. Cell 67: 767-774.

Felton, C.P., Smith, J.A., and Ehrlich, M.H., 1990. Racial differences and *Mycobacterium tuberculosis* infection (letter). The New England Journal of Medicine 322 (23): 1670-1671.

Ferguson, R.G., 1955. Studies in Tuberculosis. Toronto: University of Toronto Press.

Fine, P.E.M., 1981. Immunogenetics of susceptibility to leprosy, tuberculosis, and leishmaniasis. An epidemiological perspective. Editorial. International Journal of Leprosy and Other Mycobacterial Diseases 49 (4): 437-453.

FitzGerald, J.M., Grzybowski, S., and Allen, E.A., 1991. The impact of HIV infection on tuberculosis and its control. Chest 100: 191-197.

FitzGerald, J.M. and Gafni, A., 1990. A cost-effectiveness analysis of the routine use of isoniazid prophylaxis in patients with a positive Mantoux skin test. American Review of Respiratory Disease 142: 848-853.

Freimer, N.B., Sandkuijl, L.A., and Blower, S.M., 1993. Incorrect specification of marker allele frequencies: effects on linkage analysis. American Journal of Human Genetics 52: 1102-1110.

Frieden, T.R., Sterling, T., Pablos-Mendez, A., Kilburn, J.O., Cauthen, G.M., and Dooley, S.W., 1993. The emergence of drug-resistant tuberculosis in New York City. The New England Journal of Medicine 328 (8): 521-526.

Gardner, I.D., 1980. The effect of aging on susceptibility to infection. Reviews of Infectious Diseases 11 (Supplement 2): 801-810.

Gaudette, L.A. and Ellis, E., 1993. Tuberculosis in Canada: a focal disease requiring distinct control strategies for different risk groups. Tubercle and Lung Disease 74: 244-253.

Geiter, L.J., 1993. "Preventive therapy for tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 8. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 241-250.

Gelehrter, T.D. and Collins, F.S., 1990. "Molecular genetics: gene organization, regulation, and manipulation." Principles of Medical Genetics. Chapter 5. Baltimore: Williams and Wilkins. 76.

Glassroth, J., 1993. "Diagnosis of tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 8. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 149-166.

Goldsmith M.F., 1993. New reports make recommendations, ask for resources to stem TB epidemic. Medical News and Perspectives. Jc ırnal of the American Medical Association 269 (2): 187-191.

Goltsov, A.A., Eisensmith, R.C., Naughton, E.R., Jin, L., Chakraborty, R., and Woo, S.L.C., 1993. A single polymorphic STR system in the human phenylalanine hydroxylase gene permits rapid prenatal diagnosis and carrier screening for phenylketonuria. Human Molecular Genetics 2 (5): 577-581.

Goltsov, A.A., Eisensmith, R.C., Konecki, D.S., Lichter-Konecki, U., and Woo, S.L.C., 1992. Associations between mutations and a VNTR in the human phenylalanine hydroxylase gene. American Journal of Human Genetics 51: 627-636.

Greenberg, D.A., 1993. Linkage analysis of "necessary" disease loci versus "susceptibility" loci. American Journal of Human Genetics 52: 135-143.

Greenberg, D.A., 1992. There is more than one way to collect data for linkage analysis. Archives of General Psychiatry 49: 745-750.

Gros, P., Skamene, E., and Forget, A., 1983. Cellular mechanisms of genetically controlled host resistance to *Mycobacterium bovis* (BCG). The Journal of Immunology 131 (4): 1966-1972.

Gros, P., Skamene, E., and Forget, A., 1981. Genetic control of natural resistance to *Mycobacterium bovis* (BCG) in mice. The Journal of Immunology 127 (6): 2417-2421.

Grosset, J.H., 1993. "Bacteriology of tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 3. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 50-74.

Gusella, J.F., Tanzi, R.E., Anderson, M.A., Hobbs, W., Gibbons, K., Raschtchian, R., Gilliam, T.C., Wallace, M.R., Wexler, N.S., and Conneally, P.M., 1984. DNA markers for nervous system diseases. Science 225: 1320-1326.

Haile, R.W., Iselius, L., Fine, P.E., and Monton, N.E., 1985. Segregation and linkage analysis of 72 leprosy pedigrees. Human Heredity 35: 43-52.

Hamburg, M.A., 1992. The challenge of controlling tuberculosis in New York City. New York State Journal of Medicine 92 (7): 291-293.

Hartl, D.L. and Clark, A.G., 1989. Principles of Population Genetics. Sunderland, MA: Sinauer Associates, Inc.

Harvald, B. and Hauge, M., 1956. A catamnestic investigation of Danish twins. Danish Medical Bulletin 3 (5): 150-158.

Hasstedt, S.J., 1989. Pedigree Analysis Package Revision 3.0. Department of Human Genetics, University of Utah Medical Center, Salt Lake City, Utah, U.S.A.

Health and Welfare Canada, 1993. On tuberculosis and aboriginal peoples. Fact sheet distributed at the National Workshop on Tuberculosis, HIV, and Other Emerging Issues, Toronto, Ontario. 3-5 May 1993. Ottawa: Health Protection Branch, Health and Welfare, Canada.

Hearne, C.M., Ghosh, S., and Todd, J.A., 1992. Microsatellites for linkage analysis of genetic traits. Trends in Genetics 8 (8): 288-294.

Hedrick, P.W., 1983. "Selection: an introduction." Genetics of Populations. Chapter 4. Boston: Science Books International Inc. 119-162.

Hill, A.V.S., Allsopp, C.E.M., Kwiatkowski, D., Anstey, N.M., Twumasi, P., Rowe, P.A., Bennett, S., Brewster, D., McMichael, A.J., and Greenwood, B.M., 1991. Common West African HLA antigens are associated with protection from severe malaria. Nature 352: 595-600.

Holden, M., Dubin, M.R., and Diamond, P.H., 1971. Frequency of negative intermediate-strength tuberculin sensitivity in patients with active tuberculosis. The New England Journal of Medicine 285: 1506-1509.

Houston, S., Fanning, A., Soskolne, C.L., and Fraser, N., 1990. The effectiveness of bacillus Calmette-Guerin (BCG) vaccination against tuberculosis. American Journal of Epidemiology 131 (2): 340-348.

Jacobs, R.F. and Starke, J.R., 1993. Tuberculosis in children. Medical Clinics of North America 77 (6): 1335-1351.

Kallmann, F.J. and Reisner, D., 1943. Twin studies on the significance of genetic factors in tuberculosis. American Review of Tuberculosis 47: 549-574.

Kaufmann, S.H.E., 1993. Immunity to intracellular bacteria. Annual Review of Immunology 11: 129-163.

Kelsoe, J.R., Ginns, E.I., Egeland, J.A., Gerhard, D.S., Goldstein, A.M., Bale, S.J., Pauls, D.L., Long, R.T., Kidd, K.K., Conte, G., Housman, D.E., and Paul, S.M., 1989. Re-evaluation of the linkage relationship between chromosome 11p loci and the gene for bipolar affective disorder in the Old Order Amish. Nature 342: 238-243.

Kent, J.H., 1993. The epidemiology of multidrug-resistant tuberculosis in the United States. Medical Clinics of North America 77 (6): 1391-1409.

Khoury, M.J., Beaty, T.H., and Flanders, W.D., 1990. Epidemiologic approaches to the use of DNA markers in the search for disease susceptibility genes. Epidemiologic Reviews 12: 41-55.

Kidd, K.K., 1993. Associations of disease with genetic markers: Déjà vu all over again. American Journal of Medical Genetics (Neuropsychiatric Genetics) 48: 71-73.

King, M.-C., Lee, G.M., Spinner, N.B., Thomsom, G., and Wrensch, M.R., 1984. Genetic epidemiology. Annual Review of Public Health 5: 1-52.

Kringlen, E., 1993. Genes and environment in mental illness. Acta Psychiatrica Scandinavica Supplement 370: 79-84.

Kushigemachi, M., Schneiderman, L.J., and Barrett-Connor, E., 1984. Racial differences in susceptibility to tuberculosis: risk of disease after infection. Journal of Chronic Diseases 37 (11): 853-862.

LaBuda, M.C., Gottesman, I.I., and Pauls, D.L., 1993. Usefulness of twin studies for exploring the etiology of childhood and adolescent psychiatric disorders. American Journal of Medical Genetics (Neuropsychiatric Genetics) 48: 47-59.

Lange, K. and Weeks, D., 1990. "Linkage methods for identifying genetic risk factors." Genetic Variation and Nutrition. Simopoulos, A.P. and Childs, B., eds. Volume 63. World Review of Nutrition and Diet series. Basel: Karger. 236-249.

Lange, K., Boehnke, M., and Weeks, D., 1988. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genetic Epidemiology 5: 471-472.

Lathrop, G.M. and Ott, J., 1990. Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. American Journal of Human Genetics (Supplement 47): A188.

Lathrop, G.M., Lalouel, J.M., Julier, C., and Ott, J., 1984. Strategies for multilocus linkage analysis in humans. Proceedings of the National Academy of Sciences of the United States of America 81: 3443-3446.

Leppert, M.F., 1990. Gene mapping and other tools for discovery. Epilepsia 31 (Supplement 3): S11-S18.

Levinson, D.F. and Mowry, B.J., 1991. Defining the schizophrenia spectrum: issues for genetic linkage studies. Schizophrenia Bulletin 17 (3): 491-514.

Lifton, R.P. and Jeunemaitre, X., 1993. Finding genes that cause human hypertension. Journal of Hypertension 11: 231-236.

Lurie, M.B. and Dannenberg, A.M.Jr., 1965. Macrophage function in infectious disease with inbred rabbits. Bacteriological Reviews 29: 466-476.

Lurie, M.B., Zappasodi, P., Dannenberg, A.M.Jr., and Weiss, G.H., 1952. On the mechanism of genetic resistance to tuberculosis and its mode of inheritance. American Journal of Human Genetics 4: 302-314.

MacCluer, J.W. and Kammerer, C.M., 1991. Invited editorial: dissecting the genetic contribution to coronary heart disease. American Journal of Human Genetics 49: 1139-1144.

Mah, M.W. and Fanning, E.A., 1991. An epidemic of primary tuberculosis in a Canadian aboriginal community. Canadian Journal of Infectious Diseases 2 (4): 133-141.

Malo, D., Vidal, S.M., Hu, J., Skamene, E., and Gros, P., 1993. High-resolution linkage map in the vicinity of the host resistance locus *Bcg*. Genomics 16: 655-663.

Malo, D., Schurr, E., Epstein, D.J., Vekemans, M., Skamene, E., and Gros, P., 1991. The host resistance locus *Bcg* is tightly linked to a group of cytoskeleton-associated protein genes which include villin and desmin. Genomics 10: 356-374.

Marrack, P., and Kappler, J., 1994. Subversion of the immune system by pathogens. Cell 76: 323-332.

Martinez, M., Khlat, M., Leboyer, M., and Clerget-Darpoux, F., 1989. "Performance of linkage analysis under missclassification error when the genetic model is unknown." Genetic Analysis of Complex Traits. Genetic Analysis Workshop 5. Clerget-Darpoux, F., Falk, C.T., and MacCluer, J.W., eds. Genetic Epidemiology 6: 253-258.

Merbs, C.F., 1992. A New World of infectious disease. Yearbook of Physical Anthropology 35: 149-159.

Menzies, R. and Vissandjee, B., 1992. Effect of Bacille Calmette-Guérin vaccination on tuberculin reactivity. American Review of Respiratory Disease 145: 621-625.

Miller, M.A., 1991. A tuberculosis outbreak in a Native community: HLA linkage analysis and evaluation of diagnostic tests. Dissertation. Montreal: Department of Epidemiology and Biostatistics, McGill University.

Morris, B.J., 1993. Identification of essential hypertension genes. Journal of Hypertension 11: 115-120.

Morton, N.E., 1982. Introduction. Outline of Genetic Epidemiology. Basel: S. Karger. 1-5.

Moulding, T., 1988. "Pathogenesis, pathophysiology, and immunology." Tuberculosis. Schlossberg, D., ed. Chapter 2. 2nd ed. New York: Springer-Verlag. 13-22.

Nakahori, Y., Hamano, K., Iwaya, M., and Nakagome, Y., 1991. Sex identification by Polymerase Chain Reaction using X-Y homologous primer. American Journal of Medical Genetics 38: 472-473.

Nardell, E.A., 1993. "Pathogenesis of tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichman, L.B. and Hershfield, E.S., eds. Chapter 5. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 103-122.

Neel, J.V. and Schull, W.J., 1954. "Genetics and epidemiology." Human Heredity. Chapter 17. Chicago: The University of Chicago Press. 283-306.

O'Brien, R.J., 1993. "The treatment of tuberculosis." Tuberculosis: A Comprehensive International Approach. Reichmen, L.B. and Hershfield, E.S., eds. Chapter 11. Volume 66. Lung Biology in Health and Disease series. New York: Marcel Dekker, Inc. 207-240.

O'Brien, S.J., 1991. Ghetto legacy: can the high incidence of Tay-Sachs disease in Ashkenazi Jews be linked to historic epidemics of tuberculosis in industrial European cities? Current Biology 4: 209-211.

O'Connell, P., Lathrop, G.M., Nakamura, Y., Leppert, M.L., Lalouel, J.-M., and White, R., 1989. Twenty loci form a continuous linkage map of markers for human chromosome 2. Genomics 5: 738-745.

Ott, J., 1992. Strategies for characterizing highly polymorphic markers in human gene mapping. American Journal of Human Genetics 51: 283-290.

Ott, J., 1991. Analysis of Human Genetic Linkage. Baltimore: The Johns Hopkins University Press.

Ott, J., 1990a. "Documentation to homogeneity programs." Notes from Advanced Linkage Course, 11-15 Jan. 1993. New York: Columbia University.

Ott, J., 1990b. Invited editorial: cutting a Gordian knot in the linkage analysis of complex human traits. American Journal of Human Genetics 46: 219-221.

Ott, J., 1989. Computer-simulation methods in human linkage analysis. Proceedings of the National Academy of Sciences of the United States of America 86: 4175-4178.

Pauls, D.L., 1993. Behavioral disorders: lessons in linkage. Nature Genetics 3: 4-5.

Pausova, Z., Morgan, K., Fujiwara, M., Bourdon, J., Goltzman, D., and Hendy, G.N., 1993. Molecular characterization of an intragenic minisatellite (VNTR) polymorphism in the human parathyroid hormone-related peptide gene in chromosome region 12p12.1-p11.2. Genomics 17: 243-244.

Peto, J., 1980. "Genetic predisposition to cancer." Cancer Incidence in Defined Populations. Cairns, J., Lyon, J.L., and Skolnick, M., eds. Banbury Report 4. U.S.A.: Cold Spring Harbor Laboratory. 203-213.

Ravikrishnan, K.P., 1992. Tuberculosis: how can we halt its resurgence? Postgraduate Medicine 91 (4): 333-338.

Rieder, H.L., Cauthen, G.M., Comstock, G.W., and Snider, D.E.Jr., 1989a. Epidemiology of tuberculosis in the United States. Epidemiologic Reviews 11: 79-98.

Rieder, H.L., Cauthen, G.M., Kelly, G.D., Bloch, A.B., and Snider, D.E.Jr., 1989b. Tuberculosis in the United States. Journal of the American Medical Association 262 (3): 385-389.

Risch, N., 1992. Genetic linkage: interpreting lod scores. Science 255: 803-804.

Risch, N., 1991. A note on multiple testing procedures in linkage analysis. American Journal of Human Genetics 48: 1058-1064.

Risch, N., 1990. Genetic linkage and complex diseases, with special reference to psychiatric disorders. Genetic Epidemiology 7: 3-16.

Risch, N., 1987. Assessing the role of HLA-linked and unlinked determinants of disease. American Journal of Human Genetics 40: 15-31.

Risch, N., Claus, E., and Giuffra, L., 1989. "Linkage and mode of inheritance in complex traits." Multipoint Mapping and Linkage Based Upon Affected Pedigree Members. Genetic Analysis Workshop 6. Elston, R.C., Spence, M.A., Hodge, S.E., and MacCluer, J.W., eds. Progress in Clinical and Biological Research 329: 183-188.

Rodrigues, L.C. and Smith, P.G., 1990. Tuberculosis in developing countries and methods for its control. Transactions of the Royal Society of Tropical Medicine and Hygiene 84: 739-744.

Rosenman, K.D., 1990. Racial differences and *Mycobacterium tuberculosis* infection (letter). The New England Journal of Medicine 322 (23): 1670.

Rubin, A.L., 1993. Tuberculosis mortality decline. Letter. Science 261: 277.

Sackett, D.L., 1979. Bias in analytic research. Journal of Chronic Diseases 32: 51-63.

Sagan, L.A., 1987. The Health of Nations: True Causes of Sickness and Well-Being. New York: Basic Books.

Sandkuijl, L.A., 1993. "'Extended' sib-pair analysis." Notes from Advanced Linkage Course, 11-15 Jan. 1993. New York: Columbia University.

Sanjeevi, C.B., Narayanan, P.R., Prabakar, R., Charles, N., Thomas, B.E., Balasubramaniam, R., and Olerup, O., 1992. No association or linkage with HLA-DR or -DQ genes in South Indians with pulmonary tuberculosis. Tubercle and Lung Disease 73: 280-284.

Schull, W.J., 1993. The Raymond Pearl memorial lecture, 1992: ethnicity and disease--more than familiality. American Journal of Human Biology 5: 373-385.

Schurr, E., Malo, D., Radzioch, D., Buschman, E., Morgan, K., Gros, P., and Skamene, E., 1991a. Genetic control of innate resistance to mycobacterial infections. Immunoparasitology Today. Ash, C. and Gallagher, R.B., eds. Cambridge: Elsevier Trends Journals. A42-A45.

Schurr, E., Morgan, K., Gros, P., and Skamene, E., 1991b. Genetics of leprosy. American Journal of Tropical Medicine and Hygiene. 44(3): 4-11.

Schurr, E., Buschman, E., Malo, D., Gros, P., and Skamene, E., 1990a. Immunogenetics of mycobacterial infections: mouse-human homologies. The Journal of Infectious Diseases 161: 634-639.

Schurr, E., Skamene, E., Morgan, K., Chu, M-L., and Gros, P., 1990b. Mapping of *Col3a* and *Col6a3* to proximal murine chromosome 1 identifies conserved linkage of structural protein genes between murine chromosome 1 and human chromosome 2q. Genomics 8: 477-486.

Schurr, E., Buschman, E., Gros, P., and Skamene, E., 1989a. Genetic aspects of mycobacterial infections in mouse and man. Progess in Immunology 7: 994-1001.

Schurr, E., Henthorn, P.S., Harris, H., Skamene, E., and Gros, P., 1989b. Localization of two alkaline phosphatase genes to the proximal region of mouse chromosome 1. Cytogenetics and cell genetics 52: 65-67.

Schurr, E., Skamene, E., Forget, A., and Gros, P., 1989c. Linkage analysis of the *Bcg* gene on mouse chromosome 1: identification of a tightly linked marker. The Journal of Immunology 142: 4507-4513.

Schweinle, J.O., 1990. Evolving concepts of the epidemiology, diagnosis, and therapy of *M. tuberculosis* infection. Yale Journal of Biology and Medicine 63: 565-579.

Serjeantson, S., Wilson, S.R., and Keats, B.J., 1979. The genetics of leprosy. Annals of Human Biology 6: 375-393.

Shaw, M.-A., Atkinson, S., Dockrell, H., Hussain, R., Lins-Lainson, Z., Shaw, J., Ramos F., Silveira, F., Mehdi, S.Q., Kaukab, F., Khaliq, S., Chiang, T., and Blackwell, J., 1993. An RFLP map for 2q33-q37 from multicase mycobacterial and leishmanial disease families: no evidence for an *Lsh/Ity/Bcg* gene homologue influencing susceptibility to leprosy. Annals of Human Genetics 57: 251-271.

Singh, S.P.N., Mehra, N.K., Dingley. H.B , Pande, J.N., and Vaidya, M.C., 1983. Human Leukocyte Antigen (HLA)-linked control of susceptibility to pulmonary tuberculosis and association with HLA-DR types. The Journal of Infectious Diseases 148 (4): 676-681.

Skamene, E., 1991. Population and molecular genetics of susceptibility to tuberculosis. Clinical and Investigative Medicine 14 (2): 160-166.

Skamene, E., 1989. Genetic control of susceptibility to mycobacterial infections. Reviews of Infectious Diseases 11 (Supplement 2): S394-399.

Skamene, E., 1986. Genetic control of resistance to mycobacterial infection. Current Topics in Microbiology and Immunology 124: 49-66.

Smith, D.G., 1979. The genetic hypothesis for susceptibility to lepromatous leprosy. Human Genetics 50: 163-177.

Snider, D.E.Jr., 1993. The impact of tuberculosis on women, children, and minorities in the United States. Abstract presented at the World Congress on Tuberculosis, Bethesda, Maryland. 16-19 Nov. 1992.

Snider, D.E.Jr., 1982. The tuberculin skin test. American Review of Respiratory Disease 125 (Supplement 3): S108-S118.

Statistics Canada, 1992. Tuberculosis Statistics, 1990. Statistics Canada Catalogue 82-003S10. Health Reports 4 (2; Supplement 10): 1-50.

Statistics Canada, 1989. Tuberculosis Statistics, 1987. Statistics Canada Catalogue 82-003S. Health Reports 1: 69-79.

Statistics Canada, 1988. Tuberculosis Statistics, Morbidity and Mortality 1986. Statistics Canada Catalogue 8?-212. Ottawa: Vital Statistics and Health Status section, Health Division, Statistics Canada.

Stead, W.W., 1992. Genetics and resistance to tuberculosis. Could resistance be enhanced by genetic engineering? Annals of Internal Medicine 116 (11: 937-941.

Stead, W.W, 1989. Pathogenesis of tuberculosis: clinical and epidemiologic perspective. Reviews of Infectious Diseases 11 (Supplement 2): S366-S368.

Stead, W.W. and Dutt, A.K., 1988. "Epidemiology and host factors." Tuberculosis. Schlossberg, D., ed. Chapter 1. 2nd ed. New York: Springer-Verlag. 1-11.

Stead, W.W., Senner, J.W., Reddick, W.T., and Lofgren, J.P., 1990. Racial differences in susceptibility to infection by *Mycobacterium tuberculosis*. The New England Journal of Medicine 332 (7): 422-427.

Styblo, K., 1991. Epidemiology of tuberculosis. Vol. 24. The Hague: Royal Netherlands Tuberculosis Association Selected Papers.

Styblo, K., 1989. Overview and epidemiologic assessment of the current global tuberculosis situation with an emphasis on control in developing countries. Reviews of Infectious Diseases 11 (Supplement 2): S339-S346.

Styblo, K., 1980. Recent advances in epidemiological research in tuberculosis. Advances in Tubercle Research 20: 1-63.

Sudre, P., ten Dam, G., and Kochi, A., 1992. Tuberculosis: a global overview of the situation today. Bulletin of the World Health Organization 70 (2): 149-159.

Sutherland, I., 1976. Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli. Advances in Tubercle Research 19: 1-63.

Terasaki, P., McClelland, J.D., Parks, M.S. and McCurdy, B., 1973. "Microdroplet lymphocyte cytotoxicity test." Manual of tissue typing techniques. Publication No. 74. Washington: United States Department of Health, Education, and Welfare. 545.

Terwilliger, J. and Ott, J., 1993a. "Maximizing the lod score over models." Notes from Advanced Linkage Course, 11-15 Jan. 1993. New York: Columbia University. 1-4.

Terwilliger, J. and Ott, J., 1993b. "Nonparametric approaches." Notes from Advanced Linkage Course, 11-15 Jan. 1993. New York: Columbia University. 1-7.

Thomson, G., 1991. "Theoretical modelling and population epidemiology of complex genetic diseases." The Immunogenetics of Autoimmune Diseases. N. Farid, ed. Chapter 13. Vol. 1. U.S.A.: CRC Press. 188-203.

Thorpe, E.L.M., 1989. The social histories of smallpox and tuberculosis in Canada (culture, evolution and disease). Revised dissertation. University of Manitoba Anthropology Papers No. 30. Winnipeg: Department of Anthropology, University of Manitoba.

Trowsdale, J., 1993. Genomic structure and function in the MHC. Trends in Genetics 9 (4): 117-122.

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

Vidal, S.M., Epstein, D.J., Malo, D., Weith, A., Vekemans, M., and Gros, P., 1992. Identification and mapping of six microdissected genomic DNA probes to the proximal region of mouse chromosome 1. Genomics 14: 32-37.

Wagener, D.K., Schauf, V., Nelson, K.E., Scollard, D., Brown, A., and Smith, T., 1988. Segregation analysis of leprosy in families of northern Thailand. Genetic Epidemiology 5: 95-105.

Weber, J.L. and Wong, C., 1993. Mutation of human short tandem repeats. Human Molecular Genetics 2 (8): 1123-1128.

Weeks, D.E. and Ott, J., 1993. "SLINK: a general simulation program for linkage analysis." Notes from Advanced Linkage Course, 11-15 Jan. 1993. New York: Columbia University. 1-21.

Weeks, D.E. and Lange, K., 1988. The affected-pedigree-member method of linkage analysis. American Journal of Human Genetics 42: 315-326.

Weeks, D.E, Harby, L.D., Sarneso, C.A., and Gorin, M.B., 1993. "The affected pedigree member method of linkage analysis." Notes from Advanced Linkage Course, 11-15 Jan. 1993. New York: Columbia University.

Wijsman, E.M., 1993. "Genetic analysis of Alzheimer's disease: a summary of contributions to GAW8." Genetic Analysis Workshop 8. Issues in the Analysis of Complex Diseases and Their Risk Factors. Elston, R.C., Spence, M.A., Haines, J.L., Marazita, M.L., Pericak-Vance, M.A., Siervogel, R.M., and MacCluer, J.W., eds. Genetic Epidemiology 10 (6): 349-360.

Wijsman, E.M., 1990. "Linkage analysis of alcoholism: problems and solutions." Genetics and Biology of Alcoholism. Cloninger, C.R. and Begleiter, H., eds. Banbury Report 33. U.S.A.: Cold Spring Harbor Laboratory Press. 317-326.

Young, T. and Hershfield, E., 1986. A case-control study to evaluate the effectiveness of mass neonatal BCG vaccination among Canadian Indians. American Journal of Public Health 76: 783-786.

APPENDIX A

Questionnaire sent to the Colombian collaborators

Note.
For reasons of confidentiality, gender information and pedigree diagrams have been removed from the questionnaire

McGill

TUBERCULOSIS LINKAGE ANALYSIS QUESTIONNAIRE

TO BE COMPLETED BY COLLABORATORS

IN COLOMBIA

PREPARED BY: LUCY BOOTHROYD, B.Sc.
MASTER'S THESIS STUDENT
DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS
MCGILL UNIVERSITY

OCTOBER 31, 1992

REVISED
OCTOBER 31, 1992

TUBERCULOSIS LINKAGE ANALYSIS QUESTIONNAIRE

Total number of pages: 23, including pedigrees and maps

Questionnaire to be filled out by investigators involved in the collection and diagnoses of: **Families GV, MAM, RAV, RES, and RJA** (pedigree structures provided on pages 17 and 18)

Please read over the questionnaire and contact us if you have any questions or comments regarding this form. Our address and fax number appear on page 16. Instructions and information about the questions appear in **bold** print.

Section A. Collection Methods

This information will be used to evaluate how well the five families in this study represent all families affected with tuberculosis in Colombia. We would like to be able to describe the location of the study families in Colombia. We also want to describe in general the procedure used to collect the blood samples.

1. Please confirm whether the minimum criteria used to select the families was as follows: two or more offspring in the family were affected with tuberculosis, one parent was affected and the other parent was unaffected, and there may have been additional affected relatives. **Circle the family name beside the responses that apply.**

Criteria as above for families:	GV	MAM	RAV	RES	RJA
Different set of criteria for families:	GV	MAM	RAV	RES	RJA

Please specify the different set of criteria: _____

_____

_____

2. It is our impression that families meeting the criteria for linkage analysis are relatively infrequent (that is, many Colombian families affected with tuberculosis do not fit the selection criteria specified in question 1). Please indicate with what approximate frequency a family meeting the criteria you used for selection occurs among all affected families in the population which you serve. **Check which response applies for the selection criteria used. If only one set of criteria was used for all the families answer for the one criteria set only.**

For families in the study meeting the selection criteria suggested by us in question 1, such families occur with the frequency. **(check which response applies)**

_____ 1 in 10 **(...continued next page)**

REVISED
OCTOBER 31, 1992

_____ 1 in 100

_____ 1 in 1000

_____ other **(please specify frequency):** _____

_____ don't know

For families in the study meeting a different set of selection criteria used by you and specified in your response to question 1, such families occur with the frequency: **(check which response applies)**

_____ 1 in 10

_____ 1 in 100

_____ 1 in 1000

_____ other **(please specify frequency):** _____

_____ don't know

3. What strategy was used to find the families for this study?
Circle the family name beside the responses that apply.

Field workers knowledge for families: GV MAM RAV RES RJA

Search of a registry (for example, a hospital registry of tuberculosis patients) for families: GV MAM RAV RES RJA

Other **(please specify method and indicate families involved):** _____

_____

_____

4. Through which members were the families made known to you?
Circle the family name beside the responses that apply.

Affected children for families: GV MAM RAV RES RJA

Affected parents for families: GV MAM RAV RES RJA

Affected parents and children for families: GV MAM RAV RES RJA

REVISED
OCTOBER 31, 1992

5 Please confirm that the study families were selected in an opportunistic manner, that is, appropriate families were chosen for the study as they came to your attention. **Circle the family name beside the responses that apply.**

Selected as above for families: GV MAM RAV RES RJA

Other manner of selection **(please specify and indicate families involved):** _____

_____

_____

6. Where was each family living at the time of collection of the blood samples? **Please give the name of the village/town/city <u>and</u> province for each family if known. Please mark (with an "X" or a circle in colour) the nearest location for each family on the maps provided (pages 19 to 23) where possible (please specify the family name(s) for each marking).** Pages 19 and 20 present a map of Colombia while Page 21 to 23 feature Antioquia and Cordoba provinces for more specific indication of locations.

GV: _____

MAM: _____

RAV: _____

RES: _____

RJA: _____

7. Please describe the process of collecting the blood samples sent to us in Montréal by responding to the following four questions.

(a) **Please specify** the number and type of workers involved in collecting blood samples (for example: numbers of nurses, doctors, technicians):

_____

_____

REVISED
OCTOBER 31, 1992

(b) Please confirm the locations at which the bloods were drawn. **Check which response applies.**

_____ at hospitals and/or local health clinics only

_____ at the families' homes only

_____ both at hospitals and/or local clinics and at home

(c) Please specify the general range of time between blood collection and sample processing (that is, the time to either freezing of the samples or separation of mononuclear cells). **Fill in the blank.**

Range of hours from sample collection to processing: _____ hours

(d) Please specify when the blood samples were collected. **Fill in the blanks as indicated.**

Between the date _____ and the date _____.
(month/year) (month/year)

REVISED
OCTOBER 31, 1992

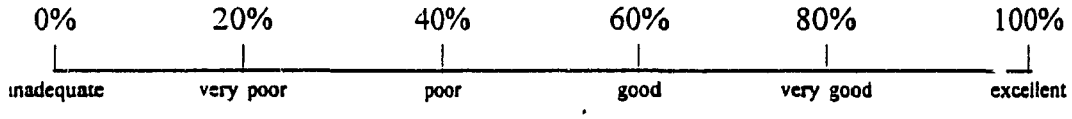
Section B: Diagnostic Methods

Here, we want information concerning the diagnosis of tuberculosis disease. We want to know which diagnostic tests were used and the confidence you have in the tests' ability to correctly indicate absence or presence of tuberculosis disease (i.e. the sensitivity and specificity of the tests). We also want to know if any family members have ever received BCG vaccination or if any individuals were given anti-tuberculosis drug treatment following the first diagnosis of tuberculosis in the family.

Another aim of this section is to assess whether any members of the families may not have developed tuberculosis because they were not exposed to the mycobacteria (that is, they were not living with the family members who were infectious). We ask you to indicate whether you know if the unaffected individuals were living with an affected family member at the time the relative was diagnosed with tuberculosis. We plan to use the skin test results as an indication of exposure to the tuberculosis infection.

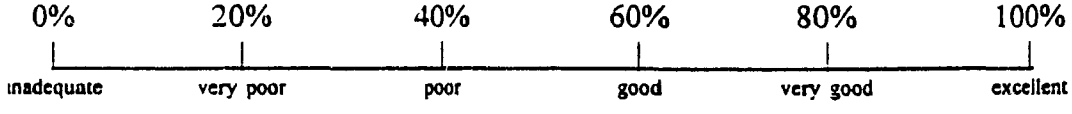
1. SENSITIVITY OF DIAGNOSTIC TESTS:
   Evaluate each available diagnostic test by **marking on the line** (with an X) where you would place your impression of the sensitivity of the diagnostic test under the conditions for the diagnoses in the five families. **The numbers over the line are the percentage of times a test will indicate disease in someone affected with tuberculosis and the words beneath describe the sensitivity level of the test. If a test is unavailable, leave the line blank.**

Sputum culture



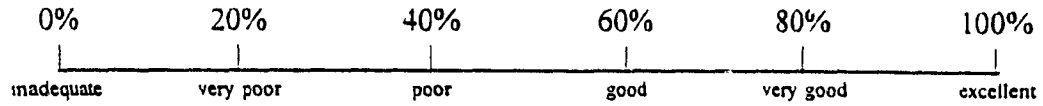
Other fluid culture (e.g. gastric fluid)

Microscope smear

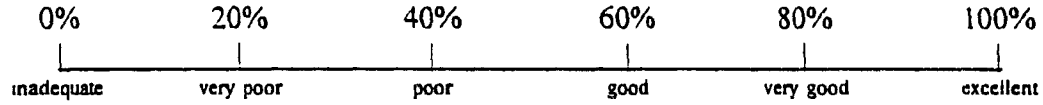


REVISED
OCTOBER 31, 1992

Chest X-ray

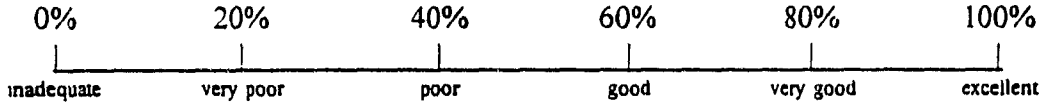


Evidence of typical clinical symptoms of tuberculosis (e.g. fever, fatigue, weight loss, cough)



Skin test

Response of symptoms to tuberculosis treatment (an individual may be considered affected with tuberculosis if his/her clinical symptoms were relieved by treatment given because he/she had contact with an infectious case)



2. SPECIFICITY OF DIAGNOSTIC TESTS:
Evaluate each available diagnostic test by **marking on the line** (with an X) where you would place your impression of the specificity of the diagnostic test under the conditions for the diagnoses in the five families. **The numbers over the line are the percentage of times a test will indicate <u>absence</u> of disease in someone <u>not</u> affected with tuberculosis and the words beneath describe the specificity level of the test. If a test is unavailable, leave the line blank.**

Sputum culture

REVISED
OCTOBER 31, 1992

Other fluid culture (e.g. gastric fluid)

Microscope smear

No evidence of typical clinical symptoms of tuberculosis (e.g. fever, fatigue, weight loss, cough)

Skin test

3. **Rank** the methods in the order of their usefulness in diagnosing tuberculosis under the conditions for the diagnoses in the five families where 1 = most useful, 2 = next most useful, and so on. **If several methods are equally useful or used together in diagnosis, rank them with the same number. If a method is unavailable, use the rank "0".**

_____ sputum culture

_____ other fluid culture

_____ microscope smear

_____ chest X-ray

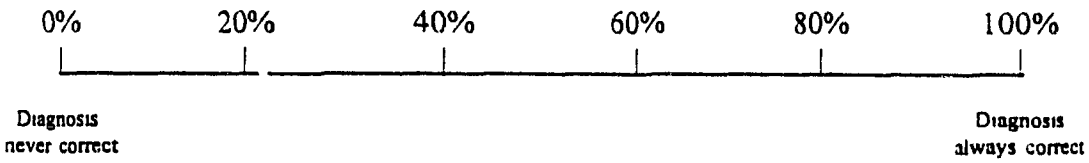
_____ evidence of typical clinical symptoms

_____ skin test

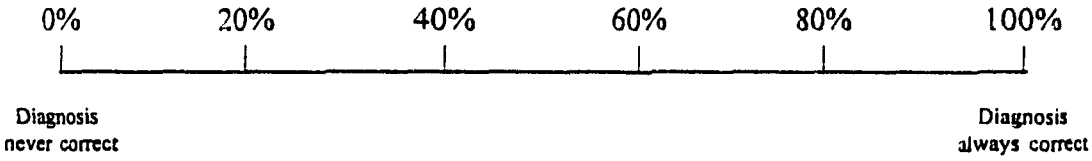
_____ response of symptoms to tuberculosis treatment

REVISED
OCTOBER 31, 1992

4. REGARDING THE DIAGNOSIS OF "AFFECTED":
Given the limitations of the available diagnostic tests, how confident are you that a person diagnosed as "affected with tuberculosis disease" under the conditions for these five families actually has tuberculosis disease? **Please indicate your level of confidence by marking the line below, where the numbers represent percentage of times the affected diagnosis is correct.**



5. REGARDING THE DIAGNOSIS OF "NOT AFFECTED":
Given the diagnostic conditions available for these study families, how confident are you that an individual diagnosed as "not affected with tuberculosis" is actually not diseased? **Please indicate your level of confidence by marking the line below, where the numbers represent percentage of times the unaffected diagnosis is correct.**



6. Please indicate the source(s) of the skin test antigen(s) used for tuberculin skin tests of the study family members if known:

_____

_____

_____

7. DIAGNOSIS, SKIN TEST, AND VACCINATION INFORMATION FOR THE INDIVIDUALS DIAGNOSED AS AFFECTED WITH TUBERCULOSIS

OCT 31/92

PLEASE COMPLETE INFORMATION FOR EACH AFFECTED PERSON WITH THE APPROPRIATE SYMBOLS AS INDICATED IN THE COLUMNS AND BELOW THE TABLE

						DURING INVESTIGATION FOR TUBERCULOSIS					PRIOR TO TUBERCULOSIS DIAGNOSIS DATE		
Individual	Sex	Date of birth (d/mo/yr)	Age at diagnosis (months or years)	Date of diagnosis (d/mo/yr)	Ever BCG vaccinated? (yes, no, or d k)	Culture (+, -, or n d) and date (d/mo/yr)	Chest X-ray (+, -, or n d) and date (d/mo/yr)	Microscopy (+, -, or n d) and date (d/mo/yr)	Tuberculosis symptoms shown? (yes, no, or d k)	Skin test result (n d or mm induration) and date (d/mo/yr)	Skin test at an earlier time (n d or mm induration) and date (d/mo/yr)	Had tuberculosis before? (yes, no, or d k)	Had drug treatment before diagnosis? (1, 2, 3, or 4)* SEE BELOW
GV-01													
GV-03													
GV-05													
GV-07													
GV-09													
GV-10													
MAM-02													
MAM-03													
MAM-05													
MAM-07													

d = day mo = month yr = year

d k = don't know

+ = positive result - = negative result

n d = not determinable, ambiguous, or not done

*use the following codes: 1 = no treatment given

2 = treatment given; individual was a suspected case

3 = treatment given, individual was asymptomatic but had contact with an infectious person

4 = treatment given, not clear whether case 2 or 3 applies

QUESTION 7 CONTINUED DIAGNOSIS, SKIN TEST, AND VACCINATION INFORMATION FOR INDIVIDUALS DIAGNOSED AS AFFECTED WITH TUBERCULOSIS

OCT 31/92

PLEASE COMPLETE INFORMATION FOR EACH AFFECTED PERSON WITH THE APPROPRIATE SYMBOLS AS INDICATED IN THE COLUMNS AND BELOW THE TABLE

						DURING INVESTIGATION FOR TUBERCULOSIS					PRIOR TO TUBERCULOSIS DIAGNOSIS DATE		
Individual	Sex	Date of birth (d/mo/yr)	Age at diagnosis (months or years)	Date of diagnosis (d/mo/yr)	Ever BCG vaccinated? (yes, no, or d k)	Culture (+, -, or n d) and date (d/mo/yr)	Chest X-ray (+, -, or n d) and date (d/mo/yr)	Microscopy (+, - or n d) and date (d/mo/yr)	Tuberculosis symptoms? shown? (yes, no, or d k)	Skin test result (n d or mm induration) and date (d/mo/yr)	Skin test at an earlier time (n d or mm induration) and date (d/mo/yr)	Had tuberculosis before? (yes, no, or d k)	Had drug treatment before diagnosis? (1, 2, 3, or 4)* SEE BELOW
RAV-01													
RAV-05													
RAV-07													
RES 01													
RES-05													
RES-16													
RES-18													
RJA-04													
RJA-05													
RJA 97													

d = day mo = month yr = year

d k. = don't know

+ = positive result - = negative result

n d. = not determinable, ambiguous, or not done

*use the following codes

1 = no treatment given

2 = treatment given, individual was a suspected case

3 = treatment given; individual was asymptomatic but had contact with an infectious person

4 = treatment given, not clear whether case 2 or 3 applies

8 DIAGNOSIS, SKIN TEST, AND VACCINATION INFORMATION FOR THE INDIVIDUALS DIAGNOSED AS NOT AFFECTED WITH TUBERCULOSIS

OCT 31/92

PLEASE COMPLETE INFORMATION FOR EACH NON-AFFECTED PERSON WITH THE APPROPRIATE SYMBOLS AS INDICATED IN THE COLUMNS AND BELOW THE TABLE

								DURING INVESTIGATION FOR TUBERCULOSIS					Skin test result prior
Individual	Sex	Date of birth* (d/mo/yr)	Date of diagnosis (d/mo/yr)	Ever BCG vaccinated? (yes,no, or d k)	Living with** an infectious relative? (yes, no or d k)	Given drug treatment*** after first diagnosis in family? (yes, no, or d k)	Had tuberculosis before? (yes, no, or d k)	Culture (+, -, or n d) and date (d/mo/yr)	Chest X ray (+, -, or n d) and date (d/mo/yr)	Microscopy (+, -, or n d) and date (d/mo/yr)	Tuberculosis symptoms shown? (yes, no, or d k)	Skin test result (n d or mm induration) and date (d/mo/yr)	to first diagnosis in family ~ (n d or mm induration) and date (d/mo/yr)
GV-02													
GV-04													
GV-06													
GV-08													
GV-11													
GV-13													
GV 14													
GV-15													
MAM 01													
MAM-04													
MAM 06													
MAM 08													

d = day mo = month yr = year

d k = don't know

+ = positive result - = negative result

n d = not determinable, ambiguous, or not done

*If unavailable, give age at diagnosis (mo or yrs) **Living with a study family member at the time the relative was diagnosed as affected

***That is, anti tuberculosis treatment given to the asymptomatic individual because contact with an infectious person was suspected

~ Most recent skin test before first tuberculosis diagnosis in the family

QUESTION 8 CONTINUED DIAGNOSIS, SKIN TEST, AND VACCINATION INFORMATION FOR THE INDIVIDUALS DIAGNOSED AS NOT AFFECTED

OCT 31/92

PLEASE COMPLETE INFORMATION FOR EACH NON AFFECTED PERSON WITH THE APPROPRIATE SYMBOLS AS INDICATED IN THE COLUMNS AND BELOW THE TABLE

								DURING INVESTIGATION FOR TUBERCULOSIS					Skin test result prior
Individual	Sex	Date of birth* (d/mo/yr)	Date of diagnosis (d/mo/yr)	Ever BCG vaccinated? (yes, no, or d k)	Living with** an infectious relative? (yes, no, or d k)	Given drug treatment*** after first diagnosis in family? (yes, no, or d k)	Had tuberculosis before? (yes, no, or d k)	Culture (+, -, or n d) and date (d/mo/yr)	Chest X-ray (+, -, or n d) and date (d/mo/yr)	Microscopy (+, -, or n d) and date (d/mo/yr)	Tuberculosis symptoms shown? (yes, no, or d k)	Skin test result (n d or mm induration) and date (d/mo/yr)	to first diagnosis in family ~ (n d or mm induration) and date (d/mo/yr)
RAV 02													
RAV-03													
RAV-04													
RAV-06													
RAV 08													
RJA-01													
RJA 02													
RJA-03													
RJA 09													
RJA 10													
RJA 96													
RJA 98													
RJA 99													

d = day mo = month yr = year

d k = don't know

+ = positive result - = negative result

n d = not determinable, ambiguous, or not done

*If unavailable, give age at diagnosis (mo or yrs)

**Living with a study family member at the time the relative was diagnosed as affected

***That is, anti tuberculosis treatment given to the asymptomatic individual because contact with an infectious case was suspected

~ Most recent skin test before first tuberculosis diagnosis in the family

QUESTION 8 CONTINUED. DIAGNOSIS, SKIN TEST, AND VACCINATION INFORMATION FOR THE INDIVIDUALS DIAGNOSED AS NOT AFFECTED

OCT 31/92

PLEASE COMPLETE INFORMATION FOR EACH NON AFFECTED PERSON WITH THE APPROPRIATE SYMBOLS AS INDICATED IN THE COLUMNS AND BELOW THE TABLE

								DURING INVESTIGATION FOR TUBERCULOSIS					
Individual	Sex	Date of birth* (d/mo/yr)	Date of diagnosis (d/mo/yr)	Ever BCG vaccinated? (yes, no, or d k)	Living with** an infectious relative? (yes, no, or d k)	Given drug treatment*** after first diagnosis in family? (yes, no, or d k)	Had tuberculosis before? (yes, no, or d k)	Culture (+, -, or n d) and date (d/mo/yr)	Chest X-ray (+, -, or n d) and date (d/mo/yr)	Microscopy (+, -, or n d) and date (d/mo/yr)	Tuberculosis symptoms shown? (yes, no, or d k)	Skin test result (n d or mm induration) and date (d/mo/yr)	Skin test result prior to first diagnosis in family ~ (n d or mm induration) and date (d/mo/yr)
RES-02													
RES-03													
RES-04													
RES 06													
RES-07													
RES-08													
RES-09													
RES-10													
RES-11													
RES-12													
RES-13													
RES-14													
RES-15													
RES-17													

d = day mo = month yr = year

d k = don't know

+ = positive result - = negative result

n.d. = not determinable, ambiguous, or not done

*If unavailable, give age at diagnosis (mo or yrs)

**Living with a study family member at the time the relative was diagnosed as affected

***That is, anti-tuberculosis treatment given to the asymptomatic individual because contact with an infectious case was suspected

~Most recent skin test before first tuberculosis diagnosis in the family

Section C: Living Conditions

In this section, we would like to evaluate the potential for transmission of the tuberculosis mycobacterium in the air between the family members in their home environment. If transmission potential is high it is more reasonable to assume that everyone in the home has been exposed.

PLEASE COMPLETE AS MUCH INFORMATION AS POSSIBLE ABOUT THE HOME OF EACH FAMILY ABOUT THE TIME OF THE FIRST TUBERCULOSIS DIAGNOSIS IN THE FAMILY (WITHIN TWO MONTHS) IF KNOWN

USE THE APPROPRIATE SYMBOLS AS INDICATED

Family	Number of persons* living in same household	Number of sleeping rooms in dwelling	Ventilation of dwelling (good or poor)**	High potential for transmission of disease in air?*** (yes or no)	Date home seen (d/mo/yr)
GV					
MAM					
RAV					
RES					
RJA					

d = day mo = month yr = year

* Include non-family members

** A well-ventilated dwelling would have many windows and perhaps a fan

*** In your opinion, based on your impression of the home

REVISED
OCTOBER 31, 1992

Section D: Tuberculosis Situation

We want to use the information provided here to estimate the frequency of a susceptibility allele and assess whether the level of exposure to infection experienced by the five families in the study is typical for families living in their area.

1. Please give a general estimate of tuberculosis incidence (number of new cases in the population per year) and tuberculosis prevalence (number of active cases in the population at one time) for the areas in which the families live. Any estimate, however "rough", will be useful. **Fill in the blanks. If the families come from several areas with varying tuberculosis situations, please provide ranges of incidence and prevalence.**

Incidence:

_____ new cases per population of 100 000 individuals during the year

_____
year

Prevalence:

_____ active cases per population of 100 000 individuals at the time

_____
month/year

2. For the time for which you have personal knowledge with the regions where the study families live, has tuberculosis incidence generally been at a stable level, steadily increasing or decreasing, or has there been periodic epidemics (large increases in incidence followed by decreases)? **Circle the family name under the appropriate responses and indicate the relevant time period in years.**

   Stable incidence for:

   GV for the time period _____

   MAM for the time period _____

   RAV for the time period _____

   RES for the time period _____

   RJA for the time period _____

(...continued next page)

REVISED
OCTOBER 31, 1992

Increasing or decreasing incidence for:

GV for the time period _____

MAM for the time period _____

RAV for the time period _____

RES for the time period _____

RJA for the time period _____

Periodic epidemics for:

GV for the time period _____

MAM for the time period _____

RAV for the time period _____

RES for the time period _____

RJA for the time period _____

3. At the time of the first diagnosis of tuberculosis in each family, indicate whether tuberculosis incidence was stable, on the rise, or on the decrease. **Circle the family name beside the responses that apply.**

Stable incidence for families:	GV	MAM	RAV	RES	RJA
Increasing incidence for families:	GV	MAM	RAV	RES	RJA
Decreasing incidence for families:	GV	MAM	RAV	RES	RJA

****

We thank you very much for your kind cooperation. We appreciate your time and effort needed to complete this questionnaire.

Lucy Boothroyd B.Sc.
Master's Thesis Program, Department of Epidemiology and Biostatistics
McGill University 1020 Pine Avenue West
Montréal, Québec CANADA H3A 1A2
(514) 937-6011 extension 4627 (telephone) 514-934-8273 (fax)

Supervisor: Dr. K. Morgan (Depts. of Epidemiology & Biostatistics and Medicine)

APPENDIX B

Pedigree diagrams and case distribution by diagnostic methods

B1 Colombian families

B2 Hong Kong families

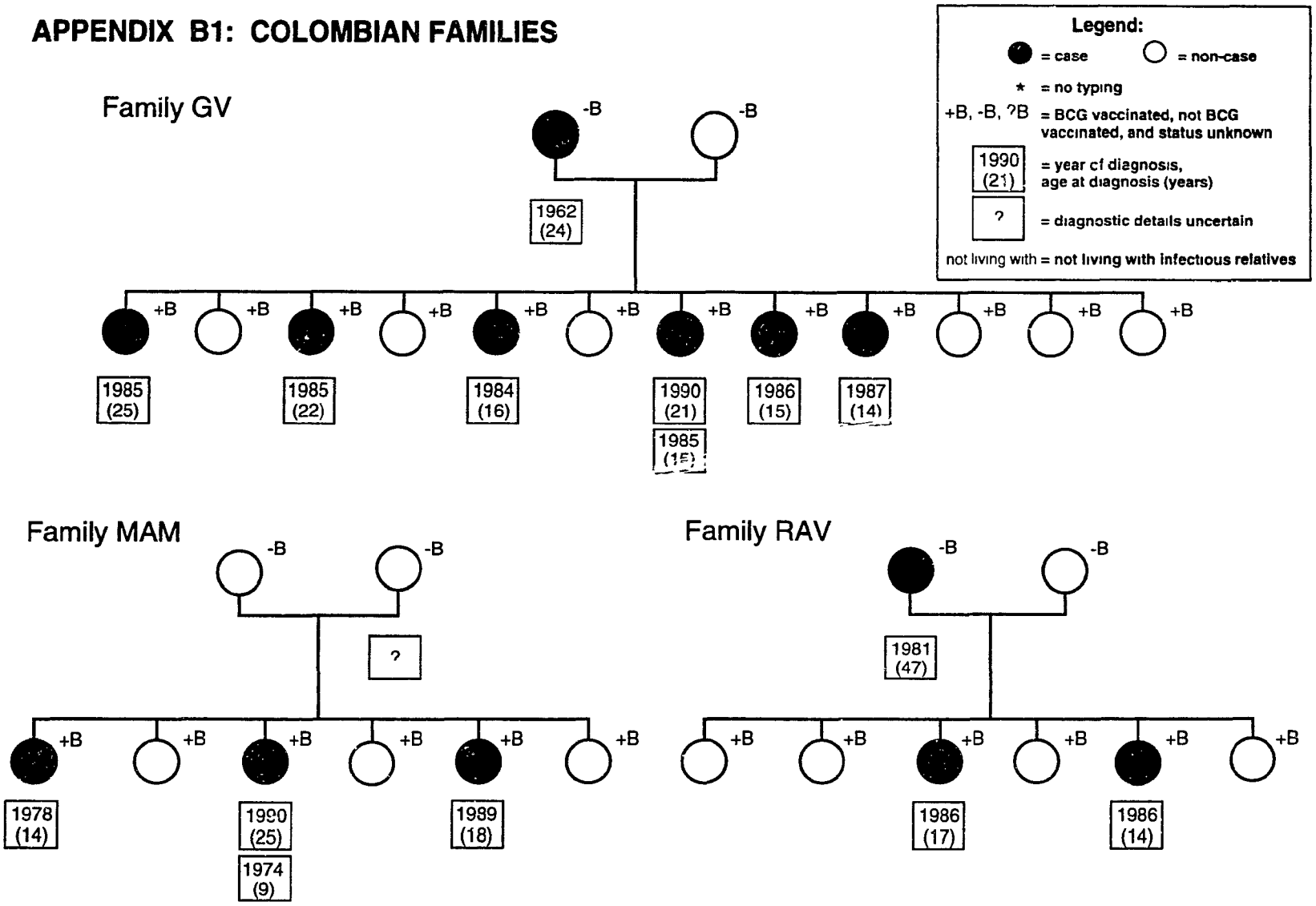
B3 Canadian family from 1989 fieldwork

B4 Canadian family from 1993 data

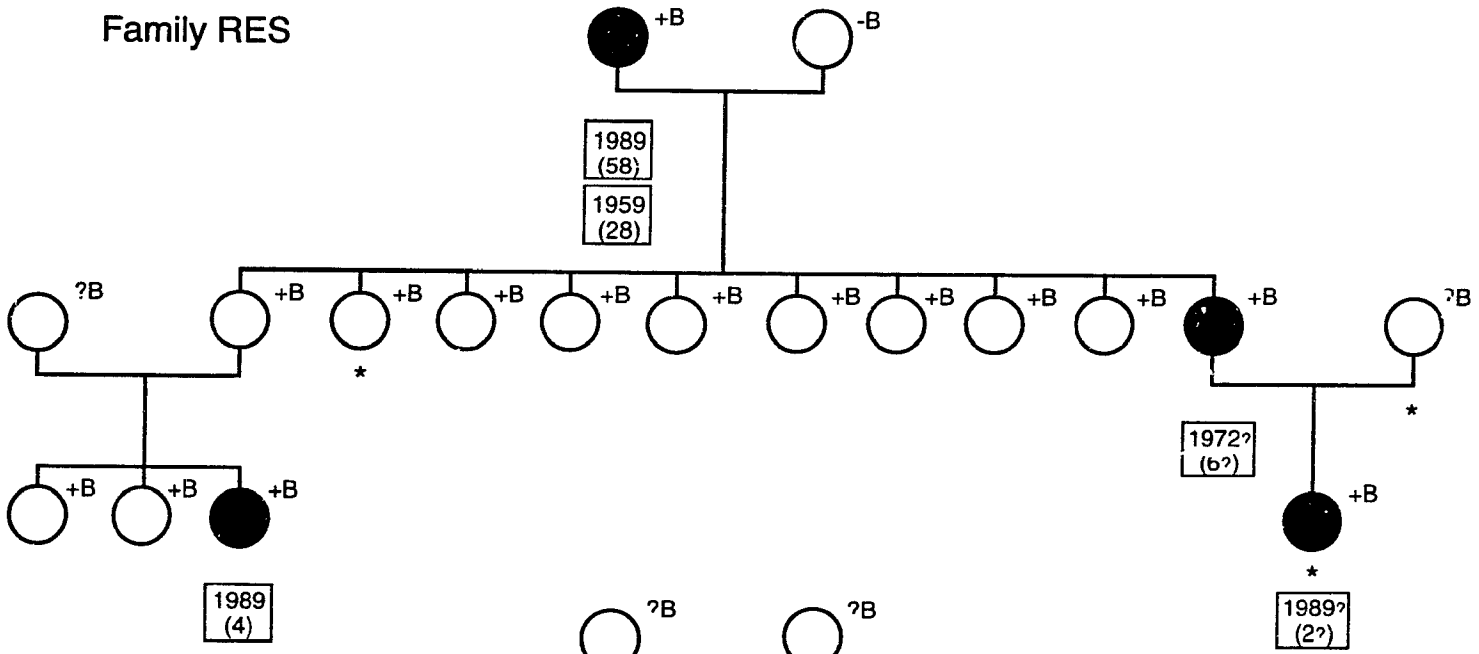
B5 Distribution of cases by diagnostic methods for affected family members analyzed for linkage in each geographical region

Note.
For reasons of confidentiality, gender information and identification numbers are not displayed in the pedigree diagrams

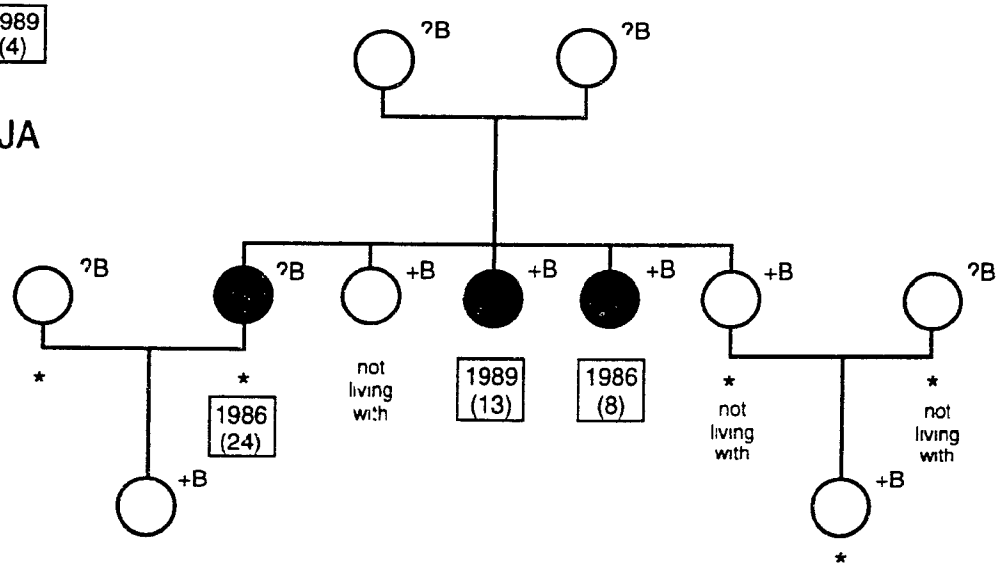
APPENDIX B1: COLOMBIAN FAMILIES



Family RES

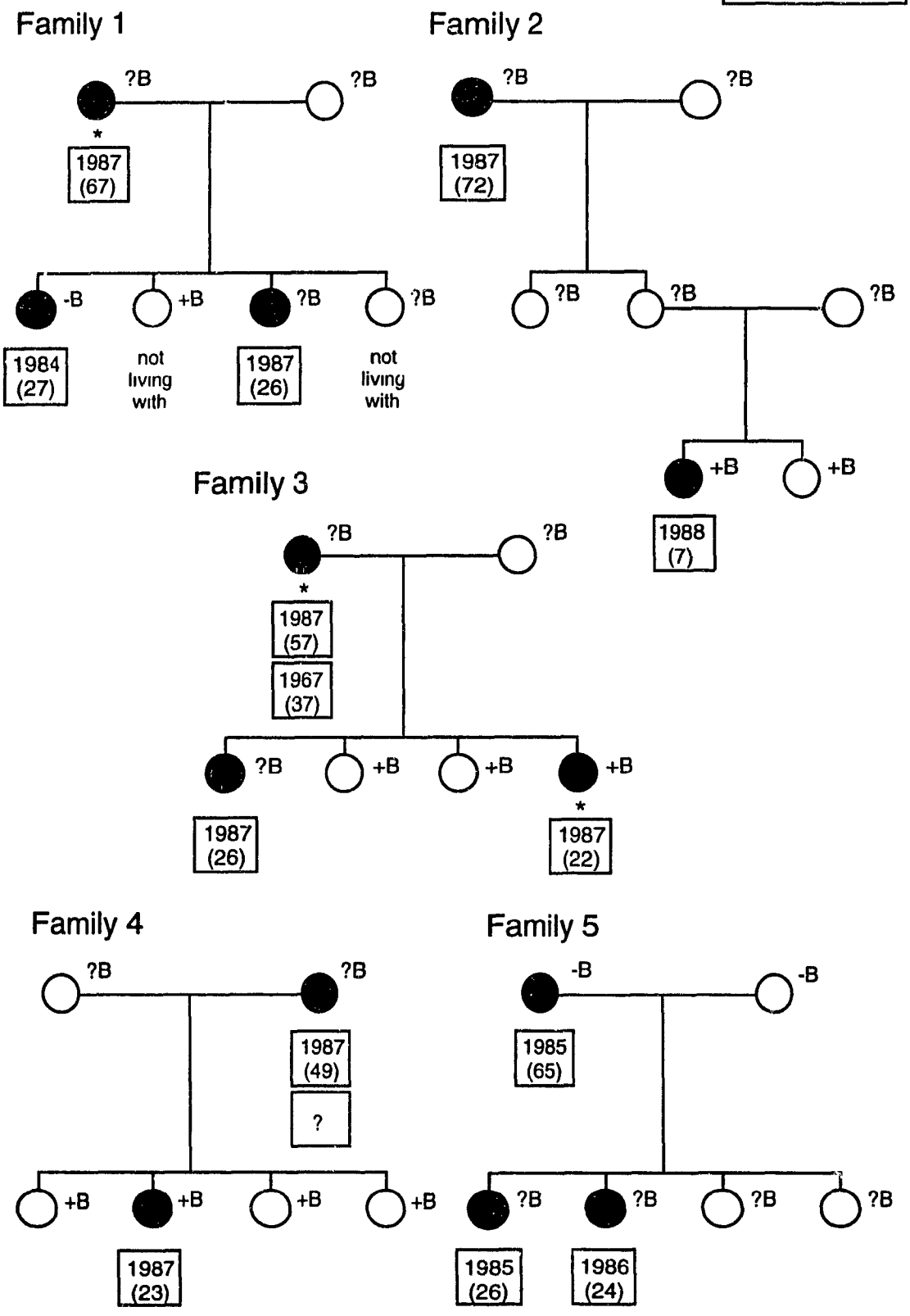


Family RJA

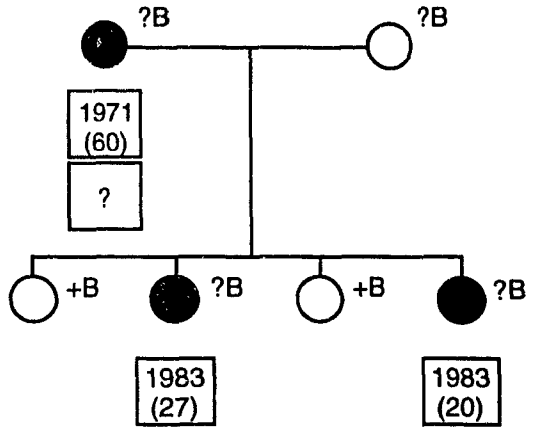


APPENDIX B2: HONG KONG PEDIGREES

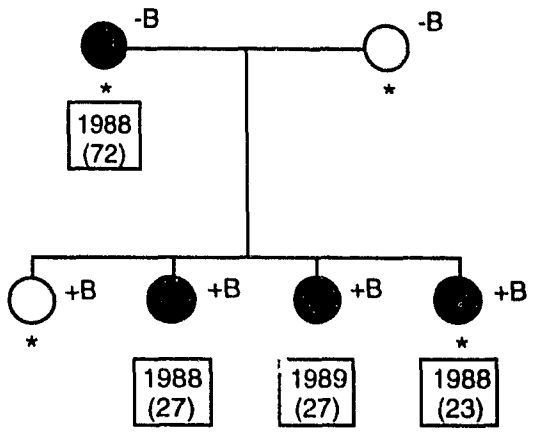
Legend as for Appendix B1



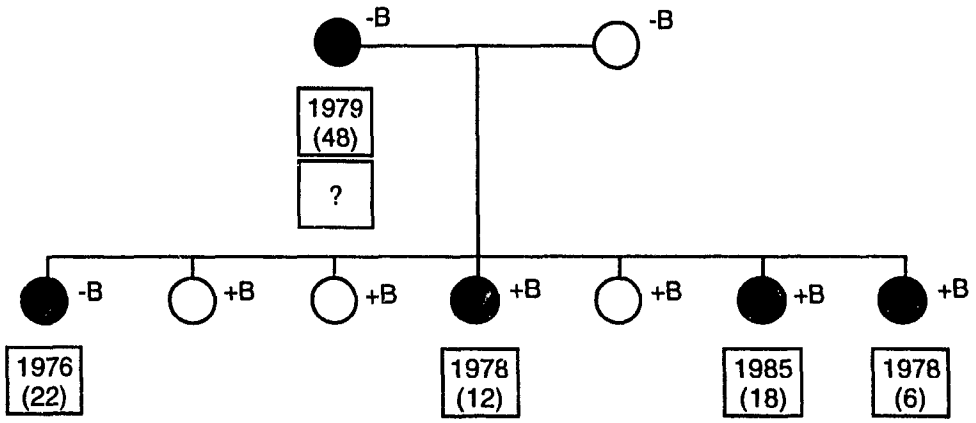
Family 6



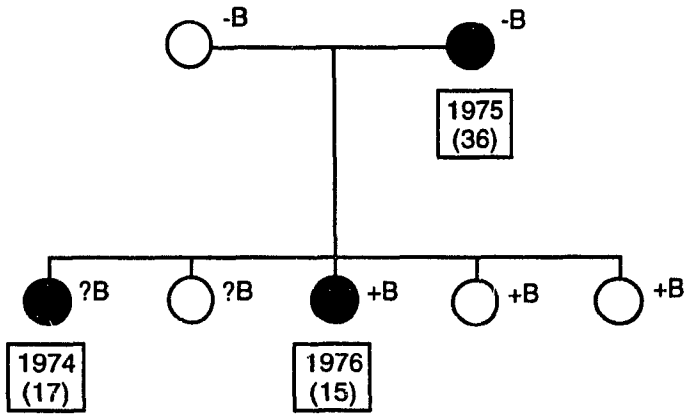
Family 7



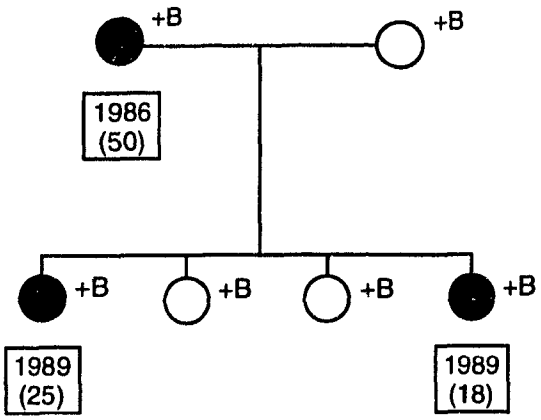
Family 8



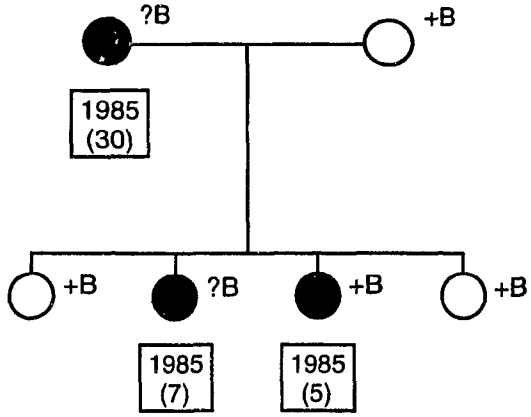
Family 9



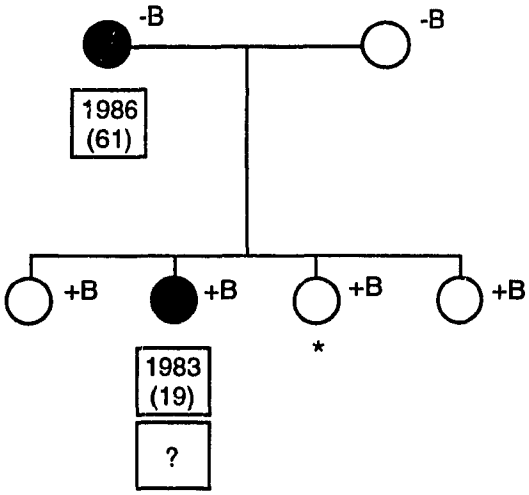
Family 10



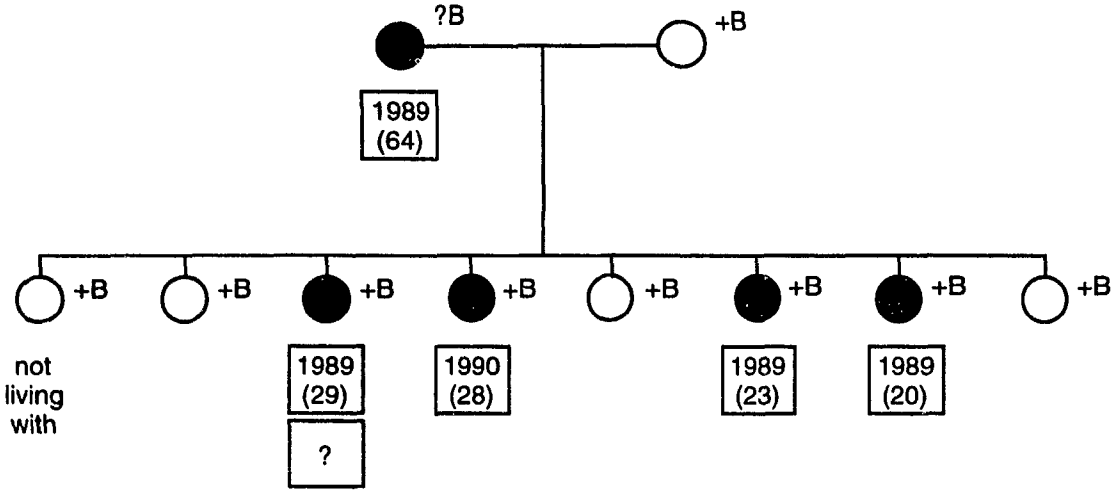
Family 11



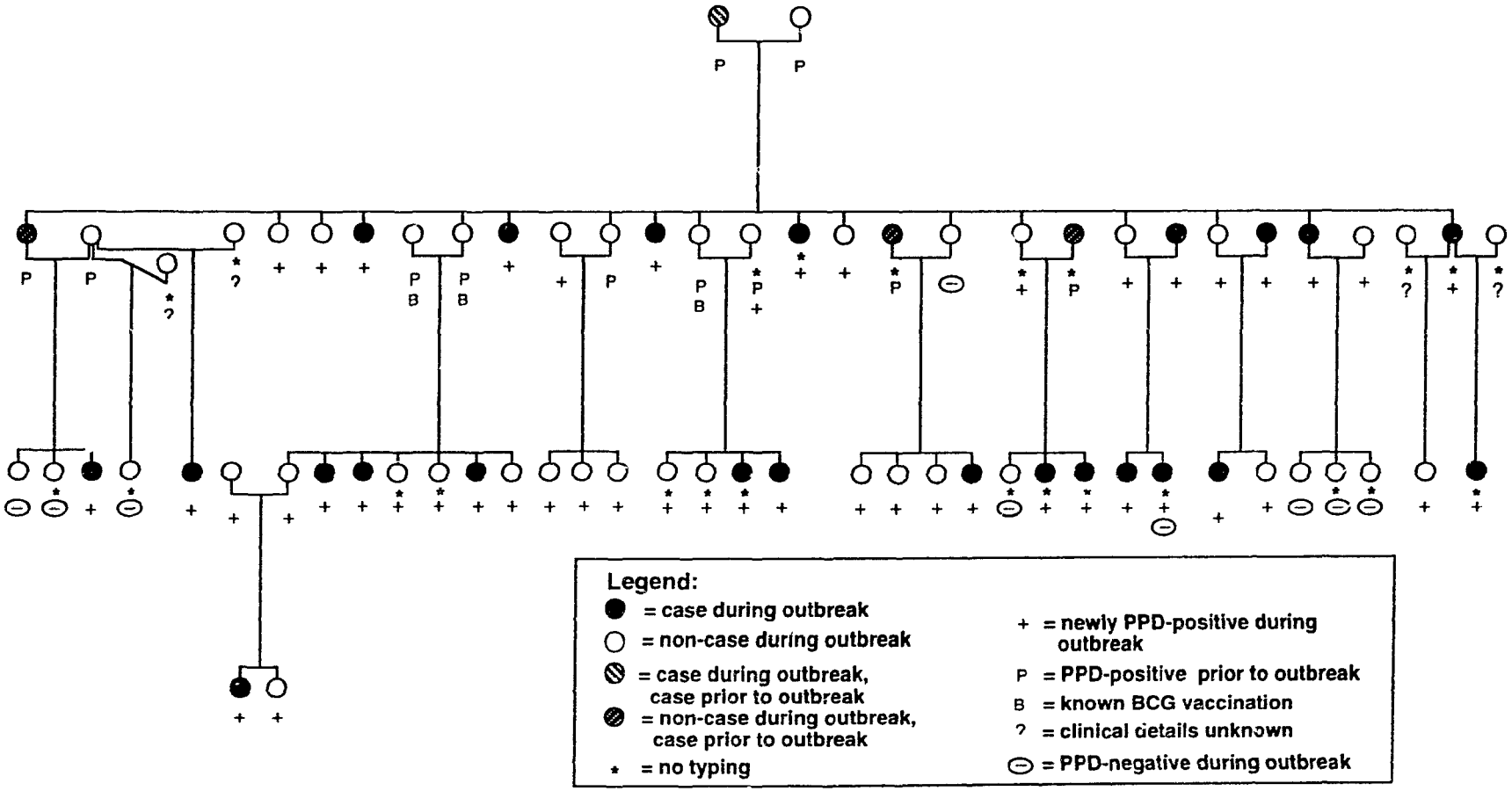
Family 12



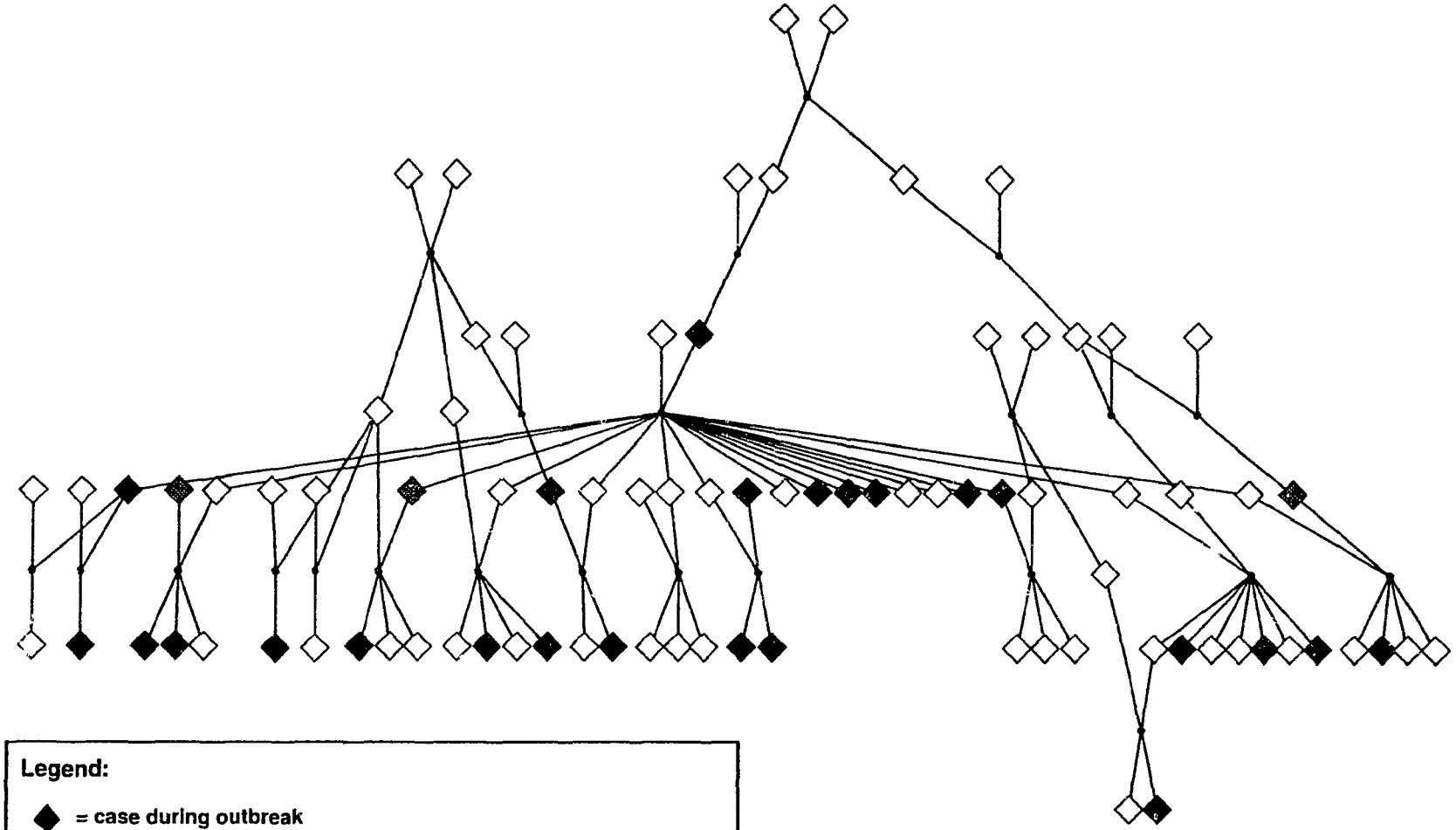
Family 13



APPENDIX B3: CANADIAN FAMILY FROM 1989 FIELDWORK



APPENDIX B4: CANADIAN FAMILY FROM 1993 DATA



Legend:

- ◆ = case during outbreak
  (note that #1 was also a case prior to outbreak)
- ◇ = non-case during outbreak
- ◈ = case prior to outbreak, non-case during outbreak

APPENDIX B5

Distribution of cases by diagnostic methods for affected family members analyzed for linkage in each geographical region

Geographical region (number of cases)	Diagnostic method				
	Smear- and culture-positive	Smear- or culture-positive (type: S=smear C=culture)	Abnormal chest X-ray (*smear- and/or culture-negative; **no microbiologic tests done)	Clinical symptoms only (no other tests)	Previous disease by historical records
Canada (22)	6	7 (7C)	6*	0	3
Colombia (15)	0	10 (9S, 1C)	1*	2	2
Hong Kong (39)	10	5 (2S, 3C)	20 (18*, 2**)	4	0

sources:
For the Canadian family: Miller (1991; personal communication), Fanning (personal communication), Mah and Fanning (1991), fieldwork notes (May, 1994)
For the Colombian and Hong Kong families: international questionnaire data

APPENDIX C

Adjustment to the preliminary model

The source for the penetrance estimate for the r/r genotype in the preliminary models was the Kallmann and Reisner (1943) twin study. The penetrance was estimated directly as the monozygous twin tuberculosis concordance rate (unadjusted for age), which was 62%. In a twin study, the monozygous twin disease concordance rate is the proportion of monozygous co-twins of an index case who have been affected with the disease. If the twin study begins with the ascertainment of cases and not twins (from, for instance, a population-based twin registry), a proportion of twins who are both unaffected in the population are not ascertained. In this situation, which was the case with the Kallmann and Reisner study, the penetrance (the conditional risk of disease given genotype) is not directly equal to the monozygous twin disease concordance rate. A more appropriate estimate of the penetrance for the preliminary models using the chosen source takes into account the sample ascertainment method, and is derived as follows.

Under the assumption of absence of phenocopies and recessive inheritance of susceptibility,

$$C = \frac{r^2}{r^2 + 2r(1 - r)}$$

where C is the monozygous twin disease concordance, r is the penetrance of the susceptible genotype, r^2 is the probability that both twins are affected, and $2r(1 - r)$ is the probability that only one twin is affected.

Solving for r,

$$C = \frac{r}{r + 2(1 - r)}$$

$$r = \frac{2C}{1 + C}$$

In the Kallmann and Reisner (1943) study, the monozygous twin concordance was 61.5% (unadjusted for age). An unbiased estimate of penetrance from this study would be the following:

$$r = \frac{2(0.615)}{1 + 0.615}$$

$$r = 0.76$$

Therefore, the estimate of penetrance for the r/r genotype in the preliminary models would be 76%, when adjusted for mode of ascertainment in the source study.

Tables 1, 2 and 3 display lod score results summed across the Colombian and Hong Kong families (at each recombination fraction) and for the Canadian family, under an adjusted preliminary model with 76% penetrance for the r/r genotype (no phenocopies; $q = 0.2$). The results are shown for selected markers with minimum lod scores ≤ -2 or maximum lod scores ≥ 1 in two-point linkage analysis with the disease trait. Figure 1 presents lod score curves for selected markers under the same model. Figures 2 and 3 show lod scores for linkage between the disease trait and the TNP-1 C marker as a function of susceptibility and marker allele frequency, in the Canadian pedigree with and without phenocopies. The penetrance of the higher risk r/r genotype was 76% for these analyses. For the phenocopy models, the penetrance of the lower risk R/R and R/r genotypes was 0.76/24 or 0.032 (3.2%), assuming a 50% phenocopy rate in the population. Figure 4 shows the effect of the addition of marriage and consanguinity loops to the Canadian pedigree on the lod scores for linkage between disease and TNP-1 C under the models with and without phenocopies.

Table 1. Lod score results under the adjusted preliminary model (no phenocopies) for selected markers, Colombian and Hong Kong families

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	**-2.11**	**-2.05**	-1.07	-0.28	0.28	0.29	0.11
CRYGP1-A	**-3.25**	**-2.99**	-1.90	-1.14	-0.43	-0.14	-0.03
CRYGP1-C	**-3.06**	**-2.61**	-1.72	-1.16	-0.54	-0.22	-0.05
FN61039-M	**-2.35**	**-2.04**	-1.27	-0.74	-0.22	-0.04	0.00
TNP-1 A	**-3.48**	**-3.03**	**-2.00**	-1.32	-0.59	-0.23	-0.06
TNP-1 C	**-7.97**	**-6.27**	**-3.86**	**-2.51**	-1.14	-0.45	-0.09
VIL 7 No. 6	**-5.60**	**-4.74**	**-3.23**	**-2.14**	-1.00	-0.41	-0.10
VIL E-62	**-6.40**	**-5.59**	**-3.68**	**-2.22**	-0.81	-0.24	-0.04
VIL E-84	**-4.96**	**-3.91**	**-2.47**	-1.61	-0.72	-0.29	-0.07
VIL pEX2(a)	**-5.23**	**-4.48**	**-3.40**	**-2.35**	-1.04	-0.40	-0.09
DESMIN	**-4.44**	**-3.41**	**-2.21**	-1.48	-0.70	-0.28	-0.07
INHB-A1	**-5.00**	**-4.13**	**-2.88**	**-2.04**	-1.02	-0.42	-0.10

Note. Lod scores were summed across all families at each recombination fraction. Indicated in bold print are minimum lod scores that significantly exclude linkage at the specified recombination fraction. Markers are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the RFLP map in section 4.5.

Table 2. Lod score results under the adjusted preliminary model (no phenocopies) for selected RFLP markers, Canadian family

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-D	-1.06	-0.52	0.32	0.75	**0.99**	0.82	0.39
FN61039-H	**-3.37**	**-2.24**	-0.96	-0.38	0.07	0.14	0.06
TNP-1 B	**2.63**	2.61	2.51	2.33	1.87	1.28	0.57
TNP-1 C	**3.46**	3.41	3.18	2.88	2.19	1.44	0.62
INHB-A1	**-1.98**	-1.46	-0.86	-0.55	-0.23	-0.06	0.01

Table 3. Lod score results under the adjusted preliminary model (no phenocopies) for selected microsatellite markers, Canadian family

Disease with marker	Lod score at specified recombination fraction						
	0.0	0.01	0.05	0.1	0.2	0.3	0.4
CRYG1-A	0.40	1.13	1.60	**1.68**	1.50	1.09	0.51
D2S157	**-6.14**	-0.74	0.46	0.87	**1.03**	0.84	0.42
D2S128	**1.69**	1.68	1.63	1.55	1.28	0.90	0.40
D2S137	**-5.90**	-1.40	-0.10	0.44	**0.79**	0.70	0.34
D2S173	**-6.00**	-1.84	-0.45	0.13	**0.50**	0.45	0.18
D2S120	**-7.29**	**-3.87**	-1.52	-0.67	-0.12	0.00	-0.05
D2S126	**-3.95**	**-2.07**	-0.30	0.38	**0.75**	0.63	0.28
PAX-3	**-5.54**	**-3.74**	**-1.99**	-1.09	-0.34	-0.08	-0.05
D2S172	**-5.54**	**-2.80**	-1.66	-0.98	-0.37	-0.14	-0.07
D2S125	**-2.98**	**-2.99**	**-3.03**	**-2.70**	-1.47	-0.73	-0.28
D2S140	**-2.94**	**-2.67**	**-1.97**	-1.44	-0.79	-0.37	-0.12

Note. Indicated in bold print are maximum lod scores in favour of linkage and minimum lod scores that significantly exclude linkage at the specified recombination fraction. Markers are listed in order proximal to distal along the long arm of chromosome 2, in accordance with the microsatellite map in section 4.5.

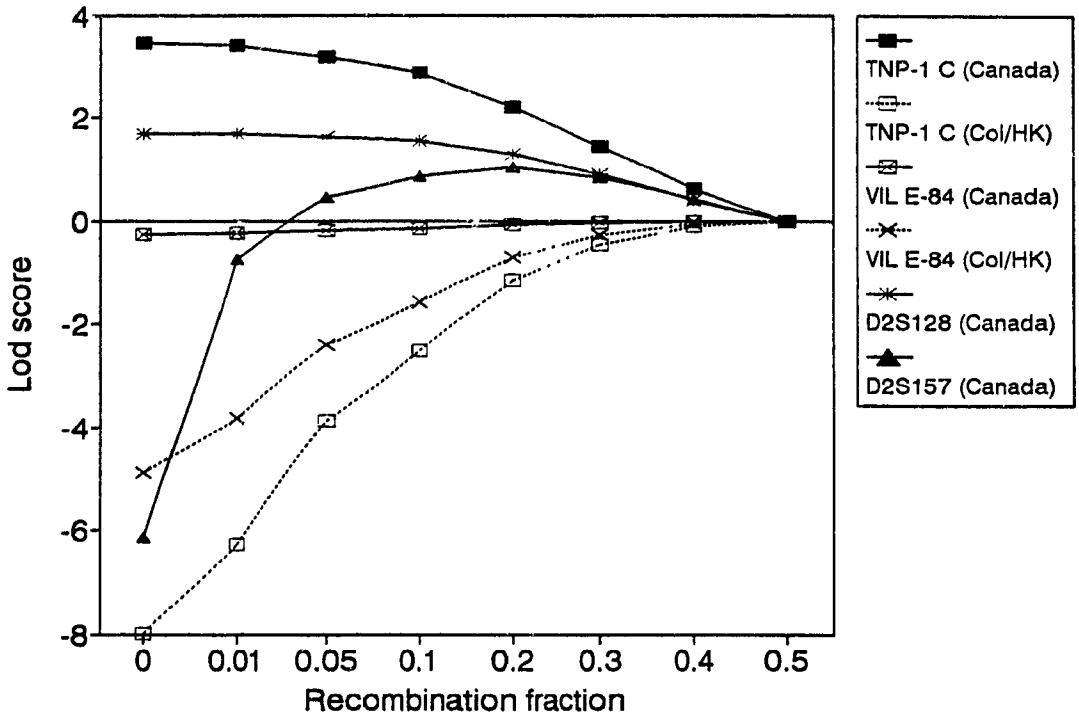


Figure 1. Lod score curve for selected markers under the adjusted preliminary model (no phenocopies)

Notes.

Canada refers to results with the Canadian pedigree

Col/HK refers to results summed over the Colombian and Hong Kong pedigrees

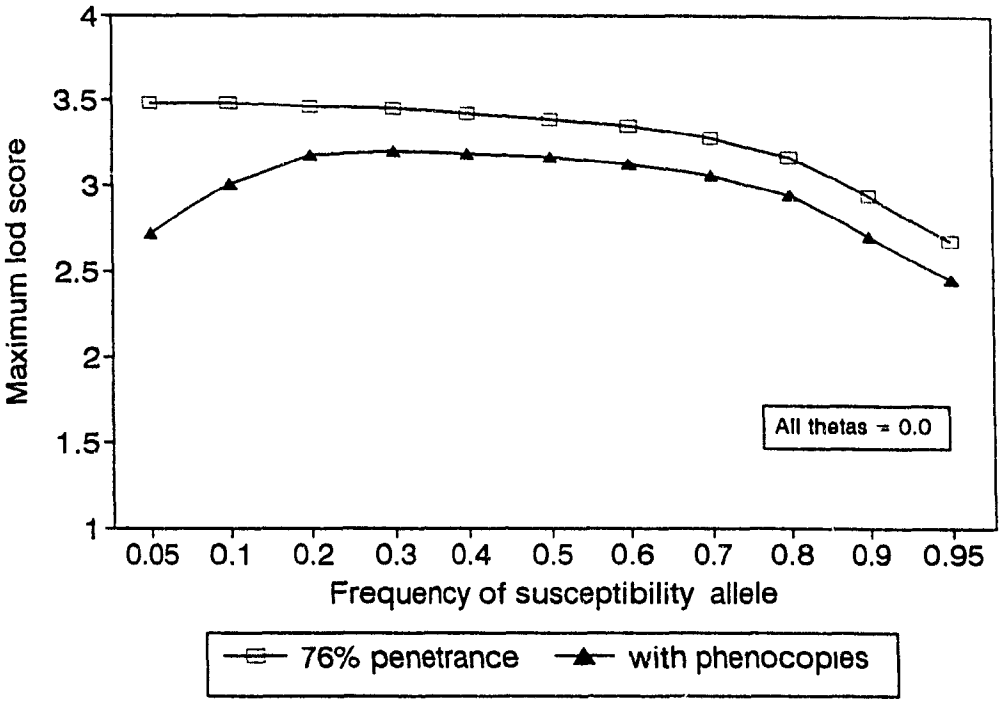


Figure 2. Lod scores for linkage between disease and TNP-1 C as a function of susceptibility allele frequency

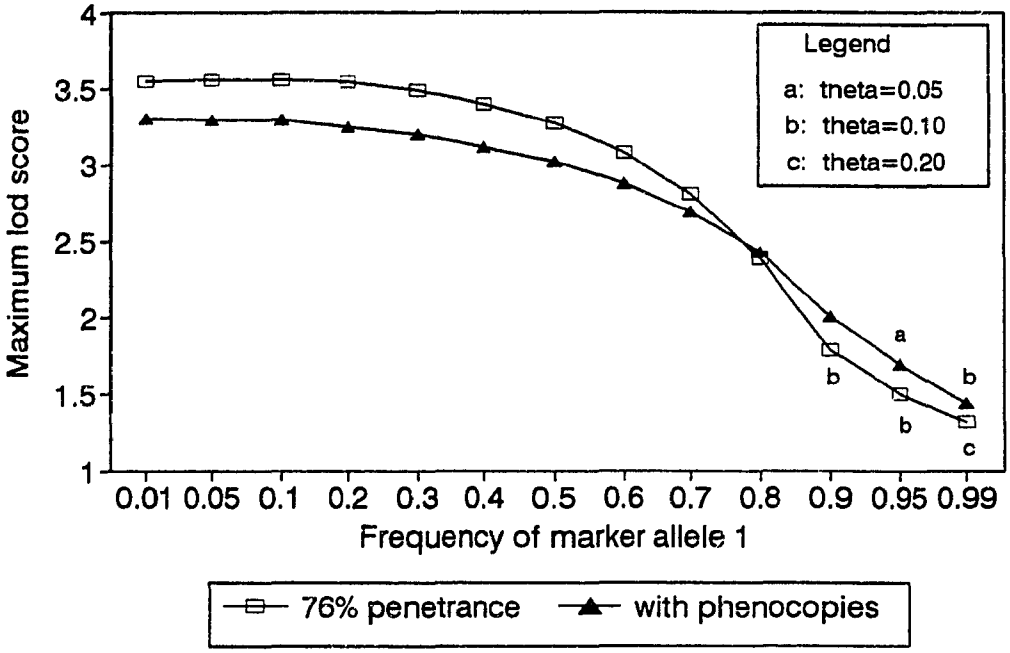


Figure 3. Lod scores for linkage between disease and TNP-1 C as a function of marker allele frequency

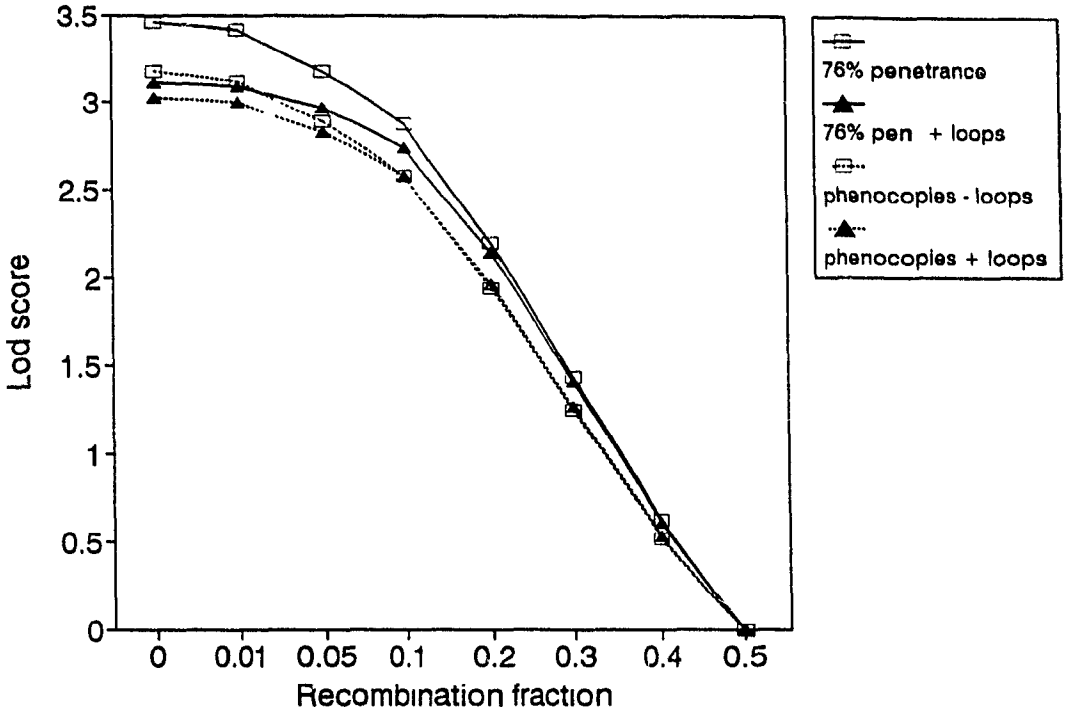


Figure 4. Effect of marriage and consanguinity pedigree loops on lod scores for linkage between disease and TNP-1 C

APPENDIX D

Lod scores for linkage between disease and TNP-1 C under the three epidemiological models in the Canadian pedigree

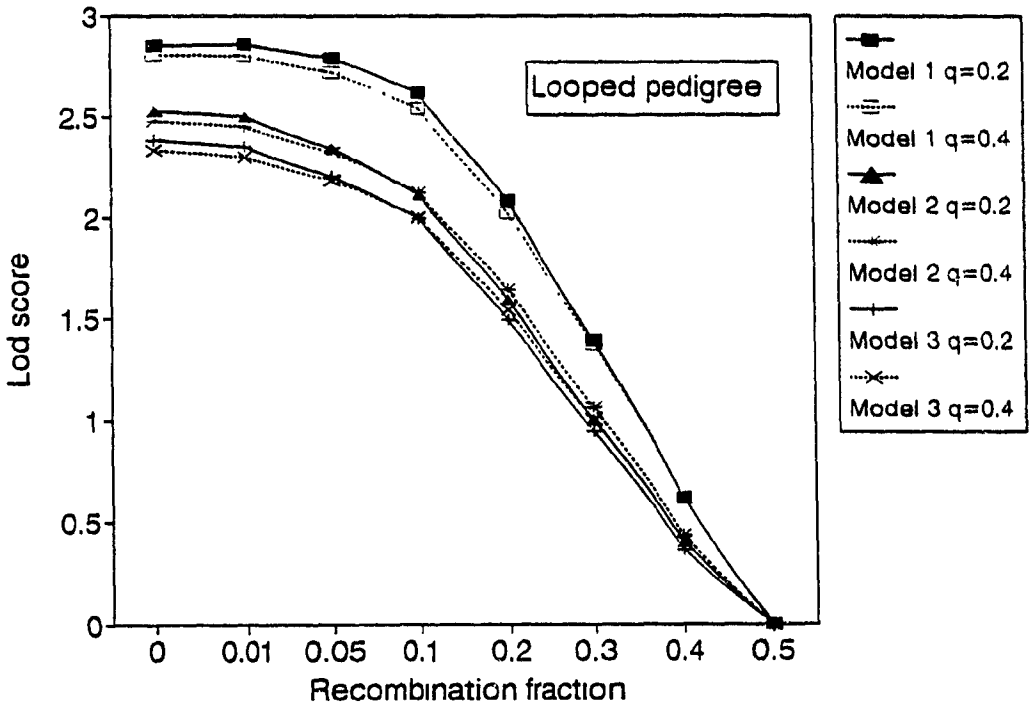


Figure 1. Lod scores for linkage between disease and TNP-1 C under the epidemiological models in the Canadian pedigree with marriage and consanguinity loops

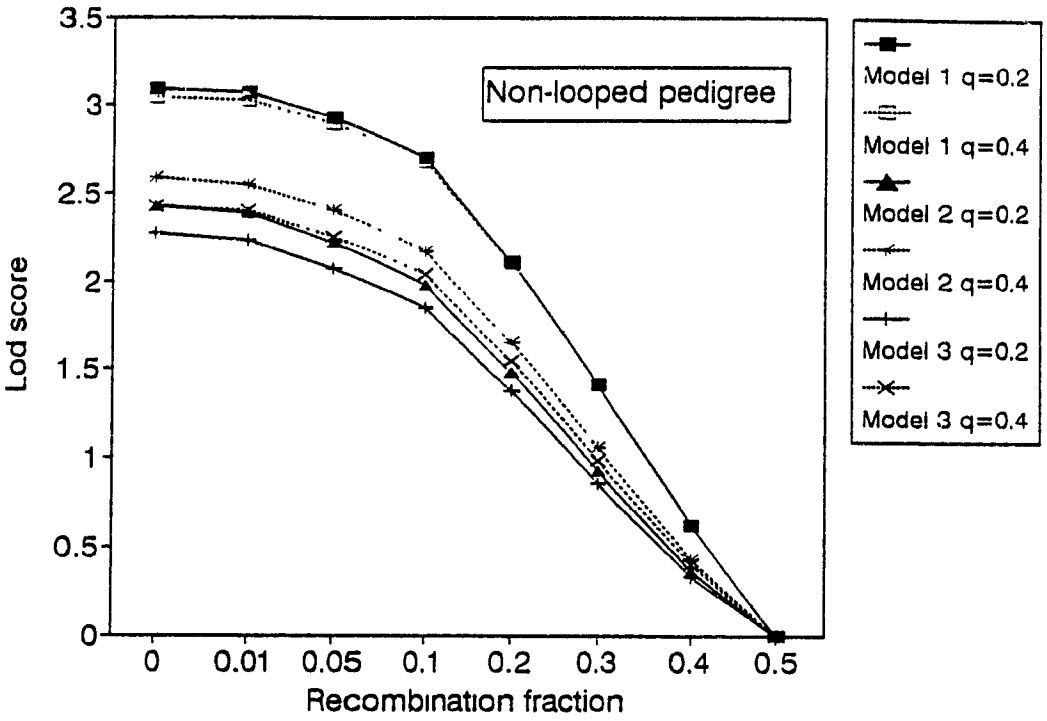


Figure 2. Lod scores for linkage between disease and TNP-1 C under the epidemiological models in the Canadian pedigree without loops

APPENDIX E

Empirical significance level of the lod score for linkage between disease and TNP-1 C, maximized over penetrance values

Methods

According to methods suggested by Ott (1991), computer simulation was used to estimate the significance level of the maximum lod score computed for disease and TNP-1 C in the Canadian non-looped pedigree, as the penetrance of the r/r genotype was varied from 50–95% (no phenocopies). Using the SLINK program (Weeks and Ott, 1993; Ott, 1989), the TNP-1 C data was simulated for the pedigree 1000 times conditional on the specified marker allele frequencies, under the assumption of absence of linkage between the trait and the marker locus (the null hypothesis). The marker allele frequencies used were estimated with unrelated founders in 12 diabetes families (n = 24), as elsewhere in this thesis. Each of the 1000 simulation replicates of marker genotype assignments was analyzed for linkage with recessive inheritance of susceptibility, $q = 0.2$, and no phenocopies, as the penetrance for the r/r genotype was varied from 50–95% in increments of 5%. The MLINK program was set up on a Sun SPARC station by Dr. Morgan to calculate lod scores for each replicate under each penetrance value, with the recombination fraction varied from 0.0 to 0.49 in increments of 0.01. The maximum lod score generated with each penetrance and the recombination fraction at which it occurred were found for each replicate. Of the ten maximum lod scores computed for each replicate, the largest (maximized-over-models) lod score was determined. The maximized lod score observed over the ten penetrance values with the observed pedigree data was compared with a distribution of the maximized lod score for each of the 1000 replicates.

Results

For the observed pedigree, the maximized lod score over the ten penetrance values for disease and TNP-1 C was 3.46 at 0% recombination. For the 1000 simulated replicates of marker genotype data, a maximized-over-models lod score of 2.88 arose once (Figure 1); this result was computed under a model with 95% penetrance for the r/r genotype. All other maximized lod scores were less than 2.0. The point estimate of the significance level of the observed maximum was 0.0 with a 95% confidence interval of (0.0, 0.003),

because the upper limit of the confidence interval is $1 - \alpha^{1/n}$ and $\alpha = 0.05$ (Ott, 1991). According to Ott (1991), the significance level associated with a maximized-over-models lod score is considered statistically significant when the upper limit of its confidence interval is less than 0.001 (the traditional 1000:1 odds for linkage). In order to meet this criterion with a 95% confidence interval 3000 replicates would have to have been analyzed.

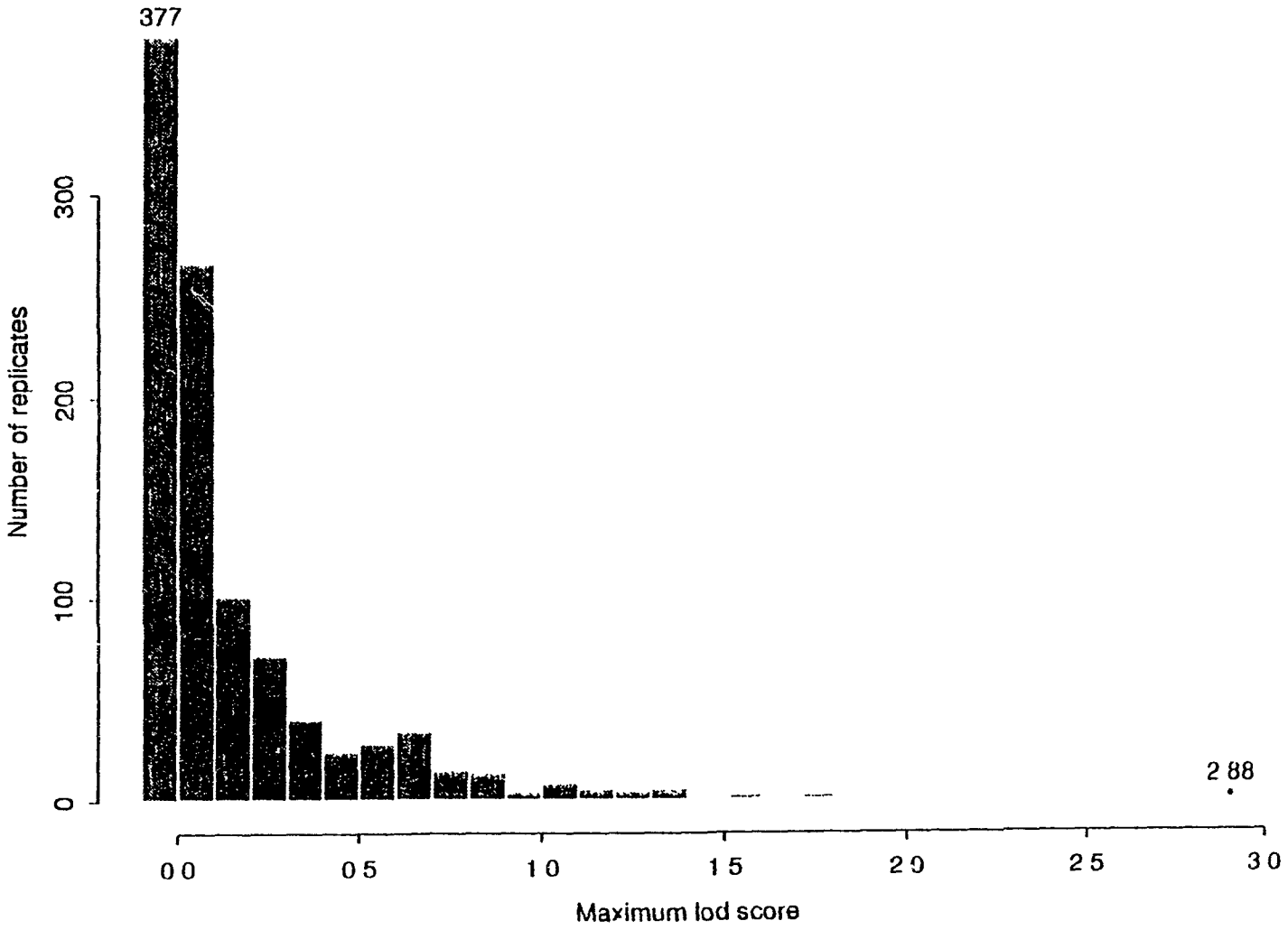


Figure 1. Lod scores maximized over 10 penetrance values for the r/r genotype in 1000 replicates (no phenocopies)