Predictive Genomics in Asthma Management Using Probabilistic Graphical Models

by

Blanca Elena Himes

BS in Physics with Specialization in Computational Physics
University of California, San Diego, 2001

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Medical Physics and Bioinformatics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

© Blanca Elena Himes, MMVII. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly
paper and electronic copies of this thesis document in whole or in part.

Author .....
Harvard-MIT Division of Health Sciences and Technology
May 1, 2007

Certified by .....
Marco F Ramoni, PhD
Assistant Professor of Health Sciences and Technology
Thesis Supervisor

Accepted by .....
Martha L Gray, PhD
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 10 2007
LIBRARIES

ARCHIVES

Predictive Genomics in Asthma Management Using Probabilistic Graphical Models

by

Blanca Elena Himes

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on May 1, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Medical Physics and Bioinformatics

Abstract

Complex traits are conditions that, as a result of the complex interplay among genetic and environmental factors, have wide variability in progression and manifestation. Because most common diseases with high morbidity and mortality are complex traits, uncovering the genetic architecture of these traits is an important health problem. Asthma, a chronic inflammatory airway disease, is one such trait that affects over 300 million people around the world. Although there is a large amount of human genetic information currently available and expanding at a rapid pace, traditional genetic studies have not provided a concomitant understanding of complex traits, including asthma and its related phenotypes. Despite the intricate genetic background underlying complex traits, most traditional genetic studies focus on individual genetic variants. New methods that consider multiple genetic variants are needed in order to accelerate the understanding of complex traits.

In this thesis, the need for better analytic approaches for the study of complex traits is addressed with the creation of a novel method. Probabilistic graphical models (PGMs) are a powerful technique that can overcome limitations of conventional association study approaches. Going beyond single or pairwise gene interactions with a phenotype, PGMs are able to account for complex gene interactions and make predictions of a phenotype. Most PGMs have limited scalability with large genetic datasets. Here, a procedure called phenocentric Bayesian networks that is tailored for the discovery of complex multivariate models for a trait using large genomic datasets is presented. Resulting models can be used to predict outcomes of a phenotype, which allows for meaningful validation and potential applicability in a clinical setting.

The utility of phenocentric Bayesian networks is demonstrated with the creation of predictive models for two complex traits related to asthma management: exacerbation and bronchodilator response. The good predictive accuracy of each model is established and shown to be superior to single gene analysis. The results of this work demonstrate the promise of using the phenocentric Bayesian networks to study the genetic architecture of complex traits, and the utility of multigenic predictive methods compared to traditional single-gene approaches.

Thesis Supervisor: Marco F Ramoni, PhD
Title: Assistant Professor of Health Sciences and Technology

Acknowledgments

I express my deepest gratitude to Marco Ramoni, Zak Kohane, and Scott Weiss for their help and support in completing this thesis and in helping with my graduate career. Each of them has provided me with an enviable graduate experience. Many thanks to Amy Berninger, who helped with data analysis, and everyone in the Channing Laboratory who helped to gather the genetic data used in this work, especially Barbara Klanderman. Helpful feedback and discussions were provided by lab members, including Rachel Badonivac, Amanda Sedgewick, and Gil Alterovitz.

I appreciate the constant help and support of my friends during graduate school. The experience was made orders of magnitude more enjoyable by them. Among those who have been with me through the good and the bad are Jenny Mu, Kathleen Sienko, Francisca Leite, Dana Hunt, and Grace Kim. Thank you to my siblings Melanie, Nancy, and William Zauscher, and to my parents Diane and Fernando Zauscher and Dave and Tonya Himes for their love and constant support. I am grateful to my husband, Nate Himes, for helpful discussions and feedback, but most of all for providing me with encouragement and love every day.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Uncovering the genetic architecture of complex traits is an important current problem whose solution would dramatically improve the health of people around the world. Complex traits are conditions with wide variability in symptoms and seemingly different mechanisms of onset and progression. Although they are heritable, they cannot be easily explained in terms of a single or few genes, and their expression is altered strongly by environmental factors. Intensely studied complex traits include aspects of cardiovascular disease, Type 2 diabetes mellitus, asthma, obesity, autism, Huntington's disease, Parkinson's disease, and Alzheimer's disease. Genes related to most of these diseases have been found, but few of them have helped in making advances in treatment and prevention.

The large amount of human genetic information that is currently available and continues to expand at a rapid pace has not resulted in a concommitant understanding of complex traits. This disappointing fact has been partly attributed to the difficulty of studying complex traits with traditional genetic techniques [1]. Association studies, the most common genetic studies, attempt to find genes that are more prevalent in individuals with a disease and therefore likely to be involved in causing the disease. A great number of association studies have been published, but their results are often never duplicated [2, 3, 4]. This lack of reproducibility has led to skepticism regarding the utility of association studies. Some have postulated that in their current form, association studies will never be able to explain complex traits because only models that account for complex gene-gene and gene-environment interactions are suitable for this task [5, 6]. Despite the difficulty of studying

complex traits, the growing availability of genetic data makes the development of better methods necessary for the elucidation of the genetic architecture of complex traits.

One of the main limitations of traditional association studies is that they investigate the relationship of single genes or assume additive effects of single genes to a phenotype. Because complex traits are caused by multiple genes, which likely interact epistatically, methods that take into account the complex interaction of many genes are likely to be more useful than one-gene-at-a-time approaches. The results of traditional association studies are lists of genes that are significant based on a statistical threshold. Validation studies involve replicating initial findings, which does little more than to increase the sample size of the original result. In most cases, results of gene association studies are not been used in a clinical setting or to motivate further biological studies. Lacking from these studies is a result that can be applied to an individual such as a quantitative risk for a phenotype or disease. Being able to give a probability for risk of a disease is highly useful in a clinical setting. Quantitative measures for an outcome are readily interpretable and decisions about how to change an individual's life can be confidently made. Predictive models give such quantitative results, assigning a probability for an outcome based on observed data. They are validated by assesing their predictive accuracy on independent populations. In addition to providing useful and testable results, multivariate predictive models are likely to provide a thorough explanation of the biological variability that is responsible for a complex phenotype by taking into account many uncommon pathways.

Probabilistic graphical models (PGMs) are a powerful technique that can overcome limitations of conventional association study approaches. Going beyond single or pairwise gene interactions with a phenotype, probabilistic graphical models are able to account for complex gene interactions. Addtionally, they can be used to make predictions of a phenotype of interest for individual subjects that allows for ascertainment of their validity. Bayesian networks, a common PGM approach, are regarded as an emerging paradigm for the analysis of complex traits [7, 8]. They have been successfully used to study gene expression data [9], protein-protein interactions [10], and pedigree analysis [11]. They also have been used to model the multigenic association and predict the occurrence of stroke in sickle cell anemia patients, demonstrating their suitability to understand the genetic basis of complex traits and predict a clinical phenotype [12]. Unfortunately, most Bayesian network discovery algorithms have limited ability to handle the large genetic datasets that are currently available.

Novel methods that are tailored for gene association are necessary for the use of Bayesian networks in large candidate gene and whole genome studies.

Asthma, a chronic inflammatory airway disease, is a serious global problem affecting 20.5 million Americans and over 300 million people around the world [13, 14]. Both the prevalence and death rate of asthma rose dramatically in the US and globally between 1960 and 2001, and have remained at high levels or continued to increase since then [15, 13, 16]. Asthma is a costly disease, as demonstrated by the increased risk of emergency room visits, hospitalization, and sick absences that are associated with it [17, 18]. Over $16 billion are spent yearly in the US on asthma-related healthcare expenses [13]. Asthma has a demonstrable genetic basis, with heritability estimates ranging from 0.36 to 0.87 [19, 20, 21, 22, 23, 24, 25]. and over 100 genes individually associated with asthma or a related phenotype [26].

Two important aspects of asthma management are exacerbations and bronchodilator response. Asthma exacerbations, commonly known as asthma attacks, are the major cause of morbidity, mortality and healthcare costs for individuals with asthma [27, 28, 29]. Exacerbation episodes involve worsening of asthma symptoms, including shortness of breath, cough, wheezing, chest pain or tightness, mucus production, or some combination of these. Uncovering the genetic basis underlying asthma exacerbations would be helpful to understand the biology of exacerbations, discover novel therapeutic targets, and identify those at risk of suffering from them.

A common clinical test that is used for the evaluation of reversible airway obstruction and the diagnosis of asthma is the bronchodilator response test. The basis of this test is to find out whether administration of a bronchodilator medication improves FEV_1. The most potent and rapidly acting bronchodilators currently available for clinical use are β_2-agonists [30]. They are used not only for bronchodilator tests, but as routine asthma therapy, despite the interpatient variability in their efficacy. Evidence for the genetic basis of bronchodilator response has been establihsed in family aggregation and gene association studies. A thorough understanding of the genetic basis of bronchodilator response would be helpful to identify patient-specific treatments, identify novel therapeutic targets, and help in the diagnosis and monitoring of asthma. Further, such a test would help establish which patients are responsive to β_2-agonists and what genetic mechanisms may be responsible for variability in patient response.

1.2 Thesis Overview

The goal of this work is to create a new Bayesian network discovery approach that can be used to study complex diseases with large genetic datasets and use the new method to create predictive models of two complex phenotypes related to asthma management: exacerbation and bronchodilator response.

In Part I, the use of PGMs in the study of complex traits is explored. Complex traits are defined, and traditional methods used to study their genetic basis are explored [Chapter 2]. The focus is on case-control gene association studies, as this approach is the most common in the investigation of complex traits. After noting the limitations of traditional association studies, PGMs are introduced and a review of their use in genetic applications is provided [Chapter 3].

A novel approach to learning Bayesian networks for the study of complex traits is described in Part II. First, Bayesian networks are explored in depth [Chapter 4], including how networks are learned from data, and how they are used to make predictions. After describing the limitations of Bayesian networks in genomic scale association studies, a new method to learn them is proposed: phenocentric Bayesian networks [Chapter 5]. This method takes advantage of a main goal of gene association studies: the desire to predict the risk for a phenotype or disease in individual subjects.

Asthma and asthma management are surveyed in Part III. Chapter 6 provides the necessary background on asthma to appreciate the complexities in defining and understanding this disease. After establishing the importance of the genetic component of asthma, features of asthma management are introduced [Chapter 7]. In particular, aspects of asthma exacerbation and bronchodilator response are addressed to understand them as complex traits.

Genetic tests for prediction of asthma exacerbation and bronchodilator response are constructed in Part IV. After describing the study population and genetic data used [Chapter 8], phenocentric Bayesian networks are used to learn predictive models of these aspects of asthma management. Details of each model are provided in Chapters 9 and 10. The predictive accuracy of the models is established as good and shown to have advantages over single-gene approaches.

The results of this work demonstrate the promise of using the phenocentric Bayesian

networks to study the genetic architecture of complex traits and demonstrates the utility of multigenic predictive methods compared to single gene approaches.

Part I

Probabilistic Graphical Models in Complex Trait Genetics

Chapter 2

Complex Traits

2.1 Definition

Complex traits are conditions with wide variability in physical manifestation and seemingly different mechanisms of onset and progression [1, 31]. Understanding them is an important problem whose solution would dramatically impact millions of people as most common disorders, which are the greatest health burden in the Western world, are complex traits. Although they are heritable, with observed familial aggregation, complex trait inheritance patterns do not follow Mendelian proportions. The complexity in heritability can be due to a strong and intricate influence of environmental factors on a simple genetic inheritance pattern, a network of complex genetic interactions that is mildly influenced by few environmental factors, or, a network of multiple and complex genetic and environmental interactions. Unfortunately, most complex traits seem to follow the last pattern. That is, they cannot be easily explained in terms of a single or few genes, and their expression is altered strongly by many environmental factors. Intensely studied complex traits include features of cardiovascular disease, Type 2 diabetes mellitus, asthma, obesity, autism, Parkinson's disease, and Alzheimer's disease.

2.2 Genetics

Uncovering the genetic architecture of complex traits is an important step towards understanding them. For many years, linkage studies in which patterns of allele segregation and disease occurrence in family pedigrees are compared, helped to elucidate the causes of many

simple Mendelian genetic traits. The traditional measure in such studies is the *LOD score*, given by [32]:

$$LOD = log\frac{L(r)}{L(0.5)},$$

where $L(r)$ is the likelihood of a disease and genetic marker occurrence as a function of genetic recombination fraction r. When $r = 0$ alleles are transmitted together, whereas when $r = 0.5$ alleles are transmitted independently. Larger LOD scores indicate that transmission of a genetic marker is associated with having a disease. Linkage analysis methods have been applied to study complex traits, but they are usually ill-suited to the task because of the complexity with which such traits are inherited. For over ten years, population-based association methods have been prefered [33] though they are recognized as a technique with limitations for the study complex traits [34]. Armed with data from the human genome [35, 36], the promise of HapMap International Project [37], and myriad genetic data gathered around the world, the expectation was that great strides would be made in solving complex traits. Indeed, genes related to many complex traits have been found, but given the large amount of human genetic information that is currently available and continues to expand at a rapid pace, the understanding of complex traits has been disappointingly slow [38, 39]. Nonetheless, association studies are the most common method used in genetic analysis and will continue to be widely used for years to come.

2.3 Association Studies

Genetic association studies attempt to find genes that are associated with a phenotype of interest. The key assumption in such studies is that a group of individuals who share a phenotype have a genotic commonality. There are two prevalent association study designs: case-control and familial. Although familial studies have the advantage of some built-in control because people with a more similar genotype with and without a disease are compared, case-control studies are often favored because it is easier to gather large cohorts from the general population than to find families with enough cooperative members. Additional advantages of case-control designs in the genetic dissection of complex traits are discussed in [40, 41]. The two most common traditional methods used in case-control studies are contingency table tests and logistic regression models.

Contingency Table Tests

Contingency tables are widely used for significance testing of categorical variables. Data are separated into rows and columns, and tested for independence or association. A generic case-control table used to find an association between a gene and occurence of a phenotype is the following:

	Major allele	Heterozygous allele	Minor allele	Total
Cases	M_1	H_1	m_1	$N_1 = M_1 + H_1 + m_1$
Controls	M_2	H_2	m_2	$N_2 = M_2 + H_2 + m_2$
Total	$n_1 = M_1 + M_2$	$n_2 = H_1 + H_2$	$n_3 = m_1 + m_2$	$N_T = N_1 + N_2$

Each entry in the table corresponds to the number of occurrences of a given genotype for cases or controls. The sums along columns and rows are referred to as column margins and row margins, respectively. The overall number of entries (N_T) is referred to as the grand total. When performing significance testing to find out whether the occurrences of genotypes are significantly different among cases and controls, we compare the actual, or observed, occurrences to expected values. The expected values are those corresponding to the null hypothesis that the distribution of each type of allele is equal for the cases and controls. That is, the probability for a control to have each genotype is equal to the probability for a case to have each genotype. The computation of expected values is performed by calculating, for each $i \times j$ cell of the table, the product of the i^{th} row margin by the j^{th} column margin, divided by the grand total. The table of expected values, corresponding to the table above is:

	Major allele	Heterozygous allele	Minor allele
Cases	$[N_1(M_1 + M_2]/N_T$	$[N_1(H_1 + H_2]/N_T$	$[N_1(m_1 + m_2]/N_T$
Controls	$[N_2(M_1 + M_2]/N_T$	$[N_2(H_1 + H_2]/N_T$	$[N_2(m_1 + m_2]/N_T$

Now that expected and observed values have been obtained, a statistical measure to test the null hypothesis that there is no difference among the genotypes of cases vs. controls is needed. A traditional measure to test this hypothesis is Pearsons χ^2:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.1}$$

where O_{ij} and E_{ij} refer to the $i \times j$ cells of the observed and expected tables, respectively. The sum, X^2, follows a chi-square distribution with an appropriate number of degrees-of-

freedom (df) for our table if the null hypothesis of equality in the distribution of genotypes among cases and controls is true. In most situations, the number of df for an $R \times C$ contingency table is/are $(R-1) \times (C-1)$. However, because Hardy-Weinberg equilibrium is assumed in the genotype example, that table has one df. Choosing an appropriate threshold of significance (often $\alpha = 0.05$) allows for the acceptance or rejection of the null hypothesis:

- If $X^2 > \chi^2(df, 1-\alpha)$, then we reject the null hypothesis. In the genotype example, the genotypes would be distributed differently among cases and controls.
- If $X^2 < \chi^2(df, 1-\alpha)$, then we accept the null hypothesis. In the genotype example, there would be no difference in the genotype distribution of cases versus controls.

Knowing the df and α allows for the calculation of a p-value corresponding to the X^2 obtained. The p-value is the area under the chi-square distribution, of the appropriate df, integrated from X^2 to infinity. A p-value is often reported as a measure of how strong the case versus control genotype distributions are similar or different.

For the chi-square test to be reasonable, certain conditions should hold. Most generally, no more than 1/5 of the cells should have expected values < 5 and no cell should have an expected value < 1. Allelic tables have previously been used to perform contingency tests, but this practice has been discouraged following results showing that they are not as robust as genotype tables in some situations (e.g. when alleles are codominant) [41].

A common alternative to analyze a contingency table is the calculation of Fisher's exact test. This test does not use approximations involved in calculating X^2 and so is often prefered. Though it is computationally more demanding, it is implemented in most statistical and mathematical software packages [42].

Contingency table measures often include calculations of the odds ratio. There are many subsets of genetic variables that are analyzed when looking for associations. Odds ratios provide estimates of how much more likely cases versus controls are to have a specific allele or genotype. Some of the most often used comparisons and their corresponding odds ratios (OR) are:

1. Allele frequency ratio in cases versus controls:

$$OR = \frac{(2M_1 + H_1)(2m_2 + H_2)}{(2M_2 + H_2)(2m_1 + H_1)}$$

2. Ratio of heterozygous genotype to major genotype in cases versus controls:

$$OR = \frac{H_1 M_2}{M_1 H_2}$$

3. Ratio of minor to major genotype in cases versus controls:

$$OR = \frac{m_1 M_2}{M_1 m_2}$$

4. Allele positivity. Ratio of minor plus heterozygous genotypes to major genotype in cases versus controls:

$$OR = \frac{(m_1 + H_1) M_2}{M_1 (m_2 + H_2)}$$

5. Armitages trend test [43]:

$$OR = \frac{\frac{H_1 M_2}{N_{12}} + \frac{m_1 H_2}{N_{23}} + \frac{4 m_1 M_2}{N_{13}}}{\frac{M_1 H_2}{N_{12}} + \frac{H_1 m_2}{N_{23}} + \frac{4(m_1 M_2 M_1 m_2)^{0.5}}{N_{13}}},$$

where $N_{ij} = n_i + n_j$ (i.e. the sum of column margins of the observed genotype table).

For (1) through (5) above, chi-square tests can be performed by making tables with the relevant subset of alleles or genotypes. Armitage's trend test, also known as the Cochran-Armitage test, is one of the most common tests used in genetic association studies [42]. It is a conservative test that does not depend on the assumption of Hardy-Weinberg equilibrium of a gene. The test consists in finding a trend among genotypes and presence of a phenotype by assuming that the alleles confer an additive effect. Therefore, the cases of a disease would be expected to have a homozygous minor genotype more often than a heterozygous genotype because having two copies of the minor allele would increase the likelihood of a disease. If there is no relationship among genotypes and the disease, then the proportion of cases and controls with each genotype is equal.

Contingency tests allows for the calculation of statistical significance of individual genes only. This is of utility when testing for the effects of singles genes, as in monogenic traits, but of less utility when trying to find a group of genes that affect a complex trait.

Logistic Regression Models

Binary logistic regression models are a type of regression that is used to determine the best predictors of a two-choice-outcome dependent variable. Given a set of categorical and/or continuous independent variables, a model with those that best describe the dependent variable is constructed. A logistic regression model can be used to predict the dependent variable given a set of independent variables. Additionally, the model can be used to measure the percent of variance in the dependent variable explained by the independents, and to rank the independent variables by their contribution towards explaining the dependent variable, including interaction effects. Interaction terms are usually no larger than the product of two independent variables as the computation time rises significantly and the model rarely improves with higher order terms.

Logistic regression models are constructed by transforming the dependent variable into a logit variable, which is the natural log of the odds of the dependent occurring. A maximum likelihood estimation (MLE) procedure is then applied to find significant independent variables. The logit transform of a binary dependent variable, Y_i, with independent variables $X_{i,1},, X_{i,p-1}$ is given by:

$$logit(\pi_i) = \beta_o + \sum_j \beta_j X_{ij},$$

where $E[Y_i] = \pi_i$. The MLE is analogous to the least squares estimation used to find the coefficients in a linear regression model. But, instead of looking at changes in the dependent variable given changes in the independent variables, the MLE looks at changes in the log of the odds of the dependent variable. The β coefficients are the weight that is given to each independent variable.

Logistic regression requires that observations are independent and that the linear relationship to the logit function holds. Because the model can be used to make predictions of the dependent variable, model accuracy can be checked by making a prediction with the independent variables used to construct the model. Predictions are usually made with log likelihood tests of the data. The goodness of fit of the model can also be assessed with Chi-square tests.

In a stepwise logistic regression, a logistic regression model is built by adding or removing independent variables according to a statistical significance test at each step. Stepwise regression procedures are useful when the number of independent variables is large. They

can be run in the forward, backward, or both selection directions. Forward selection procedures begin with a constant-only model to explain the dependent variable, and add the most significant independent variables one at a time according to a statistical criterion. The backward selection procedure begins with all of the independent variables and removes one at a time, according to a criterion by which the variable removed is deemed worst. The both selection refers to a combination of forward and backward procedures.

The most common criterion used to add/remove variables in the stepwise regression models is the likelihood ratio test, but other measures such as Raos efficient score statistic, the Wald statistic, and the conditional statistic have been developed. Most of these methods are based on the likelihood ratio test. The stepwise regression is ended when some criterion is met. In most cases, the criterion used is either (1) last step, (2) Akaike Information Criterion (AIC) [44], or (3) Bayesian Information Criterion (BIC) [45]. The last step criterion consists of updating the model until adding another variable would not significantly improve the model according to the likelihood ratio test. The AIC criterion penalizes the likelihood by the number of variables added to the model to attempt to reduce overfitting. A logistic regression is stopped when the lowest AIC is found. The AIC is given by:

$$AIC(M) = -2log[L(M)] + 2p(M),$$

where $L(M)$ is the likelihood estimate for model M, and $p(M)$ is the number of predictors used in model M (i.e. the number of degrees of freedom). The BIC criterion, also known as the Schwarz criterion, penalizes the likelihood by the number of variables added to the model taking into account the sample size of data used to construct the model. Models built with large datasets are at a higher risk of being overfitted than smaller datasets. The BIC is given by:

$$BIC(M) = -2log[L(M)] + p(M)log(n),$$

where $L(M)$ and $p(M)$ are the same as for the AIC and n is the sample size. Methods using AIC and BIC criteria are considered penalized maximum likelihood methods because they penalize the likelihood of the data by the complexity of the model used to describe it. Because the BIC takes into account the number of observations in addition to the number of model parameters, models found using the BIC tend to be more parsimonious than those obtained with the AIC criterion.

The stepwise procedure is usually recommended for exploratory purposes as it easily models noise in data. Many problems with stepwise regression models have been described, and a few will be mentioned here. There is a high likelihood of multicollinearity with larger numbers of independent variables, and stepwise procedures do poorly when faced with collinearity. A large number of subjects are required per independent variable to keep the number of fortuitous significant variables down. For instance, at a 0.05 significance level, one out of 20 independent variables is expected to be significant by chance alone. Peduzzi et al. estimated that there should be no more than one independent variable for each 10 occurrences of the smaller of the dependent variable outcomes [46]. All of the tests to evaluate the performance of the stepwise logistic regression model, which is attempting to find a best hypothesis, are based on tests that were designed to test prespecified hypotheses. As the number of independent variables rises, calculations can become intractable, especially when interaction terms are included. Therefore, logistic models of genetic data are of limited utility when data for hundreds or more genes are studied simultaneously.

2.4 Limitations of Association Studies in Complex Traits

Although successful gene associations in complex traits have been found and association study designs are more powerful to detect susceptibility variants in complex traits than linkage analysis studies [33], strong concern has surfaced over the lack of reproducibility of many association studies [47, 4, 48]. Some of the factors that are responsible for unreproducible results and other limitations of traditional association studies are discussed below [39, 49].

Ambiguous Phenotype Definition

Because complex traits are difficult to define precisely, or their definition allows for a heterogeneous group of disease processes to be classified as one, some genetic studies do not have homogeneous trait populations. Though most complex trait studies use objective phenotypes to define case versus control populations in an attempt to overcome problems related to trait definition, which minimizes within study population heterogeneity, different studies often use different criteria to define complex traits, making between-study comparisons difficult. When compared studies using different phenotype definitions appear to be

unreproducible, it cannot be determined whether the inconsistencies are due to two true associations to different phenotypes or inconsistent associations to a common phenotype.

Population Structure

When case and control populations differ by more than the phenotype of interest, then any gene association measured may be due to any of the differences among them. For instance, if the groups have different ethnic distributions, then genetic associations measured may be related to ethnic genetic heterogeneity and not the phenotype of interest. A more complicated population structure effect may be that similar alleles are expressed diferently among populations due to gene-gene and/or gene-environment interactions. In this case, similar alleles may only be found to be associated to a trait in some populations. Problems related to population structure have long been recognized and most studies attempt to reduce this error source with study designs. Methods to account for population structure in gathered data have been proposed [42]. Such methods include the use logistic regression models [50] and principal component analysis [51] with null SNPs to account for population structure.

Changes in Statistical Power

In traditional association studies, power is proportional to the number of subjects studied. Therefore, if follow-up studies have fewer subjects than initial studies, then initial associations found will not be re-measured. Additionally, intial studies may use population extremes to measure genetic differences (e.g. diseased subjects with severe symptoms/phenotype and control subjects with no symptoms or phenotype expression of any form), while follow-up studies may use more representatitive samples of the populations being studied (e.g. diseased and control subjects with variable phenotype expression). In such cases, the follow-up studies may also have less *power* to detect genetic association because the allelic distribution is more likely to be heterogeneous within the disease and control groups.

Chance

As the number of genetic variants studied gets larger, the probability of finding false-positive associations due to chance alone becomes higher. This is especially true in simple

association designs where p-value tests with little stringency are used to test for significance and with current whole-genome approaches where over 500,000 SNPs are measured at one time. Many studies with simple statistical designs fail to exclude chance as the cause of association, and publish measured associations with more confidence than they deserve.

Publication Bias

Because negative association results are rarely published, especially when they are not from follow-up studies, literature searches of association are biased towards finding false-positive results. This has a strong impact on candidate gene approaches and experimental designs. Additionally, good systematic reviews and meta-analysis studies are hidered by the unavailability of negative results. Examples demonstrating publication bias by showing the large association effects measured in small studies and the small association effects in large studies can be found in [52]. Errors due to publication bias should diminish in the near future with the advent of genome-wide association studies and the public availability of large genetic datasets.

Rare Alleles

Allelic heterogeneity occurs when a variety of genetic variants can independently cause a trait. In this situation, true but different allelic associations could explain a trait in different populations studied. Under the "common-allele, common-variant" hypothesis, it is believed that the allelic spectrum that causes common traits is small [53, 47]. This view does not suggest that allelic heterogeneity plays a significant role in explaining inconsistenices among association study results. Alternative views that consider rare alleles to be significant in explaining complex traits [54] would account for some association study inconsistencies. The common-allele, common variant view has been prevalent over the past decade, but the importance of rare alleles has become more accepted in light of recent studies where rare alleles play a significant role in determining complex traits [55].

Single Gene Additive Approach

An important limitation of traditional association studies in the study of complex traits is that they investigate the relationship of single genes or assume additive effects of single genes to a phenotype. Because complex traits are caused by multiple genes, methods that

take into account the complex interaction of many genes are more useful than one-gene-at-a-time approaches [56, 57]. Though in principle logistic regression models can account for multivariate interactions, in practice they are inadequate to do this with large genetic datasets.

Studies do not Predict Individual Outcomes

The results of traditional association studies are lists of genes that are significant based on a statistical threshold. Validation studies involve replicating initial findings, which does little more than to increase the sample size of the original result. Results of gene association studies have not often been used in a clinical setting or to motivate further biological studies. Lacking from these studies is a result that can be applied to an individual such as a quantitative risk for a trait that would be useful in a clinical setting. Such quantitative measures for an outcome are readily interpretable and decisions about how to change an individual's life can be based on them. Though logistic regression models can assign a probability to predict an outcome, few genetic studies use the models for this purpose.

Chapter 3

Probabilistic Graphical Models

3.1 Introduction

Probabilistic graphical models (PGMs) use principles from graph theory and probability theory to model complex systems with multiple interacting entities. They were developed to address high-dimensional problems that were intractable with existing methodologies. The approach consists in reducing large problems into smaller, more maneagable ones using conditional independence assumptions [58, 59]. Computational algorithms can be created to address the smaller problems, whose solutions can be joined to build a comprehensive solution. In a probabilistic graphical model, the attributes of a problem are treated as random variables and the relationships among them are described by probability distributions. The model is wholly characterized by a joint probability distribution and its corresponding graphical representation. The underlying probabilistic foundation allows PGMs to find and use complex multivariate dependencies to understand a problem, while the graphical representation is helpful to intuitively interpret the relationships among variables.

Originally, graphical models were constructed by having experts choose graphs of conditional variable dependency relations and use subjective assessments of the probability distributions that quantified the dependencies. Although this approach is sometimes used today, there are few cases where it can be applied successfully and incontrovertibly. Instead, learning methods have been developed that extract the conditional dependencies and graph models from user-supplied data. The existence of efficient algorithms to learn models from data makes PGMs a powerful tool to discover complex data dependencies.

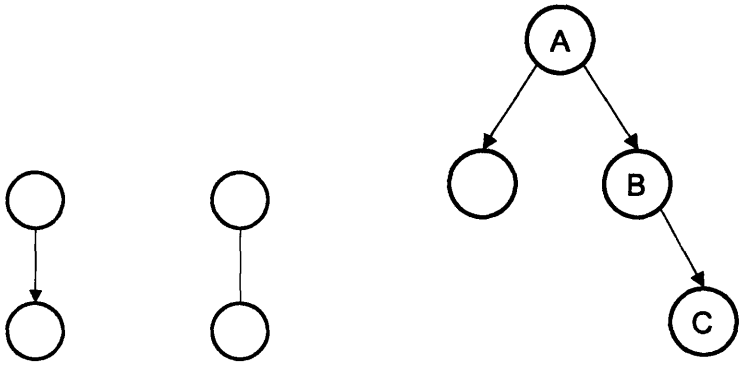
The process of learning consists of two main parts: parameter estimation and model

selection [60]. Parameter estimation refers to the process of calculating the conditional probabilities of a given model structure. This is often accomplished using a maximum likelihood approach. Model selection refers to choosing a structure that best captures the dependencies among the variables and is often performed by optimizing a score such as a Bayes' factor comparing the marginal likelihood of two models.

Once a model has been selected, a common desired task is to use the model for prediction (i.e. inference). This task serves to validate the model's ability to capture the dependencies among the data used to create it, and to test its generalizability with independent datasets. There are many ways to carry out predictions [59]. Exact calculations can be performed but are usually too computationally intense for practical purposes. Most algorithms rely on approximations and exploit a model's graphical structure to increase computational efficiency.

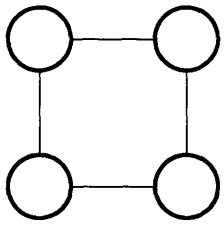
After defining some basic notions needed to describe graphs, some of the most common probabilistic graphical models are introduced. The first two, Bayesian networks and Markov random fields are stationary models, while hidden Markov models are temporal models.

In PGMs, nodes represent random variables and edges represent the probabilistic dependencies among nodes [61]. Edges between nodes can be directed or undirected [Figure 3-1(a)]. Directed edges are represented as arrows and are called arcs. Trails are sequences of edges that connect nodes in a graph. Paths are trails in which edges are followed only along directions in which arrows point. A trail of undirected edges is an undirected path, while a trail of arcs followed from arrow tails to heads is a directed path. If a path leads from node A to node C, then node A is said to be an ancestor of C, and C a descendant of A [Figure 3-1(b)]. If there is one edge between these two nodes, then node A is said to be the parent of B, and B is a child of A. When a trail begins and ends on the same node, such a trail is a cycle [Figure 3-1(c)]. Connected graphs are those which have trails between any two nodes [Figures 3-1(b) and 3-1(c)]. A tree is a connected graph with no cycles [Figures 3-1(b)]. A graph composed of arcs only is a directed graph. A directed acyclic graph (DAG) is a directed graph that contains no cycles [Figures 3-1(b)]. A clique $\mathcal{C}$ is a maximal subset of a graph's nodes in which every node is directly connected to every other node in $\mathcal{C}$.



(a) Edges representing probabilistic dependency between two nodes can be directed (left) or undirected (right)

(b) A path connects ancestor node A to descendant nodes B and C. Node A is a parent of child node B.



(c) A cycle contains at least one path connecting nodes back to themselves.

Figure 3-1: Basic concepts used in probabilistic graphical models.

3.2 Bayesian Networks

The most common PGMs are Bayesian networks [62]. Bayesian networks are DAGs that represent variables $\mathbf{Y} = \{Y_1, ..., Y_N\}$ [Figure 3.2]. Each variable has a conditional probability $P(Y_i|\Pi_i)$, where Π_i are the parents of Y_i. The joint probability distribution of the network is given by the product of each variable's conditional probability:

$$P(Y_1, ..., Y_N) = \prod_i P(Y_i|\Pi_i), \tag{3.1}$$

Bayesian networks are most often used when asymmetric probabilistic relationships exist between nodes. In some frameworks, such as artificial intelligence or medical decision-making, arcs represent causal relationships. However, arc relationships need not be causal. The development of efficient learning methods have made Bayesian networks one of the

most promising tools in data mining. Details regarding Bayesian networks are given in Section 4.

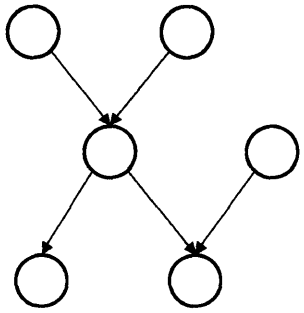


Figure 3-2: Bayesian networks are DAGs, were nodes represent variables and arcs represent conditional dependencies among variables.

3.3 Markov Random Fields

Markov random fields (MRFs) are undirected graphs where the relationships between variables $\mathbf{Y} = \{Y_1, ..., Y_N\}$ are given by potentials that capture the interactions among small subsets of variables [Figure 3.3]. The joint probability distribution is given by:

$$P(Y_1, ..., Y_N) = \frac{1}{Z} \prod_i \phi_i[C_i],$$

where $\phi_i[C_i]$ is the i^{th} potential over variable subset C_i, and Z is a normalization constant given by:

$$Z = \sum_{C \in \mathbf{Y}} \prod_i \phi_i[C_i]$$

Originally developed to model lattices of particles, MRFs are most often used to address problems in which variables are correlated and there is no clear directionality to their relationship.

3.4 Hidden Markov Models

Hidden markov models (HMMs) are representations of stochastic processes in which future states are assumed to be conditionally independent of past states given the present state.

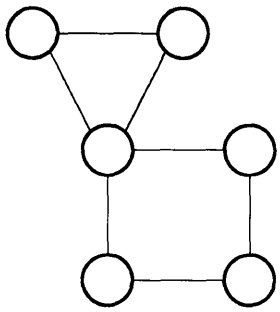


Figure 3-3: Markov random fields are undirected graphs, were nodes represent variables and edges represent conditional dependencies among variables.

Underlying an observed variable is a hidden one with unknown parameters. The purpose of the HMM is to uncover the unknown parameters using variations of the observed variable. Graphically, HMMs are represented as shown in Figure 3.4. Each top node in this figure represents the state of a hidden random variable $\mathbf{X}$ at a given time. The bottom nodes represent the observed states of random variable $\mathbf{Y}$ at a given time. The joint probability distribution for observing a sequence of L instances of $\mathbf{Y}$, $Y = y_0, ..., y_{L-1}$ is given by:

$$P(Y) = \sum_X P(Y|X)P(X),$$

where X represents every possible combination of sequences $x_0, ..., x_{L-1}$. This sum is efficiently calculated using the forward-backward procedure, which allows for predictions of an observed outcome [63]. Efficient algorithms have also been created to address the question of (1) finding the most likely sequence of $\mathbf{X}$, given model parameters, that could have generated a sequence of $\mathbf{Y}$ (Viterby algorithm), and (2) given a sequence of observed values of $\mathbf{Y}$, learn the model parameters (Baum-Welch algorithm) [64]. Problems that can be represented as an HMM can be readily solved because of the wealth of algorithms that exist to understand these models.

3.5 Genetics Applications

Probabilistic graphical models are a powerful technique that can overcome limitations of conventional association study approaches. Going beyond single or pairwise gene interactions with a phenotype, probabilistic graphical models are able to account for complex gene

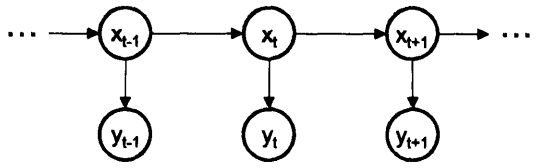


Figure 3-4: Hidden Markov Models are directed graphical models representing a time process. The observed random variable **Y** depends on a hidden random variable **X** at each time point.

interactions. Addtionally, they can be used to make predictions of a phenotype of interest for individual subjects that allows for ascertainment of their validity.

Sequence Analysis

HMMs have been used for a variety of DNA and protein sequence analysis studies [64, 65, 66]. One of the most common applications is sequence alignment for the identification of similar genes, protein-coding regions and transcription binding sites within and between species, and to find families of related sequences. Such applications have helped speed up the analyses of newly genotyped species by allowing the identification of genomic regions based on prior knowledge of previously genotyped species. They have also helped to identify the function of unknown genes within one species by identifying similarities between newly genotyped regions and known genes.

Linkage Analysis

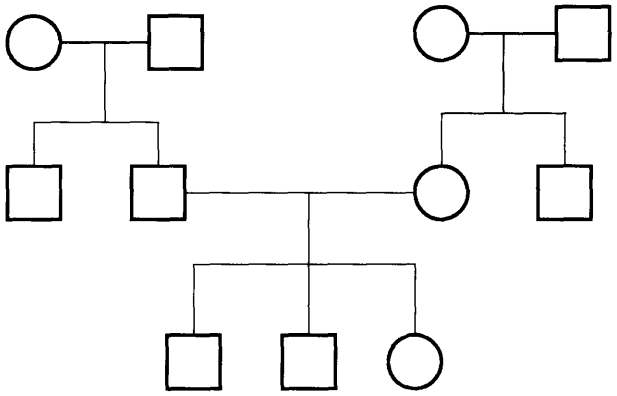
A variety of algorithms have been developed for genetic pedigree linkage analysis using probabilistic graphical models. Family pedigrees can be represented naturally as Bayesian networks with nodes representing individuals and arcs connecting parents to offspring. Figure 3-5(a) shows a traditional pedigree used in inheritance analysis. A Bayesian network representation for this pedigree is shown in Figure 3-5(b). This representation holds for genotypes undergoing Mendelian inheritance. For a system with a alleles, the genotype for each individual, G_i, has one of $a(a+1)/2$ possible states. The joint probability distribution for a network representing N individual's genotypes is given by:

$$P(g_1, ..., g_N) = \prod_{i \in \mathcal{F}} \pi(g_i) \prod_{j \notin \mathcal{F}} \tau(g_j | g_{m_j}, g_{f_j}),$$

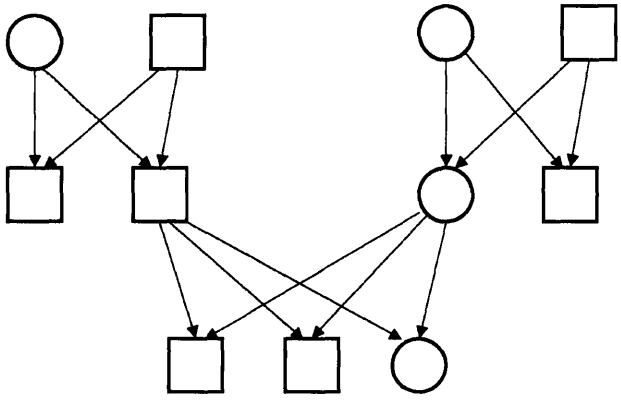
where π represent the founder genotypes (i.e. the genotypes of those without parents), $\mathcal{F}$ is the set of all founders, and τ are the transmission probabilities with which a genotype is transmitted to an individual whose mother has genotype g_{m_i} and father has genotype g_{f_i}:

$$\tau(g_i|g_{m_i}, g_{f_i}) = P(G_i = g_i|G_{m_i} = g_{m_i}, G_{f_i} = g_{f_i}).$$

More general PGM pedigree representations and their use in simple linkage analysis, detection of a quantitative trait locus (QTL), and other applications can be found in [11].



(a) A family pedigree used in traditional linkage analysis.



(b) Bayesian network representation where nodes represent individual genotypes, which depend on parental genotypes.

Figure 3-5: Bayesian network representation of a family pedigree for linkage analysis.

Multilocus Association

Most current case-control association methods that consider more than single SNPs, do so with the intention of identifying single genetic loci that cause a disease. Multiple SNPs are considered to take advanatage of linkage disequilibrium among nearby genetic markers in the identification of causal loci. Because many SNPs are likely to be in linkage disequilibrium with a causative genomic region, considering multiple markers is likely to increase the signal around true causative site [31, 67].

One approach to multilocus association is to relate haplotypes, rather than individual SNPs, to a trait or disease [68]. Haplotypes are patterns of genetic variation that tend to occur simultaneously along a chromosome [69]. The human genome is organized into blocks such that the state of some SNPs usually allows inference about the state of other SNPs in certain regions. Because complete resequencing of many individuals is not currently feasable, methods to infer haplotypes from limited genetic markers are necessary. Haplotype maps for representative populations provide a mechanism by which SNPs that uniquely identify haplotypes (htSNPs) can be selected for association studies [70]. The htSNPs can be used in association studies using conventional one-SNP-at-a-time approaches with the assumption that a greater genetic variation of the population is being accounted for than with other SNP selection procedures. Alternatively, haplotypes can be infered computationally and then associated directly to a trait using conventional statistical measures [71].

HMM models have been used to model dependencies between haplotype blocks [72] and to model haplotype ancestry along chromosomes [73, 74] to perform linkage disequilibrium mapping of genetic traits. More complex PGM for multilocus association, based on variable-length Markov chains, have been developed [75].

An alternative to using haplotype maps to model linkage disequilibrium is to consider multiple SNP interaction terms with PGMs. These methods provide an alternative to haplotype maps that can account for dependencies among SNPs that do not assume a simple physical proximity relationship to model linkage disequilibrium and consider more than pairwise associations. MRFs describing dependencies of multiple genetic markers and a trait have been used for such multilocus association studies [76]. The metric used in these studies for graphical model selection is the Bayesian information criterion:

$$BIC = log[L(G)] - \frac{log(h)}{2} df(G),$$

where $L(G)$ is the likelihood of graphical model G, h is the number of haplotypes observed, and $df(G)$ is the number of degrees of freedom of model G. The model with highest BIC is kept. An extension of this work cana perform haplotype reconstruction in addition to finding associations among genetic markers and a phenotype [77]. An alternative PGM approach has been developed for genomewide scale linkage analysis that also uses decomposable MRFs, but uses Bayesian model averaging (i.e. a Markov Chain-Monte Carlo algorithm) to select SNPs associated with a disease instead of keeping a single model with a maximal score [8].

Although the approaches discussed above consider multiple SNPs, limits are usually placed on which SNPs can interact on the basis of physical distance because the interaction under consideration is that due to linkage disequilibrium. A limitation of the above algorithms is that they require phased haplotype data. Though robust and efficient algorithms to infer haplotypes are available [78, 79, 80], the results of these procedures still have uncertainties that are often not considered when measuring association. Further, most haplotype identification algorithms do not scale to genomewide studies. Consideration of epistatic gene-gene interactions have not been implemented in these approaches, though some authors mention the ability of their methods to do so [72, 8].

Part II

A Novel Bayesian Network Approach for Complex Traits

Chapter 4

Bayesian Networks

4.1 Fundamentals

Most Bayesian network algorithms have been developed for categorical variables which are considered here. Let c_i be the number of states of Y_i and y_{ik} be a state of Y_i. The conditional dependency linking Y_i to its parents, Π_i, is mathematically defined by the conditional probability distributions of Y_i given each of the posible configurations of its parents $\pi_{i1}, ..., \pi_{iq_i}$. A node Y_i is conditionally independent of one of its non-descendants, $ND(Y_i)$, given parents Π_i that both have in common. Such conditional indepence relations allow for the factorization of the joint probability of a set of values of $\mathbf{Y}$, $\mathbf{y_k} = \{y_{1k}, ..., y_{Nk}\}$:

$$p(\mathbf{y_k}) = \prod_{i=1}^{N} p(y_{ik}|\pi_{ij}), \tag{4.1}$$

where π_{ij} are the configuration of states of Π_i in $\mathbf{y_k}$. Note that the index j of π is actually a function of i and k because the parent configuration in a set of values $\mathbf{y_k}$ is determined by the index i, which specifies the child variable and hence the parents it can have, and the index k, which specifies the states of the parent variables.

4.2 Parameter Estimation

Assume a DAG M and a sample of n cases $\mathbf{y} = \{y_1, ..., y_n\}$ are given. The sample $\mathbf{y}$ is an $n \times N$ matrix because each case, $\mathbf{y}_k$, is a row vector with each entry corresponding to the state of one of the N variables $\mathbf{y}_k = (y_{1k}, ..., y_{Nk})$. The θ parameters are to be

estimated. That is, $\theta = (\theta_{ijk}) = (p(y_{ik}|\pi_{ij}), \theta)$ are to be found. The parameter vector $\theta_{ij} = (\theta_{ij1}, ..., \theta ijc_i)$ is associated to the conditional distribution of $Y_i|\pi_{ij}$ inferred from $\mathbf{y}$.

The standard way to estimate θ is to use conjugate analysis. Let $n(y_{ik}|\pi_{ij})$ be the frequency of $(y_{ik}|\pi_{ij})$ pairs in the sample, and let $n(\pi_{ij}) = \sum_k n(y_{ik}|\pi_{ij})$ be the frequency of π_{ij}. The joint probability shown in Equation 4.1 can be written in terms of the unknown θ_{ijk}:

$$p(\mathbf{y}_k|\theta) = \prod_{i=1}^{N} \theta_{ijk},$$

where j is determined by i and k. If the cases $\mathbf{y}_k$ are independent, then the likelihood function is the product of the joint probabilities:

$$L(\theta) = \prod_{k=1}^{n} p(\mathbf{y}_k|\theta) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}.$$

The innermost products local likelihood contributions from individual parent configurations:

$$\prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}, \tag{4.2}$$

while the middle product is that of local parents-child configurations:

$$\prod_{j=1}^{q_i} \prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})}.$$

An assumption that usually holds when the likelihood can be factorized into parents-child contributions is *global independence*: the parameter vectors θ_{ij} and $\theta_{i'j'}$ associated to variables Y_i and $Y_{i'}$ are independent for $i \neq i'$. If θ_{ij} and $\theta_{ij'}$, which are associated to the distributions of Y_i, are independent given different parent contributions π_{ij} and $\pi_{ij'}$, where $j \neq j'$, then *local independence* holds and the joint prior density of can be factorized into:

$$p(\theta|I_o) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} p(\theta_{ij}|I_o).$$

When global and local independence hold and there are no missing values in the sample $\mathbf{y}$, the posterior density of θ is proportional to the product of the above factorizations of the

prior density and likelihood functions:

$$p(\theta|I_1) \propto \prod_{ij} \left\{ p(\theta_{ij}|I_o) \prod_{k=1}^{c_i} \theta_{ijk}^{n(y_{ik}|\pi_{ij})} \right\}.$$

The posterior density factorization allows for the independent update of θ_{ij}, for all i, j, which reduces the update process to local procedures. If the prior distribution of θ_{ij}, for all i, j, is a *Dirichlet* distribution, D, with hyperparameters $\{\alpha_{ij1}, ..., \alpha_{ijc_i}\}, \alpha_{ijk} > 0$ for all i, j, k, then the prior density of θ_{ij} is given by:

$$p(\theta_{ij}|I_o) \propto \prod_{k} \theta_{ijk}^{\alpha_{ijk}-1}$$

up to a constant factor. This prior is conjugate to the local likelihood (Equation 4.2), as indicated by the similar functional forms. The prior hyperparameters, α_{ijk} encode the observer's prior belief and can be thought of as representing the frequencies of imaginary cases needed to formulate the prior. The frequency of such imaginary cases in the parent configuration π_{ij} is:

$$\sum_{k=1}^{c_i} (\alpha_{ijk} - 1) = \alpha_{ij} - c_i,$$

where α_{ij} becomes the *local precision.* Further, $\sum_j \alpha_{ij} = \alpha_i$, and α_i is the *global precision* on θ_i. For consistency, it is assumed that imaginary samples have equal numbers of observations for all variables Y_i such that $\alpha_i = \alpha$. This assumption is actually necessary to enforce local and global parameter independence [81]. The marginal probabilities of $(y_{ik}|\pi_{ij})$ can be specified by the α_{ijk}:

$$E[\theta_{ijk}|I_o] = \frac{\alpha_{ijk}}{\alpha_{ij}} = p(y_{ik}|\pi_{ij}),$$

and

$$Var[\theta_{ijk}|I_o] = \frac{E[\theta_{ijk}](1 - E[\theta_{ijk}])}{\alpha_{ij} + 1}.$$

Note that for fixed $E[\theta_{ijk}]$, the variance of θ_{ijk} becomes large with small values of α_{ij}.Therefore, small α_{ij} denotes great uncertainty of the parameters. Initial ignorance can be represented by assuming $\alpha_{ijk} = \alpha/(c_i q_i)$ for all i, j, k, which reduces $p(y_{ik}|\pi_{ij})$ to $1/c_i$.

Dirichlet distributions are closed under marginalization, which means that if initially the parameters follow a Dirichlet distribution, $\theta_{ij}|I_o \sim D(\alpha_{ij1}, ..., \alpha_{ijc_i})$, then any subset of

parameters, $(\theta_{ij1}, ..., \theta_{ijs}, 1-\sum_{k=1}^{s} \theta_{ijk})$, will have a Dirichlet distribution $D(\theta_{ij1}, ..., \theta_{ijs}, 1-\sum_{k=1}^{s} \theta_{ijk})$. The parameter θ_{ijk} will follow a Beta distribution with hyperparameters α_{ijk} and $\alpha ij - \alpha_{ijk}$:

$$p(\theta_{ijk}|I_o) \propto \theta_{ijk}^{\alpha_{ijk}}(1-\theta_{ijk})^{\alpha ij - \alpha_{ijk}},$$

making marginal inference easy to carry out. When the parameter independence and prior Dirichlet distribution assumptions hold, then the posterior density of θ will remain a product of Dirichlet densities and $\theta_{ij}|I_1 \sim D(\alpha_{ij1} + n(y_{i1}|\pi_{ij}), ..., \alpha_{ijc_i} + n(y_{ic_i}|\pi_{ij}))$. The updating procedure has increased the hyperparameters by the frequency of cases observed in the sample, $n(y_{ic_i}|\pi_{ij})$, which allows for a simple calculation at each updating step. The posterior expectation and variance become:

$$E[\theta_{ijk}|I_1] = \frac{\alpha_{ijk} + n(y_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})},$$

$$Var[\theta_{ijk}|I_1] = \frac{E[\theta_{ijk}|I_1](1 - E[\theta_{ijk}|I_1])}{\alpha_{ij} + n(\pi_{ij}) + 1}.$$

The local precision has increased from α_{ij} to $\alpha_{ij} + \pi_{ij}$, demonstrating that as the frequency of parents observed increases, our parameter uncertainty decreases.

4.3 Model Selection

Given a set of models $\mathbf{M} = \{M_o, ..., Mm\}$ that are believed to contain a true model of dependence among a set of variables $\mathbf{Y}$, the best model must be identified. Initially, each model is assigned a prior probability $p(M_j|I_o)$. Let $\theta^{(j)}$ be a vector of the conditional dependencies specified by M_j. The familiar Bayesian procedure is used to compute posterior probabilities from the priors and marginal likelihood functions:

$$p(M_j|I_1) \propto p(M_j|I_o)p(\mathbf{y}|M_j)$$

To find the most probable model, the marginal likelihood, $p(\mathbf{y}|M_j)$, must be computed:

$$p(\mathbf{y}|M_j) \int p(\theta^{(j)}|M_j)p(\mathbf{y}|\theta^{(j)})d\theta^{(j)} \tag{4.3}$$

where $p(\theta^{(j)}|M_j)$ is the prior density of $\theta^{(j)}$ and $p(\mathbf{y}|\theta^{(j)})$ is the likelihood function assuming M_j is the model of dependence. Equation 4.3 has a closed form solution when assumptions analogous to those in the previous section hold:

1. All sample cases are known
2. The cases are independent given $\theta^{(j)}$
3. The prior distribution of parameters is conjugate to the sampling model $p(\mathbf{y}|\theta^{(j)})$. Specifically, $\theta_{ij}^{(j)} \sim D(\alpha_{ij1}, ..., \alpha ijc_i)$ and global and local independence of the parameters holds.

The marginal likelihood of M_j becomes:

$$p(\mathbf{y}|M_j) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(y_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})},$$

where $\Gamma(\cdot)$ is the Gamma function [82]. Thus, the marginal likelihood can be computed with the hyperparameters of the Gamma function, $\alpha_{ijk} + n(y_{ik}|\pi_{ij})$, and the local precision values, $\alpha_{ij} + n(\pi_{ij})$, of the posterior distributions of $\theta_{\mathbf{ij}}$.

When the number of variables is large, exhaustive searches through all models becomes computationally impossible. This has led to the development of heuristic methods to shorten the search process. Often, the search time is shortened by imposing some ordering among variables in the form of $Y_i \prec Y_j$, meaning that Y_i cannot be the parent of Y_j, and the parents-child dependence is used to find local answers that are then pieced into a global solution. Both of these are used in the K2 algorithm, which is a common method used in the building of Bayesian networks [83]. In this algorithm, the local contribution of a node Y_i and its parents Π_i to the overall joint probability $p(\mathbf{y}|M_j)$ is calculated using:

$$g(Y_i, \Pi_i) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}. \tag{4.4}$$

The algorithm proceeds by adding one parent at a time to node Y_i and computing $g(Y_i, \Pi_i)$. The set of parents, Π_i, is expanded to include the parent node that maximally contributes to $g(Y_i, \Pi_i)$, until the probability ceases to increase at which point the algorithm considers remaining nodes.

4.4 Prediction

Bayesian networks are usually created for the purpose of calculating the probability of an outcome or assessing some interesting feature of the problem the model addresses. How to use a selected model to make a prediction is a difficult problem. In theory, very large tables of probabilities could be created that consider all nodes in a network simultaneously. In practice, most predictive algorithms exploit the structure of networks to perform local computations, which are then joined to give an overall solution. This approach drastically reduces computational time, making Bayesian networks useful in practice.

The most common Bayesian network inference algorithm is the *clique-tree propagation algorithm* [58]. Clique-tree propagation involves two processes: compilation and propagation. The compilation process involves grouping variables into *cliques*, organizing these into *junction trees*, and assigning numerical data to their appropriate locations in the junction tree. The propagation stage involves the performance of local computations and their dissemination along the junction tree to obtain a desired global solution. A general description of this algorithm follows.

Compilation

Moralisation

The compilation process begins by turning a Bayesian network into an undirected graph by a process called *moralisation*. During this process, every arc is converted into an undirected edge, and new edges are added between every pair of parents of a node. In this new graph, the set composed of each node and its parents, $Y_i \cup \Pi_i$, has an edge between every pair of elements and is said to be *complete*. The joint probability distribution 3.1 becomes

$$P(Y_1, ..., Y_N) = \prod_i P(Y_i|\Pi_i) = \prod_{C \in \mathcal{C}} \phi_C(Y_C), \tag{4.5}$$

where $\mathcal{C}$ is the set of all cliques C, and ϕ are the potentials of each clique. The potentials are obtained by multiplying the conditional probabilities of the $Y_i \cup \Pi_i$ within each clique, which can be done easily because the graph is moral. A simple example of a moralised graph is shown in Figure 4.4.

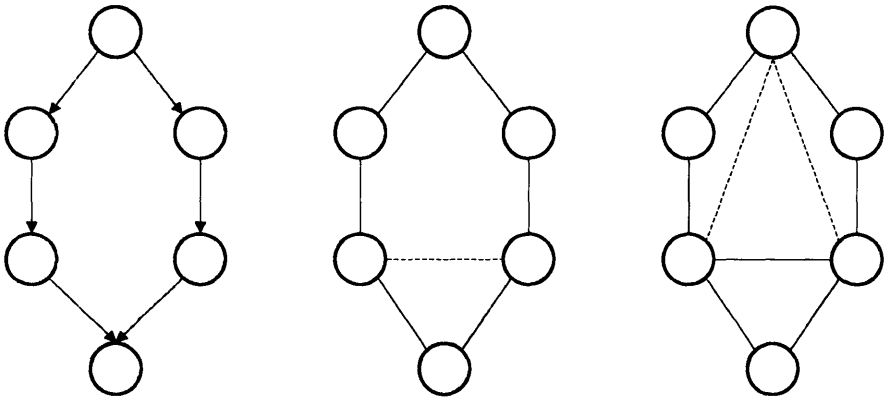


Figure 4-1: Beginning of compilation process of a Bayesian network (left). The network is *moralised* by converting directed edges to undirected edges and joining parents of each node (middle). The moral graph is triangulated by adding edges in all cycles with more than three nodes such that nonadjacent nodes are connected (right).

Triangulation

After a Bayesian network has been moralised, a process called *triangulation* is performed. Triangulation entails the addition of edges to the graph until all cycles with more than three nodes have *chords*, edges that connect two nonadjacent nodes in the cycle, by a process called *elimination*:

1. A copy the moral graph is created and an ordering of nodes is chosen
2. For each ordered node:
   (a) Fill-in edges are added between all pairs of the node's neighbors
   (b) The node and all of its adjacent edges are removed
   (c) Remaining fill-in edges are added to the original moral graph
3. The moral graph becomes a triangulated graph

The joint probability distribution remains that given in Equation 4.5, but the set of cliques C is now that in the triangulated graph. Finding an optimal triangulation, one that produces the smallest possible cliques, is an NP-complete problem [84]. However, there are various criteria used to order the nodes such that the triangulation process produces optimal results efficiently in most cases [85, 86]. Because the running time of inference algorithms is

exponential in the size of the largest clique, optimizing the triangulation step is crucial to computational efficiency.

Junction Tree Formation

In the next step of the clique-tree propagation algorithm, the cliques of the triangulated graph are identified and connected as nodes into a junction tree. For the propagation process to work properly, this tree must satisfy the property that $A \cap B \subseteq D$ for all $A, B, D \in \mathcal{C}$, where D is between A and B if it lies in the unique path from A to B. That is, elements found in cliques A and B must be in each of the cliques along the path connecting A and B. Details of the junction tree construction process can be found in [61].

Once the junction tree has been obtained, the potentials for each clique, ϕ_C are calculated. As shown in Equation 4.5, this is done by assigning the original node's conditional probabilities, $P(Y_i|\Pi_i)$, to cliques that contain node Y_i and its parents Π_i. First, all cliques are initialized to have unit potential: $\phi_C(Y_C) \equiv 1$. Then, each node Y_i is assigned to one of the cliques containing it. If $\mathcal{S}(C)$ is the set of nodes assigned to clique C, then the clique potentials are updated to be given by:

$$\phi_C(Y_C) = \prod_{Y_i \in \mathcal{S}(C)} P(Y_i|\Pi_i),$$

and the overall joint probability distribution becomes:

$$P(Y_1, ..., Y_N) = \prod_{C \in \mathcal{C}} \prod_{Y_i \in \mathcal{S}(C)} P(Y_i|\Pi_i).$$

Before propagation, it is common to update the potentials with observed data. For each clique containing an observed node $Y_i = y_i^*$, the potential is changed to ϕ_C^* as follows:

$$\phi_C^*(Y_C) = \begin{cases} \phi_C(Y_C) & \text{if } Y_i = y_i^* \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

This step, referred to as *entering evidence*, updates the joint probability to that given observed nodes Y_E^*:

$$P(Y_1, ..., Y_N|Y_E^*) = \frac{\prod_{C \in \mathcal{C}} \phi_C^*(Y_C)}{Z(Y_E^*)}, \tag{4.7}$$

where $Z(Y_E^*)$ is a normalization constant equal to the probability of the observations, $P(Y_E^*)$, and is given by:

$$Z(Y_E^*) = \sum_Y \prod_{C \in \mathcal{C}} \phi_C^*(Y_C). \tag{4.8}$$

Propagation

Construction of the junction tree permits actual calculations of the probabilities of interest, beginning with the probability of the observations given by Equation 4.8. In practice, this algorithm is not solved with brute force because the number of terms in the sum grows exponentially with the number of network nodes. A common approach is to select one clique to be the *root* of the network, and then use a *peeling* algorithm to propagate messages to the root [58]. The propagation begins at *leaves*, or cliques other than the root with only one neighbor, which are peeled off as the calculation gets passed towards the root.

Let the initial clique potentials, ψ_C, of a junction tree be given by $\psi_C = \phi_C^*, C \in \mathcal{C}$. When a message passes from leaf L to its neighbor D, the potential ψ_L is *marginalized* to $S = L \cap D$:

$$\psi_L^{\downarrow S}(Y_S) = \sum_{X_{L \setminus D}} \psi_L(Y_S, X_{L \setminus D}).$$

Before leaf L is removed from the junction tree, its neighboring clique D *absorbs* the message from L as its potential changes to $\widetilde{\psi}_D$: $\widetilde{\psi}_D = \psi_D \psi_L^{\downarrow S}$.

After all leaves other than the root, R, have been peeled, the probability of the root given the observations are all that remain to solve Equations 4.7 and 4.8:

$$\begin{aligned} P(Y_R|Y_E^*) &= \frac{\psi_R(Y_R)}{Z(Y_E^*)}, \\ Z(Y_E^*) &= \sum_{Y_R} \psi_R(Y_R), \end{aligned}$$

where ψ_R is the modified potential after all messages have been sent along the junction tree. The conditional probabilities of individual nodes in the root clique can now be obtained by summing appropriately over a maneagable number of node configurations. Variations of this propagation scheme have been created for specific applications [87]. Similarly, various computational shortcuts can be incorporated in specific situations [11].

The clique-tree propagation algorithm, and all other exact inference algorithms, are efficient with sparse graphs, but can be very slow with large graphs as their running time is

exponential in the size of the largest clique of the triangulated moral graph. Several approximate inference algorithms have been developed and are appropriate for specific applications, but many of them use elements of the exact approach outlined above.

4.5 Applications to Complex Trait Genetics

Bayesian networks are regarded as an emerging paradigm for the analysis of complex traits because of their ability to model complex multivariate dependencies and make predictions [88, 7, 8]. In addition to being used in pedigree linkage analysis as described in Section 3.5, they have been used to study gene expression data [9] and protein-protein interactions [10]. Of most relevance to this thesis, Bayesian networks have been used to study complex trait genetics in association studies.

One candidate gene association study investigated the prediction of stroke in sickle cell anemia patients using Bayesian networks [12]. Sickle cell anemia (SCA) is a monogenic recessive disease, but it is phenotypically complex. The clinical course of individuals suffering the disease has a wide range of severity and timing of symptoms. Stroke is a severe complication that affects 6-8% of sickle cell anemia patients, often before the age of 20. In order to find whether any candidate genes could serve to identify SCA patients who are likely to suffer a stroke, a Bayesian network was constructed using 108 SNPs in 39 candidate genes from 1398 SCA subjects. The resulting network contained 31 SNPs from 12 genes that interact to modulate the risk of stroke. Of these SNPs, 25 corresponding to 11 genes directly modulated stroke risk. Validation of the model in an independent population of 114 individuals had an accuracy of 98.2%. A forward logistic regression model was made for comparison to the Bayesian network model. It found that 5 SNPs from 11 genes directly modulated the risk of stroke, with a predictive accuracy of 88% in the independent population. The Bayesian network model was clearly superior to the logistic regression model.

Another study investigated the relationship among 20 SNPs in the apolipoprotein E gene and blood plasma apolipoprotein E levels (apoE) [89]. This study did not use a model for prediction, but instead searched for SNPs that were most related to apoE in bootstrapped models. Though the study was restricted to SNPs in one gene, the relationships modeled among SNPs are more complex than they would be using traditional association methods.

Chapter 5

Phenocentric Bayesian Networks

5.1 Introduction

As described in Section 4.5, Bayesian networks are a promising method to understand complex traits. Though they have been successfully used in candidate gene studies with $O(100)$ SNPs or less, conventional BN implementations do not scale up to current genomic dataset sizes. Most recent candidate gene studies use thousands of SNPs, and with the advent of whole genome association studies, methods capable of handling $O(500,000)$ are needed. Here, we describe a BN learning algorithm that focuses on prediction of a phenotype using large SNP genotype datasets for case-control association studies.

5.2 Challenges of Genomic Data Analysis

The analysis of large phenotype and genotype datasets used in the study of most complex traits requires methods more powerful than those currently available [34, 90, 91, 49]. Traditional approaches that look at one SNP or characteristic at a time, such as those described in Section 2.3 are inadequate to find the complex interactions that underlie complex traits. Multivariate methods are an improvement over univariate methods in that they examine more than single interactions between genetic and phenotypic independent variables and an outcome of interest. Logisitic regression models are the most common multivariate method used in association studies described in Section 2.3. Although they can account for interaction terms among SNPs and phenotypic covariates such as age and gender, they have serious shortcomings for the study of complex traits [92]. Because the number of parameters

needed to fit a logistic regression grows exponentially as inter-variable interaction terms are considered, logistic regression models are usually constructed with no greater than pairwise interaction terms. Even in this situation, calculations can become intractable with a large number of variables. Additionally, the independent variables in a logistic regression model are treated as covariates rather than random variables, which causes the identification of genotyping errors and missing genotypes difficult to handle in genetic association studies.

Bayesian networks are a powerful PGM technique that can overcome the limitations of conventional association study approaches as discussed in Sections 3.5 and 4.5. Going beyond single or pairwise gene interactions with a phenotype, BNs are able to account for complex multivariate interactions. Addtionally, they can be used to make predictions of a phenotype of interest for individual subjects that allows for ascertainment of model validity. BNs are better able to find relationships among a large number of variables than logistic regression models. However, the performance of BNs is challenged by large genomic datasets. The recent advent of genomewide association (GWA) studies is promising for uncovering the genetic architecture of complex traits. The magnitude of data produced in such studies eclipses data produced by earlier candidate gene and linkage studies. Initial GWA studies measured over 100,000 SNPs in hundreds of subjects [93, 94], and current GWA study sizes have increased to over 500,000 SNPs in thousands of subjects [95]. With the promise of this massive data comes the challenge of proper and efficient analysis [7]. Conventional single-SNP analysis of large datasets is able to find some of the most common or penetrant genetic variants for a trait, but is unable to provide a thorough picture of the complex genetic dependencies that produce the traits. Without more powerful multivariate methods that can scale to large datasets, the promise of fully understanding the genetic underpinnings of complex traits will likely not be fulfilled.

5.3 Learning Gene Association Bayesian Networks

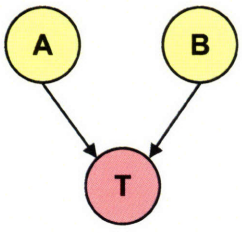
The complex dependency relationships among SNPs and a trait can be modeled with BNs. In the simplest case, one SNP modulates a trait. Figure 5.3 shows the graph and conditional probability table (CPT) for this case, where SNP A modulates a trait. According to the CPT, there is a higher likelihood for trait presence if SNP A's genotype is AA and for trait absence if SNP A's genotype is aa.

SNP A Genotype (A)		
AA	Aa	aa
0.36	0.48	0.24

Genotype	Trait (T)	
SNP A	Present	Absent
AA	0.2	0.8
Aa	0.4	0.6
aa	0.9	0.1

Figure 5-1: SNP A modulates a trait. The BN graphical representation (left) and the associated conditional probability table (right) are shown.

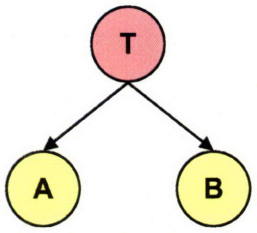
Expanding this model, let the trait be modulated by two SNPs, A and B [Figure 5.3]. The CPT's size has increased to 2×9, and the dependencies among genotypes have become more complex. With m SNPs, the CPTs size grows as 2×3^m, which becomes computationally unmanageable for large m. In this model, the assumption of independence between SNPs has been made. In many biological situations, SNPs are independent of one another so this assumption is correct. However, SNPs can be dependent through linkage disequilibrium or other biological mechanisms.



Genotype		Trait (T)	
A	B	Present	Absent
AA	BB	0.2	0.8
AA	Bb	0.3	0.7
AA	bb	0.3	0.7
Aa	BB	0.2	0.8
Aa	Bb	0.6	0.4
Aa	bb	0.6	0.4
aa	BB	0.3	0.7
aa	Bb	0.8	0.2
aa	bb	0.9	0.1

Figure 5-2: SNPs A and B modulate a trait. The BN graphical representation (left) and the associated conditional probability table (right) are shown.

For the sake of computational efficiency and to not make assumptions about SNP independence, the representation in Figure 5.3 can be changed to that of Figure 5.3. In this alternate representation, the trait is independent and it modulates SNPs A and B. The SNPs are conditionally independent given the trait and SNP independence is no longer assumed. As new SNPs are added to the model in this representation, there is linear growth of small CPTs: with m SNPs, m 2×3 tables are required.



	SNP A Genotype		
Trait (T)	AA	Aa	aa
Present	0.1	0.2	0.7
Absent	0.6	0.3	0.1

	SNP B Genotype		
Trait (T)	BB	Bb	bb
Present	0	0.1	0.9
Absent	0.7	0.2	0.1

Figure 5-3: BN representation where the trait is independent and SNPs A and B are conditionally independent given the trait. This dependency can be represented by two 2×3 conditional probability tables.

Although having trait be the parent node is computationally efficient and makes more realistic assumptions about SNP independence, the quantity of interest is the probability of trait absence or presence (T) given a genotype (G), $P(T|G)$. Using Bayes Theorem, we can find this measure by inverting the relationships between trait and SNPs. The joint probability for the representation in Figure 5.3, where the genotype is given by SNPs A and B is $P(T, A, B) = P(T)P(A|T)P(B|T)$. To obtain $P(T|A, B)$, we use $P(T|A, B) = P(T)P(T|A)P(T|B)$, where for each SNP S, $P(T|S)$ is given by Bayes Theorem:

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}$$

Thus we have an efficient basis for adding SNPs to a BN describing a trait. In addition to having SNPs that are dependent on trait, some added SNPs will be marginally independent of it.

One way to implement BN learning algorithms for genetic association models is to use this structure in which trait is the root node of the network and the genotypes are either conditionally dependent or marginally independent of it. This dependency structure can represent the association of independent as well as interacting SNPs with trait. Additionally, this structure captures complex models of dependency because the marginal likelihood measuring the association of each SNP with trait is functionally independent of the association of other SNPs with trait. In conventional BN learning algorithms, such as the K2 algorithm described in Section 4.3, relationships among all nodes are explored. The K2 greedy-search strategy considers ordered nodes in turn, and builds a list of parents based on nodes that have already been considered. Although only a subset of all nodes are considered at each step, if the number of nodes becomes too large, the algorithm becomes intractable.

5.4 A Novel Discovery Procedure

To make predictions of a node in a BN, knowing the node's parents, its children, and the parents of its children are all that is required. This set, known as a node's *Markov blanket*, directly modulates the node of interest. Thus, even if a strategy such as K2 is used to learn a full set of relations among nodes in a network, only the Markov blanket of each node is used to make predictions. This suggests a learning algorithm that focuses on finding the Markov blanket of a node, when the prediction of one node is of interest. In the case of genetic prediction of a phenotype, this strategy is optimal for finding the SNPs that best predict a phenotype. We call the corresponding network a *phenocentric Bayesian network* (PBN). Given a set of nodes S, a phenotype p, and a Bayes factor threshold bf, use the following algorithm to find the PBN:

```
PHENOCENTRIC-BAYESIAN-NETWORK(S, bf)
 1  Net ← []
 2  while max_{i∈S}[SCORE(p → i)/SCORE(i)] > bf
 3      do S ← S − i
 4         M ← p
 5         child ← i
 6         while max_{j∈S}[SCORE(M ∪ j → i)/SCORE(M → i)] > bf
 7             do M ← M ∪ j
 8         parents ← M
 9         Net ← EXPAND-NETWORK(Net, child, parents)
10  return Net
```

In this algorithm, nodes are greedily selected to become children, and then parents of each child are found. The metric to choose whether a node should be kept is Bayes factor, a ratio of likelihoods for a model with or without the node in question. The likelihood, referred to as SCORE in the algorithm, can be the log likelihood given in Equation 4.4, the Bayesian Information Criterion (BIC), or an analogous measure. The bf can be changed to adjust the stringency of model selection. In the most liberal case, $bf = 1$. The EXPAND-NETWORK refers to an algorithm that adds a *child* node and its *parents* to a network Net. The network returned is that with the highest likelihood of predicting phenotype given genotype.

The network obtained is quantified using the conditional probability distribution of each node given the parent nodes. Conditional probabilities are estimated using:

$$P(x_{ik}|\pi_{ij}) = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}},$$

where x_{ik} represents the state of the child node, π_{ij} represents a combination of states of the parent nodes, n_{ijk} is the sample frequency of (x_{ik}, π_{ij}) and n_{ij} is the sample frequency of π_{ij}. The parameters α_{ijk} and $\alpha_{ij} = \sum_k \alpha_{ijk}$ encode the prior distribution with the constraint $\sum_j \alpha_{ij} = \alpha$ for all j. The parameter α is chosen by sensitivity analysis. Predictions with the network can be performed using conventional approaches, such as those described in Section 4.4.

5.5 Conclusion

A novel approach to learn BN has been described, which focuses on learning the relationships that will best predict the outcomes of a given node. Addressing the need for better analytic methods for the study of complex traits, the PBN approach is tailored for gene association studies where the goal is to successfully predict a trait given a set of genetic markers (i.e. SNPs). Tailoring PGM and especially BNs for the study of complex traits is the most promising approach for modeling traits that accounts for their complex genetic underpinnings and has a quantitative metric to assess their predictive accuracy for individuals.

Part III

Asthma Management Features as Complex Traits

Chapter 6

Asthma

6.1 Definition

Classically, asthma is recognized by signs and symptoms including shortness of breath, cough, and wheezing. These findings are not very specific and can be attributed to many other respiratory disorders. According to the Global Initiative for Asthma Management and Prevention, asthma is a chronic inflammatory disorder of the airways in which many cells play a role, in particular mast cells, eosinophils, and T lymphocytes. In susceptible individuals this inflammation causes recurrent episodes of wheezing, breathlessness, chest tightness, and cough particularly at night and/or in the early morning. These symptoms are usually associated with widespread but variable airflow limitation that is at least partly reversible either spontaneously or with treatment. The inflammation also causes an associated increase in airway responsiveness to a variety of stimuli [14].

6.2 Impact

Asthma is a serious global problem affecting 20.5 million Americans and over 300 million people around the world [13, 14]. Its high and rising prevalence in most of the world has resulted in asthma being refered to as an epidemic [16]. Figure 6-1 shows the estimated burden of asthma around the world according to 2004 estimates [96]. Asthma is a costly disease, as demonstrated by the increased risk of emergency room visits, hospitalization, and sick absences that are associated with it [17, 18]. Over $16 billion are spent yearly in the US on asthma-related healthcare expenses [13].

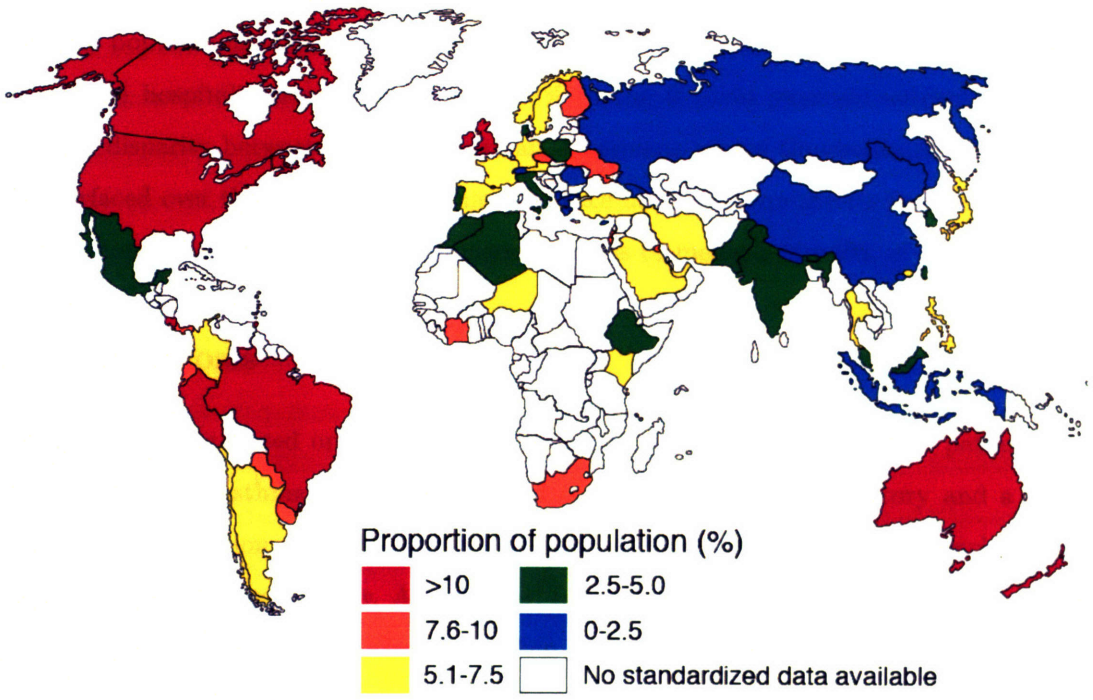


Figure 6-1: Prevalence of asthma around the world [14].

Both the prevalence and death rate of asthma rose dramatically in the US and globally between 1960 and 2001, and have remained at stable levels or slightly decreased since then [15, 13]. According to a CDC study looking at US asthma data from 1982 to 1992, self-reported asthma rates increased by 42% (1995). This same study found that in the 5 to 34 year age group, the rate increased by 52% (from 34.6 to 52.6 per 1000). Because this age group is thought to have the most accurate asthma diagnosis, this increase emphasizes that asthma prevalence rose during these years. According to other estimates, the number of asthma cases reported between 1980 and 1995 increased by 75% (from 30.7 to 53.8 per 1000), and asthma cases in children under the age of five increased by 160% [15]. The prevalence and morbidity of asthma have also been found to be on the rise worldwide [14]. In most countries who keep records, asthma prevalence has been rising through 2002 [16]. Explanations for the increased prevalence of asthma have been proposed. They include increased rates of diagnosis due to increased physician and public awareness of asthma symptoms and increased environmental factors that contribute to asthma. However, it is more likely that asthma is more prevalent due to increased risk factors than increased diagnosis [97].

In the US, asthma mortality and hospitalization rates have increased, particularly in minority populations [15]. From 1980 to 1993, the asthma death rate increased 118% and the annual hospitalization rate increased 28% in the 0 to 24 year old subject category. A racial disparity between deaths in African Americans versus Caucasian Americans has also surfaced over this time period. This difference has been hypothesized as being due to socioeconomic factors, including differences in access to medical care [98, 99].

6.3 Diagnosis

Asthma is not diagnosed on the basis of a single clinical test or a definitive pathological process. Instead, asthma is diagnosed based on a thorough clinical history and a series of pulmonary function tests, including measures of lung volume, airflow, bronchodilator response, and diffusing capacity. Age, gender, family history, and race are not highly useful in the diagnosis of asthma. Although it often manifests initially in children, asthma can occur at any age [100]. Differences in rates of asthma based on gender in different age groups have been observed, but these differences are not significant enough to preferentially suspect asthma in a patient of a particular gender [101]. Sufficient evidence exists to support the heritability of asthma, and about 50% of children with asthma have a positive parental history [102, 103]. However, the predictive value of a family history has not been studied and so is not helpful in the diagnosis of asthma. Studies on the incidence of asthma by race have found little difference of clinical utility for diagnosis, other than the finding that asthma is rare in Inuit populations [104, 105].

Clinical Lab Tests

Routine clinical lab tests that can be performed to help diagnose an asthmatic include chest radiographs and allergy tests. Chest radiographs primarily help rule out other causes of signs and symptoms suggestive of asthma or to find evidence of asthma complications. Allergy tests are performed because some patients who develop asthma are thought to do so as a consequence of underlying atopy (i.e. allergic tendency). In blood tests, elevated eosinophil count and serum immunoglobulin E (IgE) concentration are supportive of atopy. However, extremely elevated eosinophil percentage ($> 15\%$) suggests diagnoses other than asthma such as parasitic infections and pulmonary infiltrates. Extremely elevated IgE

levels ($> 1000ng/mL$) suggest asthma as well as eczema and allergic bronchopulmonary aspergillosis, two conditions associated with asthma. Identification of specific allergens is also possible with the use of skin tests and radioallergosorbent tests (RAST). To perform an allergy skin test, a set of antigen-containing solutions are injected via individual lances into the epidermis of a subject. Allergic reactions are identified as those where a large wheal and flare reaction form within 10-15 minutes after the antigen is administered. RAST tests are tests in which a subjects blood is exposed to particular antigens, and then IgE antibody levels are measured around such antigens. Many airborne allergens are known to cause asthma, including dust mite antigen, cat and dog danders, cockroach antigen, pollens, and mold spores. Association of these allergens to asthma can sometimes be elicited from a patients history, in which case the skin or RAST tests provide further evidence that a person has allergic asthma. Knowing what substances a subject is allergic to helps in determining potential asthmatic triggers and taking steps towards avoiding them.

Pulmonary Function Tests

Lung dysfunction in asthmatic patients includes (1) variable airflow limitation that is reversible with bronchodilator administration and (2) airway hyperresponsiveness, which is an excessive decrease in airflow in response to specific stimuli. Monitoring airflow is therefore essential to diagnose and track asthmatic patients. Pulmonary function tests (PFTs) are tests designed to measure parameters that are related to airflow. These tests provide an objective measure of pulmonary capacity that is correlated with disease severity. Measures obtained with PFTs include spirometry values, bronchial hyperresponsiveness and reversible airway obstruction.

Spirometry

Spirometry is the most useful and readily available pulmonary function test. It consists of a series of pulmonary function measures that are performed with a dedicated instrument. Because the instrument costs approximately $2000, the tests are usually performed in a clinical setting. The most usefule measures obtained include forced expiratory volume in one second (FEV_1) and forced vital capacity (FVC).

FEV_1, the total volume of air exhaled after 1 second, is the most important measure used to determine airway obstruction [106]. It decreases linearly with worsening airway

obstruction, and increases with relief of airway obstruction.A conventional classification scheme is: borderline obstruction greater than 80% predicted FEV_1; mild obstruction 60-80% predicted FEV_1; moderate obstruction 40-60% predicted FEV_1; severe obstruction less than 40% predicted FEV_1. Serial FEV_1 values can be measured to assess pulmonary status in a single patient over time, although these are usually performed using peak expiratory flow rate measures by a patient at home. FEV_1 is known to be dependent on gender, age, height, and race. Measures of a patients FEV_1 can be compared to predicted values from established normal populations to determine airway obstruction. Commonly used equations to calculate predicted FEV1 are those by Knudson et al [107].

The FVC is the total volume of air exhaled during a maximal exhalation. Because it does not fall as much as FEV_1 with obstructed airways, FVC is a less useful measure of airway obstruction. FEV_1 is dependent on the volume of inhaled air: if a subject does not inhale maximally before exhaling, the FEV_1 will fall. Therefore, FVC measures provide a standard value to correct FEV_1 for amount of air inhaled. A measure composed of these two, which has been shown to be very sensitive to airflow limitation, is the FEV_1 to FVC ratio (FEV_1: FVC) [106, 108]. Spirometry values are dependent on subject effort and cooperation. A properly trained person should coach a patient through the spirometry test for more reliable results. The reproducibility of FEV_1, FVC, and FEV_1: FVC have been determined to be 5% or less [106, 108]. Therefore, spirometry values are an important objective measure that can help in the diagnosis of asthma.

Bronchial Hyperresponsiveness

During bronchoprovocation testing, a patient is stimulated with a known bronchoconstrictor, such as inhaled metacholine, and pulmonary function tests before and after the provocation are compared to determine bronchial hyperresponsiveness. The performance of normal subjects and asthmatics is significantly different in such a test. Besides inhaled metacholine, substances used to stimulate bronchoconstriction include exercise, hyperventilation of cold and/or dry air, and inhalation of histamine. Bronchoprovocation testing can also be used to test for suspected asthmatic precipitants in a patient. The most common metacholine bronchoprovocation test consists in (1) performing spirometry before and after inhalation of a negative control (e.g. saline), (2) administering five dosed metacholine inhalations followed by a spirometry test after two minutes, (3) if the FEV_1 decreases by 20% or less

than the initial value, five metacholine inhalations are repeated at a higher dose followed by another spirometry test after two minutes, (4) step (3) is repeated until the FEV_1 decreases by more than 20% of the initial value or until the highest metacholine dose is administered. If the FEV_1 drops by more than 20% at low doses of metacholine, then the subject is said to have bronchial hyperresponsiveness. The result of a bronchoprovocative test is usually reported as the dose of metacholine (or stimulating agent) administered that resulted in a decrease in FEV_1 of 20%. In the normal population, studies have estimated that 7% of individuals will have bronchial hyperresponsiveness. Diseases other than asthma can lead to positive bronchoprovocation tests, including allergic rhinitis. Because the false negative rate of this test is estimated to be less than 5%, negative test results are valuable in ruling out asthma in an individual.

Reversible Airway Obstruction

Reversible airway obstruction is a classic, but not necessary, finding in asthma. The most common test to evaluate it is a bronchodilator test. The protocol of such a test in most labs involves (1) obtaining a pre-bronchodilator set of spirometry values, (2) taking two metered-dose inhaler inhalations of a rapidly acting beta-agonist (e.g. albuterol), (3) waiting 10-20 minutes to allow the beta-agonist to take effect, and (4) obtaining a post-bronchodilator set of spirometry values. An increase in FEV_1 following bronchodilator administration is a typical finding in an asthma patient. One recommendation for a significant bronchodilator response in an adult is that FEV_1 or FVC increase by 12% and at least 200mL (1991). In some cases, a patient may not show FEV_1 improvement after bronchodilator administration despite a subjective feeling of improved breathing capacity. This improvement may be measurable by other parameters, such as lung volume. Therefore, although FEV_1 often increases after bronchodilator administration, a lack of change does not necessarily imply that the bronchodilator had no effect. Bronchodilator response is further discussed in section 7.4.

6.4 Environmental Risk Factors

Many environmental factors contribute to the incidence and severity of asthma, including some that are currently unknown. Strong links to individual environmental factors can be

difficult to find because such a wide range of environmental and genetic factors contribute to asthmatic status. Nonetheless, several environmental contributors involved in asthma are known. They include: exposure to indoor allergens, outdoor air pollution, respiratory infections, smoking and exposure to tobacco smoke.

Indoor Allergens

Indoor allergens, including dust mites, animal allergens, cockroach allergen, and endotoxin, play a significant role in the onset of asthma. Exposure to such allergens has increased as indoor living (i.e. houses that are isolated from outdoor air) has become more widespread in the West. Concomitantly, the prevalence of asthma has risen in developed countries [97, 109]. However, studies have found no association between increased levels of exposure to dust mites early in life and the developement of childhood asthma [110, 111]. Perhaps some individuals are more likely to develop asthma after indoor allergen exposures because of an underlying genetic predisposition.

Outdoor Air Pollution

Outdoor air pollution is known to be associated with lung disease, but it has not been clearly associated with asthma. Population studies after the German reunification are some of the best large-scale natural experiments that can look at the effects of different environments on genetically similar individuals. Lack of association of air pollution to asthma in this population provides strong evidence for no association, but the issue remains controversial because of the association between pollution and respiratory illness [112, 113]. In the US, levels of pollution were found to be correlated to bronchitis and chronic cough, but not to asthma [114]. However, exposure to pollutants is associated with asthma exacerbations [114, 115]. Other studies confirmed that asthmatics react differently to pollutants than non-asthmatics: asthmatics reacted (i.e. wheezed) to lower concentrations of inhaled sulfur dioxide, but reacted similarly to ozone and nitrous oxide [116, 117]. These studies also investigated whether outdoor air pollution was associated with asthma incidence, and the authors concluded that the answer is negative.

Respiratory Infections

Respiratory infections are known to exacerbate asthma [118, 119] and there is an association between viral respiratory infections and asthma development in adults [97, 120]. In children, some studies have provided evidence that lower rates of respiratory infection lead to increased asthma and atopy prevalence [121]. Conversely, frequent respiratory infections during childhood seem to decrease the likelihood of developing asthma later in life [122].

Tobacco Exposure

As with outdoor air pollution, smoking and exposure to tobacco smoke is related to pulmonary illness. Some studies have found a relationship between smoking and the development of asthma [123]. In adults, exposure to tobacco smoke has been associated with a slight increase in asthma, odd-ratio of 1.39 [124]. Secondhand smoking is especially harmful in the development of asthma in the case of smoking mothers of young children [125, 126]. Such children are twice as likely to develop asthma than their peers with non-smoking mothers. Prenatal exposure to maternal smoking is also associated with an increased risk of developing asthma [127, 128].

6.5 Pathogenesis

Inflammation is known to be an early event in asthma. Infiltrating cells are usually found in airway biopsies of newly diagnosed asthmatics [129]. Figure 6-2 shows some of the known components of airway inflammation at a cellular level [130]. Some of the major characteristics of this inflammation are infiltration of airway wall by eosinophils and lymphocytes,

Eosinophil and Lymphocyte Infiltration of Airway Wall

A study investigating the infiltrate distribution in the airways found that a higher inflammatory cell density is found in smaller airways, which may help explain the characteristic peripheral airway obstruction of asthma [131]. Specifically, in the large airways ($> 3.0mm$), the infiltrate locates mostly to the region between the basement membrane and smooth muscle layers. In smaller airways ($< 3.0mm$), infiltrate locates more to regions between the smooth muscle layer and alveolar attachments than to the region between basement membrane and smooth muscle. Further evidence for the obstruction being caused by infiltrates

is provided by studies that correlated the amount of eosinophils within the airway wall with asthma severity [132, 133]. Eosinophils involved in asthma are in an activated state, but the mechanisms of activation are poorly understood. One hypothesis is that stimulation of high affinity IgE receptors (FcERI) and/or low affinity IgE receptors (FcERII) leads to the eosinophil activation.

Many of the lymphocytes that infiltrate airways are of a TH2 subtype, which are known to produce IL3, IL4, IL5, and GM-CSF, but not interferon-gamma, and to express CCR3 [134]. Production of IL4 and IL5 further increase an allergic response by contributing to the formation of mast cells, the differentiation of TH2 lymphocytes, and the differentiation and chemotaxis of eosinophils [135]. The release of cytokines promotes the differentiation of plasma cells that can produce IgE against specific antigens [136]. The IgE molecules, which are transported through the circulation, attach to mast cells and eosinophils via the FcERI [137]. When the mast cells are reexposed to antigen, they secrete more mediators and cytokines, which perpetuate the asthmatic response.

Some transcription factors are known to be involved in asthma. Members of the cytokine signaling family known as signal transducers and activators of transcription (STATs) are known to be constitutively activated in asthma [138]. In addition to eosinophils and lymphocytes, neutrophils are found in the lung airways of severe asthmatics [139].

Inflammation of Airway Parenchymal Cells

The phenotype of cells that are normally present in airways, as opposed to infiltrates, becomes more inflammatory. The most prominent change is the sensitization of mast cells by IgE towards specific antigens [140]. Airway smooth muscle becomes hypertrophic and hyperplastic [141] and airway epithelium becomes thickened and dysplastic, which causes loss of the normal pseudostratified columnar arrangement [142].

Airway Remodeling

Noncellular components of the airway wall undergo inflammatory changes. Collagen is deposited at the basement membrane [143] and the loose areoloar connective tissue in the spaces between the epithelium and smooth muscle and outside the smooth muscle expands [144]. Both of these changes create a thickened airway wall that contributes to airway hyperresponsiveness by contributing to airway constriction [145].

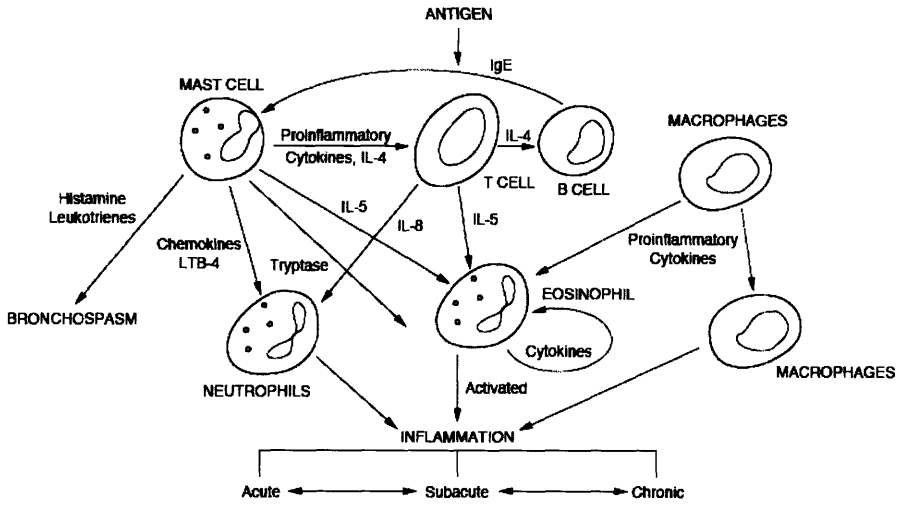


Figure 6-2: Cellular mechanisms of airway inflammation [130].

The above description of airway inflammation has been found to be incomplete by further asthmatic studies. When an IL5 antibody was administered to patients with mild asthma, airway eosinophilia was nearly eliminated but airway responsiveness did not improve [146]. Administration of IL12 into a cohort of asthmatics caused a reduction of eosinophils but no airway responsiveness change [147]. These findings put into question the role of eosinophils in causing the airway hyperresponsiveness that is characteristic of asthma. A more complete picture on the pathogenesis of asthma is likely to emerge with further studies.

6.6 Genetics

Asthma has a demonstrable genetic basis, with heritability estimates ranging from 0.36 to 0.87, and is known not to be transmitted in a simple Mendelian pattern [19, 20, 21, 22, 23, 24, 25]. Twin studies have attributed a greater genetic than environmental component to asthma [22, 23, 24, 25]. However, asthma results from the interaction of multiple genes with environmental and developmental factors, making it a prototypical complex disease.

Over 100 genes have been individually associated with asthma or a related phenotype [26]. Of these genes, 25 have been associated in six or more populations and 54 have been associated in two to five populations. Two of the regions with the strongest evidence for involvement in asthma by association studies are: chromosome 11q13 (to bronchial

hyperresponsiveness and total serum IgE) and a region on chromosome 5q (to total serum IgE) [148]. In addition to the genes found by linkage and association studies, positional cloning has found four genes that are associated with asthma: ADAM33 [149], DPP10 [150], PHF11 [151], and GPRA [152]. Some of the genes that are known to reside in 5q31-q33 are the beta-2 adrenergic receptor [153], some cytokine genes (IL4, IL5, IL9, IL13), glucocorticoid receptor 1 (GRL1), and GM-CSF [154]. Although there is some evidence that variants of the beta-2 adrenergic receptor gene correlate with IgE levels [155] and nocturnal asthma [156] in known asthmatics, the variants alone have not been found to be predictive of asthma. A case-control study found that the IL4 receptor was associated with high total or specific IgE levels [157].

Chapter 7

Asthma Management

7.1 Overview

The main goals of asthma management are to prevent and treat asthma exacerbations, episodes of worsening asthma symptoms, and to help patients lead as normal a life as possible by optimizing lung function and minimizing symptoms that interrupt daily activities. Although pharmacologic therapy is an important component of asthma management, good asthma control is also achieved by a combination of routine monitoring of patient symptoms and lung function, controlling asthma triggers, and patient education [158]. Performing these latter tasks is important to ensure that a patient has an optimal medication regimen.

Patient Monitoring

Routine monitoring of asthma symptoms and lung function are helpful to prevent the worsening of asthma [158]. Symptoms that should be routinely monitored include how often reliever medications are taken, how often patients wake up during the night with asthma symptoms, how many school or work absences due to asthma have taken place, how many times asthma interfered with daily activities, and how many exacerbations have occured. An increase of one or many of these symptoms suggests that asthma is poorly controlled. Pulmonary function monitoring helps quantify the amount of airflow obstruction in a patient. Serial measurements in an individual patient serve to determine how well asthma is being controlled.

Controlling Asthma Triggers

Identifying and avoiding triggers of asthma, is an important component of successful asthma management that often results in reduced exacerbations and medication use. Common triggers are those described in the Environmental Risk Factor Section 6.4: indoor allergens, outdoor pollutants, respiratory infections, and tobacco smoke. Other common triggers include physical activity, emotional stress, cold air, gastroesophageal reflux, and medications such as aspirin [158].

Patient Education

Patient education, which has been shown to reduce asthma hospitalization rate, improve daily function, and improve patient satisfaction [159, 160], involves patients in the monitoring of their symptoms and pulmonary function, identifying and avoiding asthma triggers, and properly using asthma medications.

7.2 Pharmacologic Therapy

There are two broad categories of asthma drugs: reliever drugs and controller drugs [161]. Reliever drugs attempt to reverse acute bronchoconstriction. The most common type of reliever drugs, which are the treatment of choice for mild asthmatics, are β_2-agonists (e.g. albuterol, metaproterenol, pirbuterol, levalbuterol). The β_2-agonists act by relaxing bronchial smooth muscle via β_2-adrenergic receptor activation. More details regarding β_2-agonists are in Section 7.4. For moderate or severe asthma, reliever drugs are often combined with controller drugs. Controller drugs reduce the severity of airway inflammation and obstruction. The most common types of controller drugs are inhaled corticosteroids (e.g. budesonide, beclomethasone, flunisolide, fluticasone) and leukotriene modifiers (e.g. montelukast, zafirlukast, zileuton). Corticosteroids are potent anti-inflammatory agents that affect T-lymphocyte responses. Their binding to cytoplasmic glucocorticoid receptors leads to their translocation to cell nuclei and the transcription of anti-inflammatory genes. Corticosteroid administration decreases airway inflammation and hyperresponsiveness by inhibiting inflammatory cell recruitment and production of cytokines [162]. Although corticosteroids and β_2-agonists are different forms of therapy that have unique pathways of action, there is growing evidence that these molecular pathways overlap [163, 164, 165]. One of the effects

they have in common is the ability to activate glucocorticoid receptors [166, 167]. Synergistic effects between the two drugs are due partly to the increased nuclear translocation of corticosteroid-activated glucocorticoid receptors after administration of long-acting β_2-agonists [168]. Clinical trials have demonstrated the efficacy of using both corticosteroids and long-acting β_2-agonists in the treatment of asthma [169, 170]. Leukotrienes are potent biochemicals released from mast cells, eosinophils and basophils, that contract airway smooth muscle, increase vascular permeability, increase mucus secretions, and attract and activate inflammatory cells [171]. Leukotriene modifier drugs attempt to reverse the effects of leukotrienes, which helps counteract the inflammatory response in asthma patients [172].

Asthmatics do not respond uniformly to therapy, as studies of medication efficacy have found [173, 174, 175, 176]. Further, as many as one-half of asthmatic patients do not respond at all to the most efficacious current asthma therapies, namely beta-agonists, corticosteroids, and leukotriene modifiers [175, 177, 178]. There are various factors that lead to the variability in drug reponse, but a substantial portion is thought to be due to genetic differences. The identification of patients that would respond to a specific treatment, which would greatly enhance treatment efficacy, is currently hindered by the lack of a definition of what it is to be a non-responder and the mechanism of such resistance.

Antiasthmatic and bronchodilator drugs are some of the most highly prescribed medications in the US [179]. According to 2003-2004 National Center for Health Statistics records, they were the second most prescribed drugs in children after penicillins, and the third in all age groups after antidepressants and non-steroidal anti-inflammatory drugs. Although these drugs are prescribed for more respiratory conditions than just asthma, particularly in older adults, these statistics highlight how common asthma drugs are despite the heterogeneity of their efficacy.

7.3 Asthma Exacerbation

Asthma exacerbations, commonly known as asthma attacks, are the major cause of morbidity, mortality and healthcare costs for individuals with asthma [27, 28, 29]. Therefore, they are one of the main targets f asthma management. Exacerbation episodes involve worsening of asthma symptoms, including shortness of breath, cough, wheezing, chest pain or tightness, mucus production, or some combination of these. In 2004, 11.7 million Amer-

icans (3.9 million children under 18) had an asthma attack [13]. This comprises 57% of the 20.5 million Americans who are estimated to have asthma. American Lung Association data gathered between 1997 and 2004 consistently show that children 5-17 years old have the highest exacerbation rates [13]. Indeed, asthma is the third leading cause of hospitalizations in children, occurring an estimated 198,000 times per year [180]. Exacerbations present similarly in both sexes, but the rates are higher in young boys than in girls during childhood and are higher in adult women than adult men [181].

Pathophysiology

The airway narrowing causing airway obstruction in asthma exacerbations is due to a combination of smooth muscle contraction, thickening of airway walls, and secretions within airway lumen. A sudden onset of these events throughout the tracheobronchial tree, resulting in a severe reduction in airflow, is what comprises an asthma attack. After an attack, the obstruction of airways reverses from larger to smaller ones. Initially, usually after hours to days, the trachea, mainstem bronchi, lobar bronchi, and segmental bronchi reopen. The smaller peripheral airways may not return to normal until after weeks or months of time have elapsed.

Determining the exact cause of exacerbations is difficult. Many exposures have been linked to asthma exacerbations, but causal mechanisms remain unclear. The majority of exacerbations are associated with respiratory viral infections (RVIs), especially rhinovirus [182, 183, 184]. In particular, respiratory viruses are found in over 80% of children with exacerbations [183]. Knowing that viruses are involved in asthma exacerbations has motivated studies to identify the cellular mechanisms through which exacerbation takes place. Chemokines have been identified as important mediators of respiratory viral infection [185]. For instance, CCL5 and CXCL8 are related to exacerbation of allergic asthma. In asthmatic subjects with rhinovirus infection, CXCL8 levels correlated with severity of symptoms [186]. Pollutants [114, 115] and allergens [187, 188] have also been linked to asthma exacerbations, mostly by increasing the propensity towards exacerbations due to RVIs.

Treatment and Prevention

The main approach to managing exacerbations is to prevent them. This is accomplished by managing asthma and by attempting to reverse worsening symptoms before a severe exacerbation develops. Most patients have plans with instructions to increase their medications at home as their symptoms worsen. Plans include a threshold of symptoms for which a patient should seek medical help by either calling their physician or going to an emergency room or hospital. Once they occur, exacerbations are usually treated with a combination of short-acting beta-2 agonists and systemic corticosteroids until symptoms subside [161]. The inhaled beta-2 agonist helps to relax the smooth muscle in the airways, while the corticosteroid helps reverse inflammation.

Identification of At-Risk Patients

A subset of asthma patients suffers from frequent exacerbations. The proper identification this group is of clinical importance both to monitor it more carefully and to treat it more aggressively [189]. Additionally, it is important to identify those patients who are not at risk of exacerbations in order to not overmedicate them. The most severe group of exacerbators, those with near-fatal and fatal asthma, has been well described [190]. Predictors of this type of asthma, which results in respiratory arrest and/or death, are increased medication use (beta-agonists, oral steroids, and oral theophylline) and a history of hospital and/or intensive care unit admissions and mechanical ventilation. Prior emergency department assessment visits and use of inhaled corticosteroids have not been found to be predictors of severe asthma.

Some asthma patients suffer from frequent exacerbations that are not as severe as those of near-fatal and fatal asthmatics. These exacerbations are still a serious and costly health problem that interferes with patient's lives. Attempts to characterize patients who suffer from frequent exacerbations have been made, but this group is still not well understood. In one study, patients with multiple exacerbations per year compared to those with one exacerbation per year were more likely to be on higher doses of inhaled and oral corticosteroids, be hospitalized, have chronic sinusitis and be intolerant to non-steroidal anti-inflammatory drugs [191]. Medical conditions that have been associated with frequent exacerbations include severe nasal sinus disease, gastro-esophageal reflux, recurrent respiratory infections,

psychological dysfunction, and obstructive sleep apnea [192].

The preceding studies demonstrate that there is no good way to identify asthma patients who will have exacerbations, until the patients have an established history of exacerbations. Some of the findings, such as the fact that frequent exacerbators are more likely to be hospitalized than other asthmatics, does not help determine what patients are exacerbators until after they have develped a serious medical problem. What would be most useful is to identify who is at risk for exacerbations before a pattern of disease has been established. This would reduce the number of exacerbations and the morbidity, mortality and cost associated with asthma.

Genetic Basis

The underlying genetics of asthma exacerbations is unknown. Because exacerbations occur in some patients with asthma and not others who have been exposed to similar environments, a genetic basis that predisposes some individuals to exacerbations is likely. For example, subtle differences in a variety of chemokine genes could predispose some individuals to respond differently to viruses, differences in interleukin genes could predispose some individuals to repond more aggressively to stimuli, or differences in many groups of genes could create the end-effect of asthma exacerbations in some individuals but not others. Uncovering the genetic basis underlying asthma exacerbations would be helpful to understand the biology of exacerbations, discover novel therapeutic targets, and identify those at risk of suffering from them.

7.4 Bronchodilator Response

Bronchodilator responsiveness is a common clinical test that is used for the evaluation of reversible airway obstruction and the diagnosis of asthma. The basis of this test is to find out whether administration of a bronchodilator medication improves FEV_1 as described in Section 6.3. The physiological response to a bronchodilator is a complex trait, involving intricate interactions among airway epithelial and smooth muscle cells and nerves. In a single individual, repeated tests are quite variable if not performed in a pulmonary funciton test laboratory by a properly trained test administrator [193]. When properly performed, single bronchodilator tests are appropriate to assess airway responsiveness and effectiveness

of a bronchodilator drug in a patient although some measurement variability remains [194]. The lack of response to a bronchodilator does not always imply that a patient will not benefit from bronchodilator therapy [195].

Quantification of Response

In order to meaningfully interpret bronchodilator response tests, a reliable quantitative measurement has to be established. Studies have been conducted to establish the best comparison of pre- and postbronchodilator measures, which are usually FEV_1 and FVC. The most common ones are absolute change (ΔFEV_1), percent change from initial value ($\Delta FEV_1\%init$), and change in percent predicted ($\Delta FEV_1\%pred$), which are defined as follows:

$$\Delta FEV_1 = Postbronchodilator\ FEV_1 - Prebronchodilator\ FEV_1$$

$$\Delta FEV_1\%init = \frac{Postbronchodilator\ FEV_1 - Prebronchodilator\ FEV_1}{Prebronchodilator\ FEV_1} \times 100\%$$

$$\Delta FEV_1\%pred = Postbronchodilator\ FEV_1\%pred - Prebronchodilator\ FEV_1\%pred$$

Controversies still exist regarding which of these measures is most accurate and reliable. Because prebronchodilator FEV_1 is dependent on variables including a person's age, gender, height, and race, a way to normalize the change in FEV_1 to account for a person's baseline pulmonary function is generally favored. Of the above three measures, $\Delta FEV_1\%init$ and $\Delta FEV_1\%pred$ account for prebronchodilator lung function. In children, $\Delta FEV_1\%pred$ has been proposed as the best measure because it provides a measure that is independent of age, height, and prebronchodilator lung function [196]. Other studies have found that $\Delta FEV_1\%pred$ should be used because it is best at differentiating asthmatics from others [197] and because it is less dependent on prebronchodilator lung function and has the highest reproducibility among measures [198]. However, the recommended clinical measure is percent change from baseline: $\Delta FEV_1\%init$ [194, 199].

Response Thresholds

After choosing a measure for bronchodilator response, a criterion to differentiate bronchodilator responders from non-responders must be determined. Because bronchodilator response is a continuous variable, the values for such a threshold are arbitrary. Studying

the distribution of bronchodilator responsiveness in asthmatics and non-asthmatics provides reasonable estimates for thresholds to best differentiate these populations. A common choice is to use the upper 95th percentile of a sample measures from normal non-asthmatic subjects, who compared to asthmatics, do not have an increase of airflow after bronchodilator administration. This threshold for the $\Delta FEV_1\%pred$ in one population was 9% [200]. The most current clinical threshold for improvement of $\Delta FEV_1\%init$ is 12% or greater than 200mL, while improvements less than 8% or less than 150mL are considered to be within measurement variability [194, 199].

β_2-Agonists

The most potent and rapidly acting bronchodilators currently available for clinical use are β_2-agonists [30]. Their primary effect is to stimulate β_2 receptors on the surface of airway smooth muscle cells, which via an increase in intracellular cyclic AMP levels, relaxes airway smooth muscles and reduces bronchoconstriction. These drugs are the primary drugs used in bronchodilator tests and are routinely used for the pharmacologic management of asthma.

For the treatment of acute asthma exacerbations, β_2-agonists are the incontrovertible drug of choice, but controversies exist regarding the chronic use of these bronchodilators as a maintenance therapy in asthma. Two studies in the early 90s reported that chronic use of beta agonists was associated with increased mortality, decreased asthma control, and lower efficacy than inhaled corticosteroids [201, 202]. These studies lead to great concern regarding the safety of beta agonists, but more recent studies have found no or weak association of chronic beta agonist use to mortality [203, 204, 205, 206]. Similarly, most studies of the effect of chronic beta agonist use on asthma symptoms have found no evidence for increased complications (e.g. exacerbations) or decreased asthma control [207, 208]. Some concern regarding the safety of beta agonists still exists and patients who take them are monitored carefully. Treatement with inhaled corticosteroids is generally favored over treatment with beta agonists in most mild and moderate asthma patients because the former are more effective at reducing symptoms than the latter [209, 210]. When treatment with a corticosteroid alone does not help decrease symptoms significantly, combination long-acting beta agonist and corticosteroid therapy has been found to be effective [211, 212].

Regardless of the daily therapy choice for a patient, β_2-agonists remain the favored rescue medication. Therefore, understanding the effectiveness of β_2-agonists remains an

important question for most asthma patients because even if they do not take beta agonists regularly, they likely use them for worsening symptoms and/or exacerbation treatment. A better understanding of bronchodilator response tests would be helpful to establish what patients benefit from β_2-agonist therapy.

Genetic Basis

Evidence for the genetic basis of bronchodilator response has been establihsed in a family aggregation study and genetic association studies. Familial aggregation of bronchodilator response was established in a study of 1,161 families in a rural community in China that found correlations of adjusted $\Delta FEV_1\%init$ values in parent-offspring pairs [213].

Genetic variants of the beta-2 adrenergic receptor have been shown to change the bronchodilator response of individuals in four separate study populations [214, 215, 216, 217]. Although the reanalysis of two prospective studies found that polymorphisms of this gene can partially predict the patient response to inhaled albuterol [218, 219], these results have not been incorporated in clinical practice because they are not considered strong enough by clinicians.

A better understanding of the genetic basis of bronchodilator response would be helpful to identify patient-specific treatments, identify novel therapeutic targets, and help in the diagnosis and monitoring of asthma. Further, such a test would help establish what patients are responsive to β_2-agonists and what genetic mechanisms may be responsible for variability in patient response to such drugs.

Part IV

Phenocentric Bayesian Networks in Asthma Management

Chapter 8

Data and Methods

8.1 Subject Population

The Childhood Asthma Management Program (CAMP) is a multi-center, longitudinal, randomized, double-blinded clinical trial that followed 1,041 asthmatic children (5-12 years of age) for approximately four years [220]. The subjects were assigned to one of three treatment groups: budesonide, nedocromil, placebo. Subjects were selected for having mild to moderate asthma, which was assessed as those having had asthma symptoms and/or medication in six or more months of the previous year without requiring more than one asthma hospitalization or five or more prednisone bursts, having a history of intubation for asthma, an FEV_1 less than 65% of normal, or any other pulmonary disease. All subjects had an initial methacholine bronchoprovocation test resulting in 20% FEV_1 reduction to ensure a more objective definition of asthma was followed. Informed consent was obtained from all CAMP participants and their parents. Following completion of the clinical trial, an additional 922 subjects have been followed for an additional 6 years in the CAMP Continuation Study (CAMPCS). The studies were approved by the Institutional Review Board of the Brigham and Womens Hospital.

8.2 Clinical Data

Follow-up visits occurred every four months and spirometry was performed twice yearly. Data collected included responses to questions regarding asthma symptoms and severity and medications used (e.g. "How many times have you called the doctor since the previous

visit?", "How many overnight hospitalizations due to asthma have occurred since the last visit?", "How often have you used albuterol?"). Spirometry was performed according to American Thoracic Society recommendations with a volume-displacement spirometer, and airway responsiveness was assessed by methacholine challenge with the Wright nebulizer tidal breathing technique [220]. Spirometry and methacholine testing were performed by pulmonary function technicians trained and certified specifically for the CAMP protocol and procedures. Spirometry performed met or exceeded the American Thoracic Society (ATS) standards. Spirometry and methacholine testing were performed at least 4 h after the use of a short-acting bronchodilator and 24 h after the last use of a long-acting bronchodilator. Postbronchodilator (two puffs albuterol by metered-dose inhaler) measurements were taken at each spirometry session (18). After administration of the bronchodilator, the minimal elapsed time before the postbronchodilator test was 15 min. Equations used to predict the average value of lung function measures for age, sex, and height were race-corrected according to Coultas and coworkers [221] for Hispanics, and according to Knudson and coworkers [107] for all other ethnic groups. Total blood eosinophils were counted by center-specific methods. Serum total IgE was measured by radioimmunosorbent assays from blood samples collected during the CAMP screening sessions. Genetic data was collected for 968 children and 1518 parents, representing 582 complete nuclear families.

8.3 Genetic Data

Candidate Genes

Candidate genes were selected to be genotyped based on a previously identified association to asthma or a related phenotype. Some of the candidate genes, involved in innate immunity and pharmacogenetic pathways, are described below. A full list of genotyped genes (n = 441), named according to the September 2006 NCBI data (Build 36.2), is given in Table A.1. The biological pathways representing the genes according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [222] are given in Table A.2.

Innate Immunity Genes

Genes involved in innate immunity included chemokines, interleukins, toll-like receptors and other transmembrane proteins. Chemokines are pro-inflammatory cytokines that induce

chemotaxis in nearby responsive cells, especially leukocytes, which then recruit monocytes, neutrophils, and other cells involved in innate immunity. The Eotaxin gene family (eotaxin/CCL11 [223], eotaxin-2/CCL24 [224], eotaxin-3/CCL26 [225]), is composed of C-C chemokines that are potent eosinophil chemoattractants that act via a common receptor (CCR3) found primarily on eosinophil cell surfaces [226]. Eotaxins are involved in the recruitment of peripheral blood eosinophils into the lung during acute allergic inflammation [227, 228]. Interleukins (ILs) are cytokines that are involved in a wide range of immune responses. IL10 is an anti-inflammatory cytokine that inhibits the secretion of other proinflammatory cytokines (e.g. IL1, IL6, IL8, IL12), making it an immunosuppressant of T cells, monocytes, and macrophages (i.e. it suppresses the TH1 phenotype) [229]. It has been shown to have lower in vitro production from macrophages and mononuclear cells of asthmatics [230, 231] and have lower mRNA levels in bronchoalveolar lavage cell pellets of asthmatics [232]. IL10 variants have been associated to asthma phenotypes (i.e. FEV_1 percent of predicted and IgE levels) in children [233]. In the presence of IL4 in vitro, IL10 regulates IgE production [234]. Additional interleukins that may play a role in asthma include IL8, a C-X-C chemokine that potently chemoattracts and activates neutrophils [235]. The Toll-like receptor (TLR) group of transmembrane proteins is a highly conserved set of innate immune pattern recognition receptors. Motifs present in microbial antigens (e.g. bacterial DNA, lipoglycans, lipoproteins) stimulate TLRs in antigen-presenting cells, including tissue macrophages, blood monocytes, and dendritic cells, which leads to the activation of cytokines and other genes that mediate immune responses and initiate the transition from innate to acquired immunity [236, 237]. TLR10 [238] genotypes have been associated with asthma diagnosis.

Glucocorticoid and β_2-Agonist Interaction Genes

The β_2-agonist drugs act by binding to β_2 adrenergic receptors (β_2ARs) on smooth muscle cells. These receptors are G-protein coupled receptors that act by adenylyl cyclase activation leading to increased levels of cAMP and protein kinase A (PKA) activation. The activated PKA phosphorylate a variety of targets, which act to decrease intracellular Ca^{2+} thereby causing muscle relaxation. One β_2AR SNP has been associated with decreased pulmonary function response to albuterol treatment [218, 239] and an increased frequency of asthma exacerbations [219]. A study looking at β_2AR haplotypes based on 13 SNPs

found that bronchodilator response was related to the haplotype pair but not to individual SNPs [217]. The CAMP study also investigated β_2AR haplotype and SNP relationships to bronchodilator response and found that regions of this gene that may be significant in β_2-agonist treatment response [178]. Although the above studies demonstrate that β_2AR genotype plays a role in success of β_2-agonist therapy, a set of SNPs that accurately predicts successful treatment has not been identified.

The corticotropin-releasing factor type 1 and 2 receptors, CRHR1 and CRHR2, are G-protein coupled receptors involved in the Hypothalamic-Pituitary-Adrenal (HPA) axis. Corticotropin-releasing hormone (CRH) is the endogenous hormone that binds the CRHR receptors to activate the HPA [240]. The HPA is sometimes referred to as the stress axis because it plays a major role in the stress response. In addition to their primary role in stress, glucocorticoids are involved in modulating the immune system [241, 242]. Specifically, it has been shown in a mouse model that in the absence of CRH, endogenous glucocorticoid production decreases and airway inflammation increases [243]. Polymorphisms of the CRHR1 gene have been associated with differences in therapeutic response of asthmatics to glucocorticoids [244].

SNP Selection

SNP selection was performed such that a small set of SNPs distinguished the common haplotypes of the genes of interest [245]. Haplotypes were inferred using Bayesian methods as implemented in PHASE [79]. SNPs that distinguished the most common haplotypes were identified using the BEST algorithm [246]. Only haplotypes that were found in the Caucasian population at a frequency of 5% or greater were considered because the study population is composed of Caucasian subjects only. Rare SNPs (minor allele frequency ¡5%) were considered for genotyping if the SNP led to a nonconservative amino-acid change, implying potential functional significance. The number of SNPs genotyped for each gene, assigned according to the September 2006 NCBI data (Build 36.2), are shown in Table A.1. In addition to the SNPs in this table, data for 466 intergenic SNPs is available. Most of these SNPs are in linkage disequilibrium with candidate genes.

Genotyping

Most selected SNPs were genotyped with an Illumina BeadStation 500G using the GoldenGate assay, an allele-specific hybridization reaction. Briefly, $250ng$ of genomic DNA is obtained for each subject for the multiplex reaction. Three oligonucleotides are designed for each SNP loci. Two 5' oligonucleotides (P1' and P2') are specific for the two SNP alleles. The 3' base of each oligonucleotide is complementary to one of the two SNP alleles, allowing hybridization with one of the two alleles. The third oligonucleotide, 3' to the SNP, is specific for each SNP locus. The three oligonucleotides contain universal PCR primer sites. The 3' oligonucleotide includes an address sequence that will hybridize to a specific silica bead. After allele-specific hybridization and extension, the extended product serves as a template for a PCR reaction using 2 fluorescent-labeled primers (P1' and P2') and one unlabeled primer (P3'). The PCR products are then hybridized to the bead labeled with a complementary sequence to the address sequence on oligonucleotide P3'. To achieve a high-level multiplex assay consisting of thousands of SNPs per sample, the Illumina platform uses a bead-based fiber-optic array for each DNA sample [247]. For a 1536 SNP array, there is a matrix of 50,000 individual fibers so that each SNP is represented approximately 30 times [248]. For each SNP, a sequence complementary to the address tag on P3' has been hybridized to a silica bead randomly assembled into the matrix of optical fibers. After array manufacture, a series of DNA hybridizations is used to decode the location of each randomly-located bead [249]. Following hybridization of PCR products to the beads, the array is scanned at two different wavelengths, and the fluorescent output for each SNP is recorded. Software is used to integrate fluorescent signals to obtain bead location in each array, thereby deciphering the genotype for each SNP of each subject. The genotype calls are highly accurate, with call rates over 99.5% and greater than 99.5% reproducibility between duplicate samples [248].

Subjects missing more than 5% of SNP data, and SNPs missing in more than 5% of subjects were dropped. Hardy-Weinberg equilbrium was checked in all SNPs among control subjects using the exact procedure in [250] with a $p = 0.01$ significance threshold, and those that were not in equilibrium were droped. SNPs with minor allele frequence (MAF) less than 5% were dropped. Missing alleles were imputed marginally from the HWE distribution among controls.

8.4 Traditional Association Tests

The Cochran-Armitage test for trend as implemented in the SAS FREQ procedure is used to measure the association of single SNPs to a phenotype of interest [251]. Variables with exact p-values less than 0.05 are reported as significant.

Binomial logistic regression models are built in SAS with the LOGISTIC forward stepwise procedure [251]. This model finds covariate variables most strongly associated with a response variable according to Fisher's scoring criterion. Regressors with $p < 0.05$ are reported as significant.

8.5 Predictive Validation

The predicted probability of an outcome of interest, given evidence in a Bayesian network, is calculated using the clique algorithm implemented in Bayesware Discoverer [61]. Goodness of fit is assessed using fitted values, by predicting phenotype in each subject used to construct the network. Network robustness is assessed via a twentyfold cross-validation in which each of twenty non-overlapping data subsets, obtained by randomly splitting the original dataset, is used as an independent dataset while the remaining 19 subsets are used to learn the network dependencies. Fitted values or cross-validation predicted probabilities are compared to actual phenotypes using receiver operator characteristic (ROC) curves.

ROC curves are plots of *sensitivity* versus $1 - specificity$ that are commonly used to evaluate the goodness of tests [252]. By varying the classification threshold where cases are differentiated from controls, a series of sensitivity and specificity pairs are obtained by comparing predicted and actual subject phenotypes. The area under an ROC curve (AUROC) is used as a measure of accuracy. Tests that are perfect at differentiating cases and controls have an AUROC of 1.0. When a test is no better at classifying cases and controls than doing so randomly, the AUROC is 0.5. Based on these extremes, a conventional scheme used to classify the predictive accuracy of tests is:

AUROC	Rating
0.5-0.6	Fail
0.6-0.7	Poor
0.7-0.8	Fair
0.8-0.9	Good
0.9-1.0	Excellent

For each predictive model, ROC curves are created by comparing the predicted to the actual phenotypes. Convex hulls are estimated using the Qhull algorithm [253] as implemented in Matlab (The Mathworks, Inc., Natick, MA 01760), and the area under the convex hull is obtained using the trapezoidal rule.

Chapter 9

Asthma Exacerbation

9.1 Overview

Asthma exacerbations, commonly known as asthma attacks, are the major cause of morbidity and mortality in asthma [27, 28, 29]. Exacerbation episodes involve worsening of asthma symptoms, including shortness of breath, cough, wheezing, chest pain or tightness, mucus production, or some combination of these [Section 7.3]. As the primary reason for asthma hospitalizations and emergency room visits, they account for a large portion of asthma healthcare expenses. In 2004, 11.7 million Americans (3.9 million children under 18) had an asthma attack [13]. This comprises 57% of the 20.5 million Americans who are estimated to have asthma. American Lung Association data gathered between 1997 and 2004 consistently show that children 5-17 years old have the highest exacerbation rates [13]. Indeed, asthma is the third leading cause of hospitalizations in US children, occurring an estimated 198,000 times per year [180].

The underlying genetics of asthma exacerbations is unknown. Twin studies have attributed a greater genetic than environmental component to asthma [22, 23, 24, 25]. Because exacerbations occur in some patients with asthma and not others in similar environments, a genetic basis that predisposes individuals to exacerbations is likely. Uncovering the genetic basis underlying asthma exacerbations would be helpful to understand the biology of exacerbations, discover novel therapeutic targets, and identify those at risk of suffering from them. In this work, a genetic predictive model of asthma exacerbations was created with PBN [Section 5] using data from the CAMP trial [Section 8.1].

9.2 Phenotype Definition

A cohort of Caucasian CAMP subjects with available genetic data were selected to create the predictive model of exacerbation. These subjects are not part of the steroid treatment group of CAMP and were followed during CAMPCS. The clinical data used to define exacerbation in these subjects is responses to the questions "How many times have you had an overnight hospitalizion for asthma since the last visit?" and "How many times have you had an emergency room visit for asthma since the last visit?" that were gathered over 10 years during trial visits or some CAMPCS phone interviews. Subjects are classified as cases (i.e. exacerbators) if they have at least one overnight hospitalization and controls (i.e. non-exacerbators) if they do not have emergency room visits or hospitalizations during the observation period. A total of 290 subjects, 83 cases and 207 controls, meet the criteria outlined.

9.3 Model

The genetic data available for the subjects included 2443 SNPs from 350 candidate genes and 399 intergenic loci. All of these SNPs are in Hardy-Weinberg Equilibrium among controls and have minor allele frequencies greater than 0.05. A PBN was learned from the genetic data [Figure 9-1]. In this network, 132 SNPs from 55 genes and 28 intergenic loci are found to be predictive of exacerbation [Table A.3].

9.4 Predictive Accuracy

The model's goodness was assessed using fitted values by predicting exacerbation for each subject used in the PBN construction. The corresponding area under the ROC curve (AUROC) is 0.97. Model robustness was tested by performing a 20-fold cross-validation, in which the original dataset was split into 20 subgroups and each subgroup was used as an independent dataset while the remaining subgroups were used to learn the PBN parameters. The AUROC for the cross-validation procedure was 0.84 [Figure 9-2], which demonstrates that the network has good predictive accuracy.

To compare the performance of the PBN to a single-gene approach, datasets created using information of SNPs from one gene at a time were created. The PBN was used to

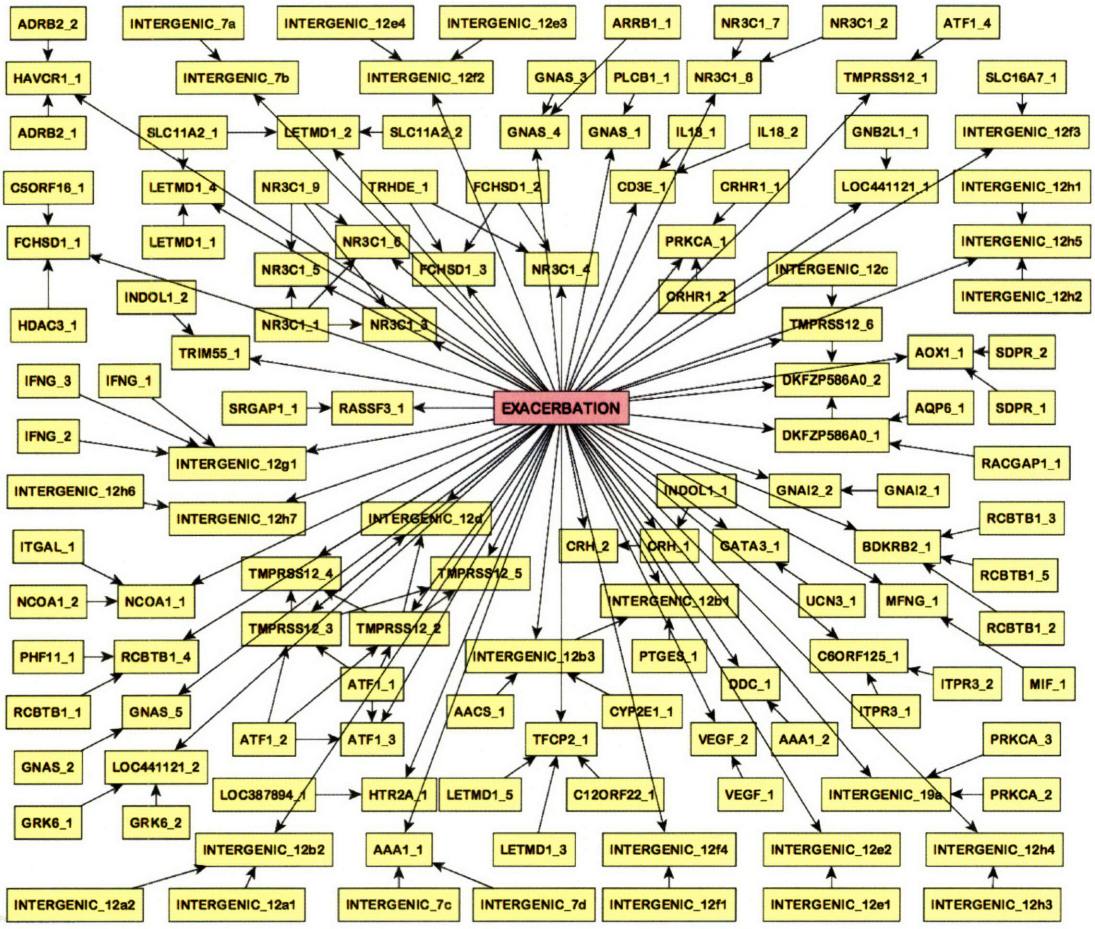


Figure 9-1: Phenocentric Bayesian network of asthma exacerbation. Exacerbation is predicted by 132 SNPs from 55 genes and 28 intergenic loci.

classify cases and controls using this single-gene data, and corresponding AUROC of fitted values were obtained. As Figure 9-3 shows, the predictive accuracy of individual genes is nearly random (0.50) in most cases. Using all of the genes is far better than individual ones.

9.5 Biological Interpretation

Biological pathways corresponding to the genes in the PBN were found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [222]. Additionally, genes in the glucocorticoid and beta-agonist pathways were analyzed because of the importance of these paths in asthma management. Pathways with several represented genes are shown in Table 9.5. Among the genes in the PBN, four have been associated to asthma in previous

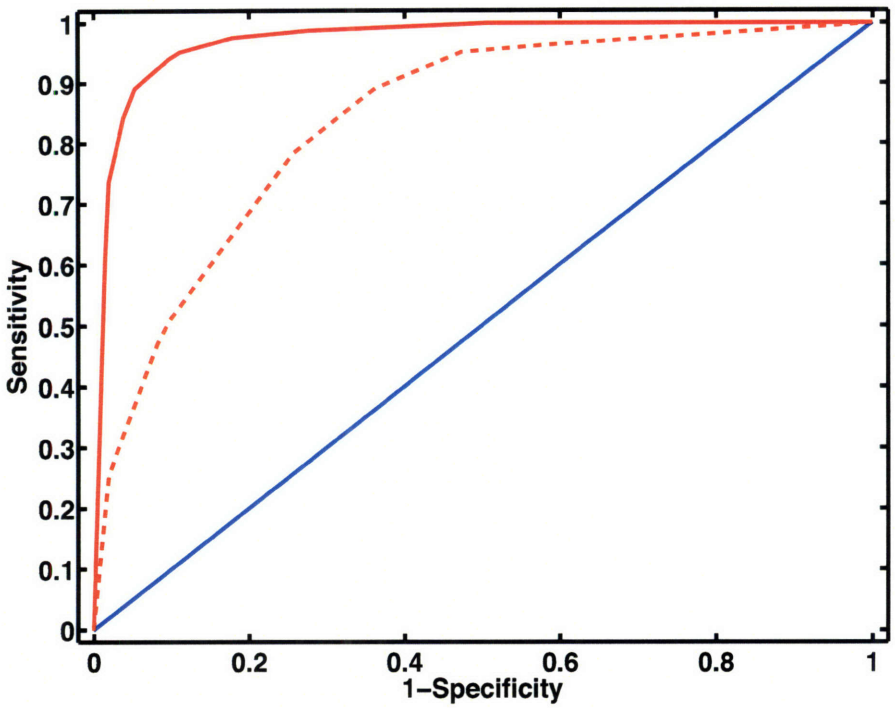


Figure 9-2: Predictive accuracy of exacerbation PBN. The area under the ROC curve corresponding to fitted values is 0.97 (red, solid) and to 20-fold cross-validation is 0.84 (orange, dashed).

studies (IFNG, NR3C1, HAVCR1, ADRB2) and ten to immune processes (PHF11, CD3E, IL18, IFNG, NR3C1, MIF, HAVCR1, PRKCA, ADRB2, BDKRB2).

The G-protein coupled receptor (GPCR) and calcium signaling pathways are two highly represented pathways in the exacerbation PBN [Figure 9-4]. Four GPCRs are represented, each with potential involvement in asthma exacerbation. The serotonin (HTR2A) and bradykinin (BDKRB2) receptors are known to be involved in airway cell contractility [254, 255, 256]. The beta-2 adrenergic receptor (ADRB2) is involved in bronchial smooth muscle dilation [30]. Changes in these three GPCRs could lead to increased exacerbations due to changes in airway contractility. The corticotropin releasing hormone receptor (CRHR1) is involved in glucocorticoid synthesis, which mediates inflammatory responses. Deficiency of CRH or CRHR1 has been shown to lead to inflammatory response changes [257, 243]. In addition to GPCR genes, two ligands (CRH and UCN3) and several downstream effectors (PRKCA, PLCB1, ITPR3, GNAS, GNAI2, ATF1) are represented [222, 258]. Some of the downstream effectors are part of calcium signaling, which is significant because alterations in calcium homeostasis can increase airway smooth muscle contractile responses [259]. Overall,

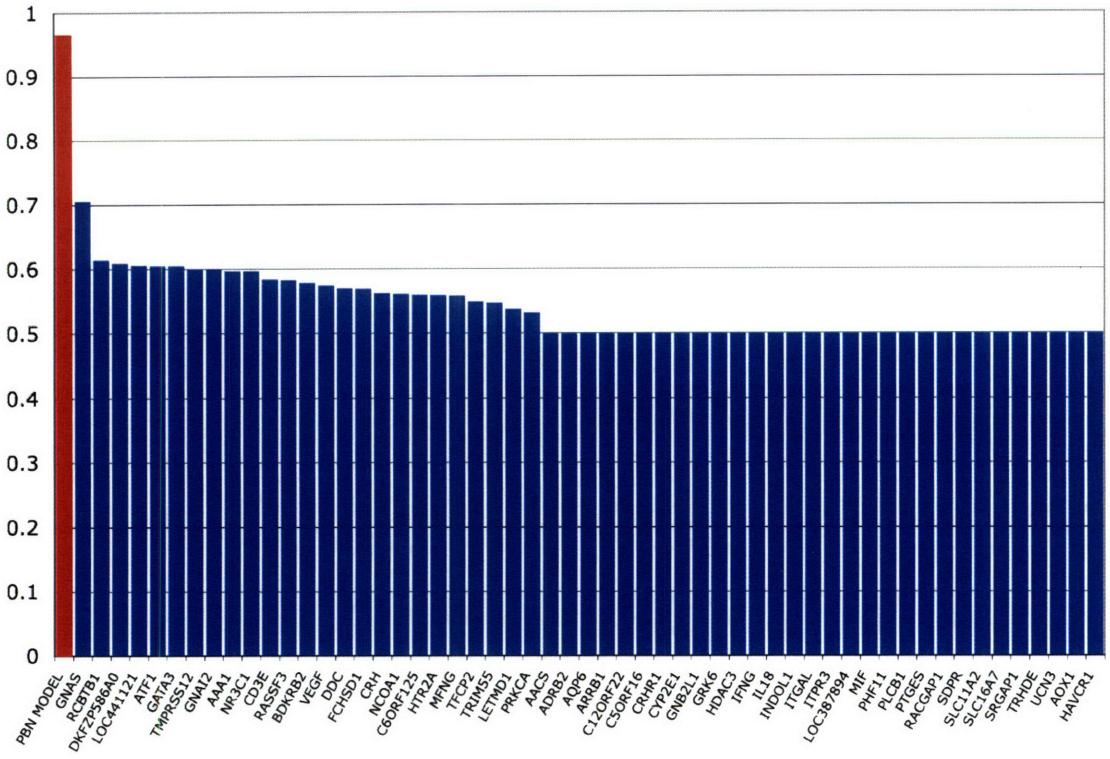


Figure 9-3: Predictive accuracy of single genes in the exacerbation PBN. The area under the ROC curve corresponding to fitted values of a dataset consisting of single gene information shows that the full PBN exacerbation model (red) has far better predictive performance than single genes (blue).

differences in these GPCR pathways could lead to asthma exacerbations through changes in airway cell contractility and/or changes in inflammation.

Another group of highly represented SNPs in the exacerbation PBN are from genes related to steroid and beta-agonist pathways [Figure 9-5]. Both of these pathways are targets for the most common asthma drugs: corticosteroids and beta-agonists [Section 7.2]. The genes involved include receptors for both of these drugs (NR3C1, ADRB2), which are normally activated by endogenous ligands [166, 167]. The beta-two agonist receptor (ADRB2) is also a GPCR shown in Figure 9-4. Differences in steroid and beta-agonist pathways suggests that differential response to endogenous or administered steroids and beta-agonists could lead to increased exacerbations through changes in inflammatory response or bronchodilation. These pathways overlap with the GPCR and calcium signaling pathways.

Pathway Name	Represented Genes
G-protein coupled receptor (GPCR) signaling	CRH, UCN3, HTR2A, CRHR1, ADRB2, BDKRB2, PRKCA, PLCB1 ITPR3, GNAS, GNAI2, ATF1
Calcium signaling pathway	TPR3, GNAS, HTR2A, PRKCA, PLCB1, ADRB2, BDKRB2
Glucocorticoid	NR3C1, NCOA1
Beta-agonist	DDC, ADRB2, GRK6, ARRB1
Phosphatidylinositol signaling system	ITPR3, GRK6, PRKCA, PLCB1
Long-term depression	GNAI2, ITPR3, CRH, GNAS, CRHR1, PRKCA, PLCB1
Gap junction	GNAI2, ITPR3, GNAS, HTR2A, PRKCA, PLCB1
Tyrosine metabolism	MIF, AOX1, DDC
Neuroactive ligand-receptor interaction	UCN3, NR3C1, CRH, HTR2A, CRHR1, ADRB2, BDKRB2

Table 9.1: Biological pathways that are represented by genes in the exacerbation PBN.

9.6 Traditional Association Test Results

The Cochran-Armitage trend test was performed to do a traditional single-SNP association analysis of the genetic markers to exacerbation. This analysis found that 85 SNPs from 44 genes and 23 intergenic loci are individually associated to exacerbation at a $p < 0.05$ significance level [Table A.4]. Association results of this type, routinely reported in the literature, are verified by replication studies. However, translation of this type of results to a clinical setting is slow because they are not useful for individual risk assessment. Of the 44 genes that are significant in the trend test, 19 are in part of the exacerbation PBN: AAA1, ATF1, BDKRB2, DDC, DKFZP586A0, FCHSD1, GNAS, GNB2L1, HAVCR1, HTR2A, LETMD1, LOC441121, MFNG, NCOA1, NR3C1, PLCB1, PRKCA, RASSF3, TMPRSS12. Of the 19 in common with the PBN, 16 were genes whose individual predictive accuracy is higher than 0.50 in Figure 9-3. Therefore, even the genes whose statistical association is significant according to the test, have low individual predictive accuracy. Using the combination of genes in the PBN provides a more robust framework to differentiate exacerbators from non-exacerbators than using single genes.

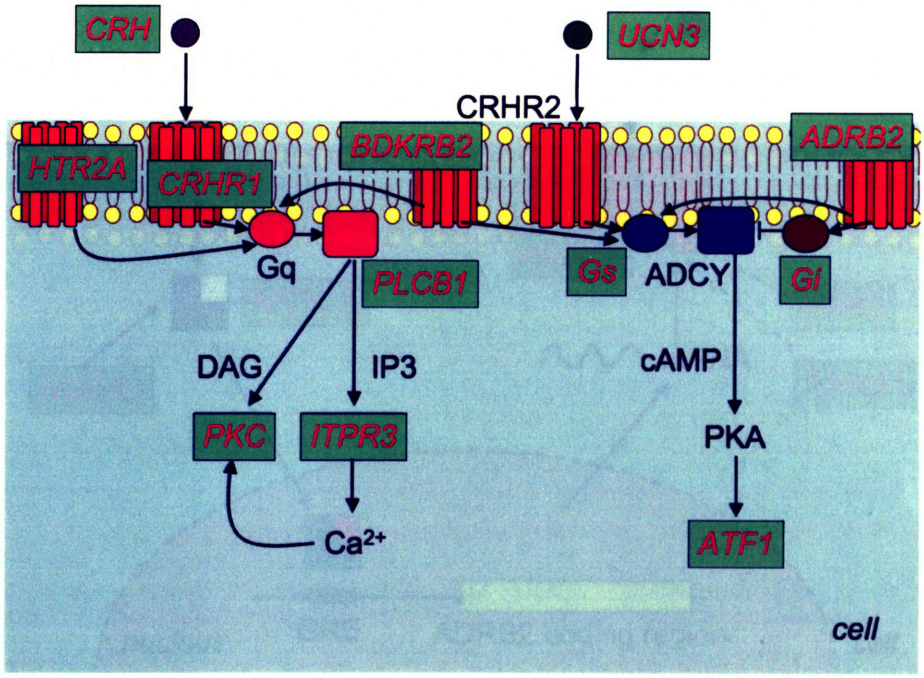


Figure 9-4: G-protein coupled receptor and calcium signaling pathway genes in the exacerbation PBN. Genes that are represented in the PBN are in boxes.

9.7 Comparison to Clinical Model

As an alternative to a genetic test, would a model made of clinical variables be better to predict asthma exacerbations? When a new asthma patient is observed, there is no way to know that the patient will have exacerbations. Once a history of exacerbations is established, then it can usually be assumed that the patient will have future asthma attacks. In clinical practice, it would be most useful to know that a person will be a future exacerbator before a history of attacks has been established. A genetic test could be performed for this purpose. An alternative test using clinical variables would only use data available at a first visit, prior to long observations.

Using the CAMP subjects from which the PBN was learned, clinical variables available at the beginning of the CAMP trial (i.e. *pre-trial*) were analyzed and used to build a stepwise logistic regression model. Characteristics of pre-trial clinical variables for the subjects are shown in Table 9.2. Fisher's exact test on categorical values found that parent(s) smoking and positive skin test(s) are individually associated with exacerbation. The Kruskal-Wallis test found that the distributions of age, height, weight, BMI, eosinophilia, $Ln(PC_{20})$, pre-

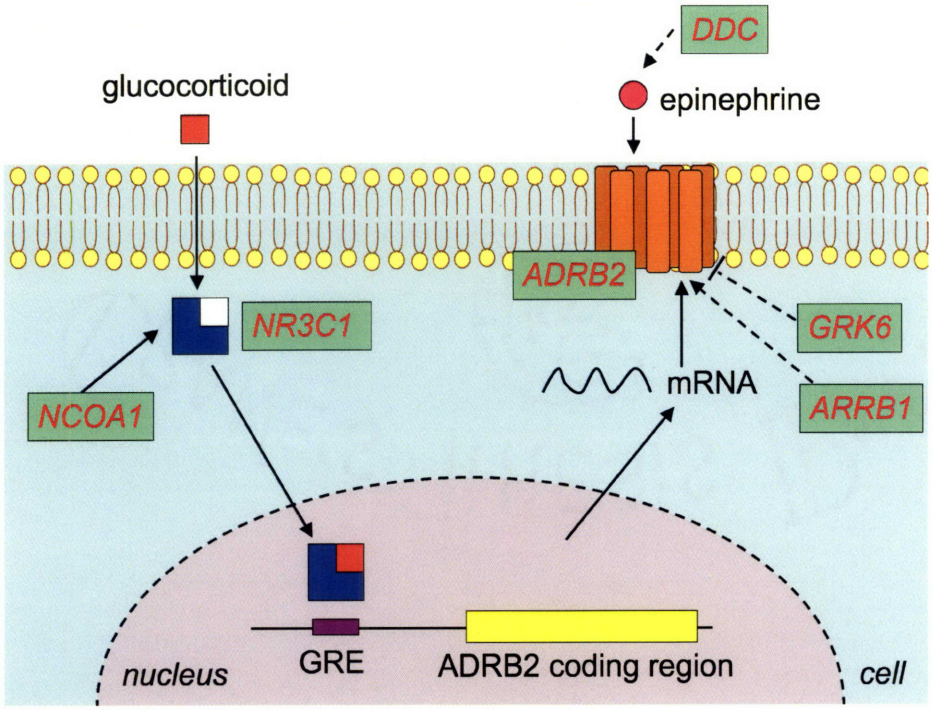


Figure 9-5: Glucocorticoid and beta-agonist pathway genes in exacerbation PBN. Genes that are represented in the PBN are in boxes.

bronchodilator FEV_1, and bronchodilator response measured either as $\Delta FEV_1\%init$ or $\Delta FEV_1\%pred$ are significantly different in cases and controls.

A predictive model of asthma exacerbation was built using a forward stepwise logistic procedure with the pretrial clinical data. This model found that bronchodilator response defined as $\Delta FEV_1\%init$, age, parent(s) smoking, and IgE levels are predictive of exacerbation [Table 9.7]. The goodness of this model compared to the PBN model was assessed by comparing AUROC curves of fitted values of the subjects used in the model construction. As Figure 9-7 shows, the genetic model's AUROC (0.97) is significantly higher than the clinical variable model (0.73), p-value < 0.001 [260].

A second clinical variable model of exacerbation was created with Bayesian networks. The continuous variables were made into categorical ones by splitting each into quartiles, and a Bayesian network was learned using Bayesware Discoverer [12]. In the resulting model, age, BMI, mother smoking while pregnant, bronchodilator response defined as $\Delta FEV_1\%init$, pre-bronchodilator $FEV_1 : FVC$, positive skin test, and IgE levels are predictors of exacerbation [Figure 9-6]. Despite the discretization procedure, the predictive accuracy of the Bayesian network (AUROC of fitted values equal to 0.74) is higher, although

	Total ($N = 290$)	Cases ($N = 83$)	Controls ($N = 207$)	p-value
Gender, no (%)				.1821
Male	179 (61.72)	46 (15.86)	133 (45.86)	
Female	111 (38.28)	37 (12.76)	74 (25.52)	
Age, years	9.1 (2.2)	8.2 (2.2)	9.5 (2.1)	.0001
Age at asthma onset, years	3.2 (2.6)	2.8 (2.2)	3.4 (2.7)	.2113
Height, cm	134 (14)	129 (14)	136 (14)	.0002
Weight, kg	34 (12)	30 (11)	35 (12)	.0001
BMI, kg/cm^2	18 (3)	17 (3)	18 (3)	.0002
Mother smoking while pregnant, no (%)				.0632
Yes	41 (14.21)	17 (5.90)	24 (33)	
No	247 (85.76)	66 (22.92)	181 (62.85)	
One or both parent(s) smoke, no (%)				.0024
Yes	120 (41.38)	46 (15.86)	74 (25.52)	
No	170 (58.62)	37 (12.76)	133 (45.86)	
Positive skin test(s), no (%)				.0468
Yes	246 (84.83)	76 (26.12)	170 (58.62)	
No	44 (15.17)	7 (2.41)	37 (12.76)	
Total serum IgE, $log(IU/mL)$	2.6 (0.7)	2.8 (0.6)	2.5 (0.7)	.0770
Eosinophilia, $log(cells/mm^3)$	2.5 (0.6)	2.6 (0.6)	2.4 (0.6)	.0003
$Ln(PC_{20})$, $ln(mg/mL)$	0.2 (1.2)	-0.04 (1.2)	0.3 (1.2)	.0455
Pre-BD FEV_1, L	1.7 (0.5)	1.5 (0.5)	1.8 (0.5)	.0001
Pre-BD $FEV_1 : FVC$	80 (8)	79 (9)	80 (8)	.3395
Pre-BD $FEV_1\%pred$	95 (14)	95 (16)	96 (13)	.6944
BDR $\Delta FEV_1\%init$	.11 (.11)	.15 (.16)	.09 (.08)	.0008
BDR $\Delta FEV_1\%pred$	9.9 (8.3)	13 (11)	8.5 (6.5)	.0003

Table 9.2: CAMP pre-trial clinical data in exacerbation cases and controls. Mean (standard deviation) are reported unless otherwise noted. Categorical variable p-values are for Fisher's exact tests, continuous variable p-values are for Kruskal-Wallis tests. BD is bronchodilator; BDR is bronchodilator response.

not statistically different (p-value 0.080), than that of the logistic regression [Figure 9-7]. The genetic PBN model has much higher predictive accuracy than the clinical variable Bayesian network model (p-value < 0.001).

In 2003, the CAMP research group published a report in which risk factors for hospitalization of asthma prior to enrollment in the clinical trial were identified [261]. Performing univariate analysis on data for all 1041 children who enrolled in the trial, they found that younger age of asthma onset, longer duration of asthma, greater number of positive allergy skin tests, higher serum IgE level, greater peripheral blood eosinophilia, greater recent inhaled corticosteroid use, greater airway obstruction, greater airway hyperresponsiveness, and lower patient intelligence quotient (IQ) were all associated with prior asthma hospital-

Step	Variable Entered	p-value
1	Bronchodilator response $\Delta FEV_1\%init$	0.0001
2	Age	0.0005
3	One or both parent(s) smoke	0.0016
4	IgE levels	0.0054

Table 9.3: Pre-trial clinical variables selected by forward stepwise regression to model exacerbation.

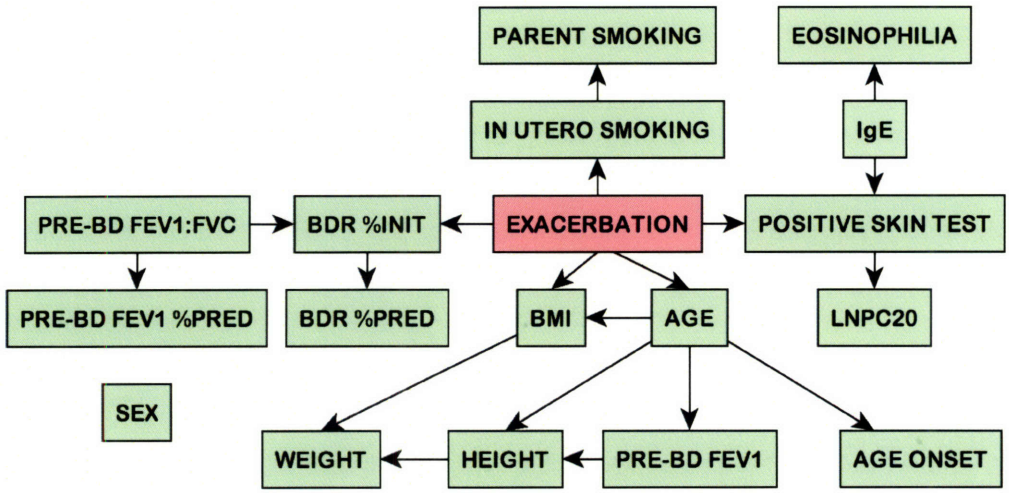


Figure 9-6: CAMP pre-trial clinical variable exacerbation Bayesian network. Age, BMI, mother smoking while pregnant, bronchodilator response defined as $\Delta FEV_1\%init$, pre-bronchodilator $FEV_1 : FVC$, positive skin test, and IgE levels are predictors of exacerbation.

ization. Using all of these variables to construct a forward multivariate logistic regression, they found that younger age of asthma onset, longer duration of asthma, recent use of inhaled corticosteroid, greater airflow obstruction, and lower patient IQ were significant risk factors for prior asthma hospitalization. Another study using CAMP subjects found that $FEV_1\%pred$ was associated to asthma exacerbations defined as subjects with oral steroid use, emergency department visits and/or hospitalizations [262]. Our one-variable-at-a-time analysis results, which were obtained using a subset of all CAMP subjects and data obtained just prior to the clinical trial beginning, are consistent with most of these findings. However, some inconsistencies appear to be present among the studies. It is difficult to compare the studies in terms of what variables are "better" than others to differentiate exacerbators from non-exacerbators because there is no measure of predictive accuracy. What is most useful, especially in a clinical setting, is a model that can be validated with predictive testing in individual subjects. As the stepwise logistic regression model and Bayesian network

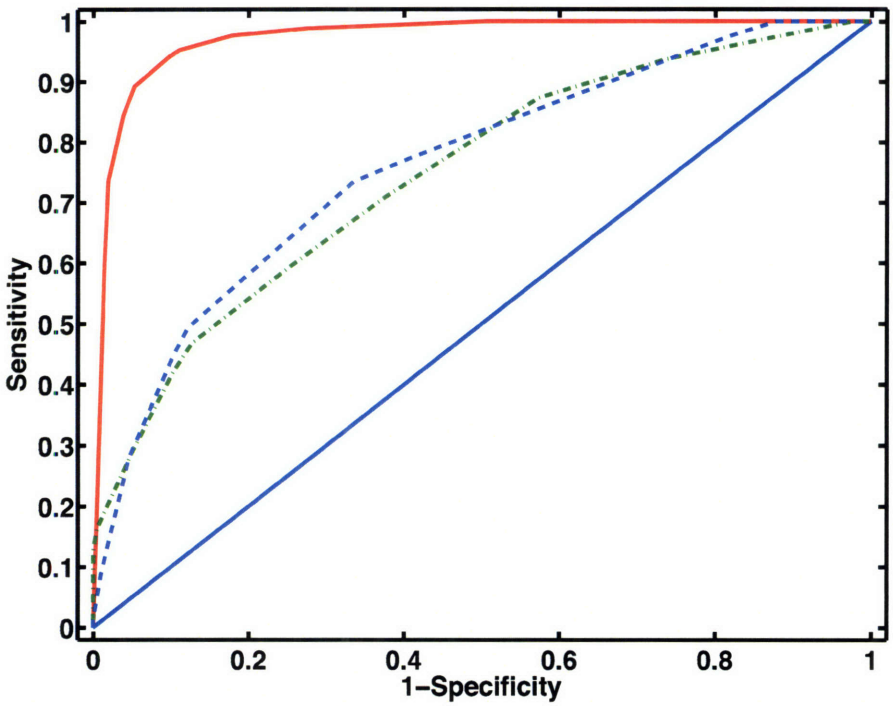


Figure 9-7: Comparison of clinical variable exacerbation forward logistic regression model (FLR; dash-dot, green), Bayesian Network (BN; dash,purple) and genetic exacerbation phenocentric Bayesian network (PBN; solid, red). The genetic model is better than the clinical models as demonstrated by AUROC of fitted values of 0.97 for PBN versus 0.73 for FLR and 0.74 for BN. Curves for PBN are significantly different from clinical model AUROCs (p-value $< .0001$). The BN and FLR AUROC are not significantly different from each other (p-value 0.080).

results demonstrate, the predictive accuracy of clinical variables known at the beginning of the CAMP trial is low. The genetic PBN model is significantly better at predicting exacerbation and would be an easier and more objective test to administer to new patients.

9.8 Conclusion

A successful genetic predictive model of asthma exacerbations was built using PBNs. This model found that 132 out of 2443 SNPs from 55 out of 350 genes and 28 out of 399 intergenic loci are predictive of asthma exacerbation. The model's predictive accuracy is good, as established with fitted values (AUROC 0.97) and a 20-fold cross-validation (AUROC 0.84). The multivariate exacerbation model was compared to a single-gene approach, and the superior predictive accuracy of the multivariate approach was demonstrated. Comparison

of the genetic PBN to a forward logistic regression model and a Bayesian network built with CAMP clinical data shows that the genetic model has better predictive accuracy (AUROC of 0.97 versus 0.73 and 0.74 for fitted values, respectively) and would be a more objective and easily administered clinical test. The genes represented in the exacerbation PBN suggest biological pathways that could be involved in mediating exacerbations via changes in inflammation and airway cell contractility.

Chapter 10

Bronchodilator Response

10.1 Overview

Measurement of bronchodilator response (BDR) is a common clinical test for the evaluation of reversible airway obstruction and the diagnosis of asthma [Section 7.4]. This test is administered to determine whether administration of a bronchodilator medication improves FEV_1. It consists in a pre-bronchodilator measurement of FEV_1, administration a bronchodilator (e.g. albuterol), and a follow-up measurement of FEV_1. In asthmatics compared to non-asthmatics, there tends to be a large change in FEV_1. The physiological response to a bronchodilator is a complex trait, involving intricate interactions among airway epithelial and smooth muscle cells and nerves.

The most potent and rapidly acting bronchodilators currently available for clinical use are β_2-agonists [30]. Their primary effect is to stimulate β_2 receptors on the surface of airway smooth muscle cells, which via an increase in intracellular cyclic AMP levels, relaxes airway smooth muscles and reduces bronchoconstriction. These drugs are the primary drugs used in bronchodilator tests and are routinely used for the pharmacologic management of asthma despite their variable efficacy among patients.

Evidence for the genetic basis of BDR has been established in family aggregation and gene association studies but a comprehensive understanding of this response has not been realized. Unraveling the genetic basis of bronchodilator response would be helpful to identify patient-specific treatments, identify novel therapeutic targets, and help in the diagnosis and monitoring of asthma. Further, a predictive test would help establish what patients are responsive to β_2-agonists and what genetic mechanisms may be responsible for variability

in patient response to such drugs. In this chapter, a genetic predictive model of BDR was created with PBNs [Section 5] using data from the CAMP trial [Section 8.1].

10.2 Phenotype Definition

Bronchodilator response is a more complicated trait to define than asthma exacerbation because it is a continuous trait, with no clear boundary separating *responders* from *non-responders*. Further, a variety of BDR definitions are used by different studies and in different clinical settings. In this work, the definition chosen for BDR is $\Delta FEV_1\%init$. This definition is the most commonly used one in a clinical setting. Because there is no clear boundary to distinguish cases from controls, arbitrary thresholds must be chosen [Section 7.4]. Two extreme regimes of BDR, according to clinical standards, were selected to define cases and controls. By selecting extreme regimes, the genetic signal from each group should be maximal.

A cohort of Caucasian CAMP subjects with available genetic data were selected to create the predictive model of BDR. These subjects are not part of the steroid treatment group of CAMP and have at least 9 out of 11 BDR test data available. Subjects are classified as cases (i.e. responders) if they have a mean BDR of 12% or greater and controls (i.e. non-responders) if they have a mean BDR less than 8%. A total of 308 subjects, 113 cases and 195 controls, meet the criteria outlined.

10.3 Model

Genetic data available for these subjects includes 2444 SNPs from 350 candidate genes and 402 intergenic loci. All of these SNPs are in Hardy-Weinberg Equilibrium among controls and have minor allele frequencies greater than 0.05. A PBN was learned from the genetic data [Figure 10-1]. In this network, 163 SNPs from 55 genes and 22 intergenic loci are found to be predictive of BDR [Table A.5].

10.4 Predictive Accuracy

The model's goodness was assessed using fitted values by predicting BDR for each subject used in the PBN construction. The corresponding area under the ROC curve (AUROC) is

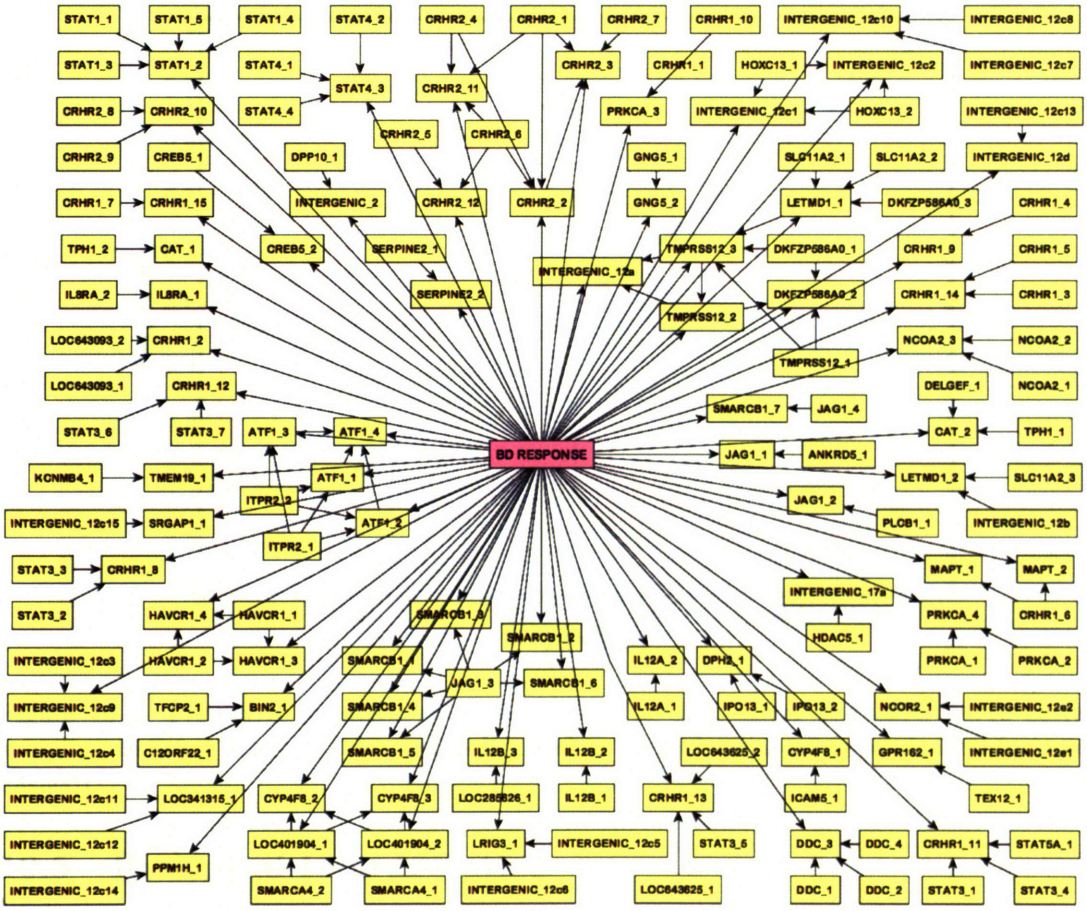


Figure 10-1: Phenocentric Bayesian network of BDR. The BDR is predicted by 163 SNPs from 55 genes and 22 intergenic loci.

0.94. Model robustness was tested by performing a 20-fold cross-validation, in which the original dataset was split into 20 subgroups and each subgroup was used as an independent dataset while the remaining subgroups were used to learn the PBN parameters. The AUROC for the cross-validation procedure was 0.80 [Figure 10-2], which demonstrates that the network has good predictive accuracy.

To compare the performance of the PBN to a single-gene approach, datasets created using information of SNPs from one gene at a time were created. The PBN was used to classify cases and controls using this single-gene data, and corresponding AUROC of fitted values were obtained. As Figure 10-3 shows, the predictive accuracy of individual genes is nearly random (0.50) in most cases, whereas using all of the genes has an AUROC of 0.94.

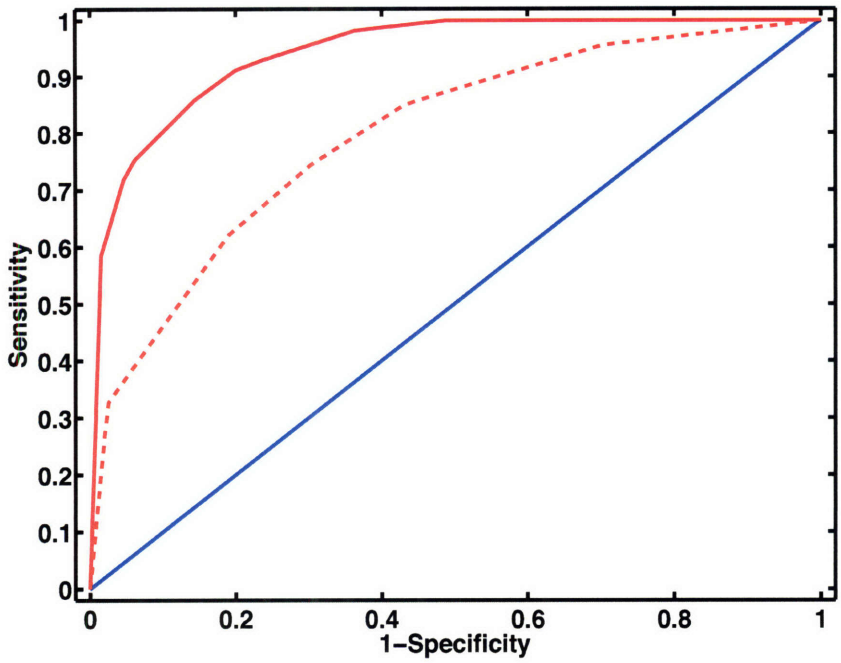


Figure 10-2: Predictive accuracy of BDR PBN. The area under the ROC curve corresponding to fitted values is 0.94 (red, solid) and to 20-fold cross-validation is 0.80 (orange, dashed).

10.5 Biological Interpretation

Biological pathways corresponding to the 55 genes in the PBN were found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [222]. Pathways with several represented genes are shown in Table 10.5. Some of these pathways are related to neuronal signaling, indicating that bronchodilators may be activating these pathways differently in those who respond and do not respond to bronchodilators. Additionally, some of the pathways are cell-signaling pathways, suggesting involvement of a cellular response to bronchodilators. Among the genes in the PBN, three have been associated to asthma in previous studies (DPP10,IL12B,IHAVCR1) and four to immune processes (PRKCA,DPP10,IL8RA,IL12B).

The G-protein coupled receptor (GPCR) and calcium signaling pathways are highly represented in the BDR PBN [Figure 10-4]. In particular, both corticotropin releasing hormone receptors are represented (CRHR1, CRHR2). The corticotropin releasing hormone receptor (CRHR1) is involved in glucocorticoid synthesis, which mediates inflammatory responses. Deficiency of CRH or CRHR1 has been shown to lead to inflammatory response

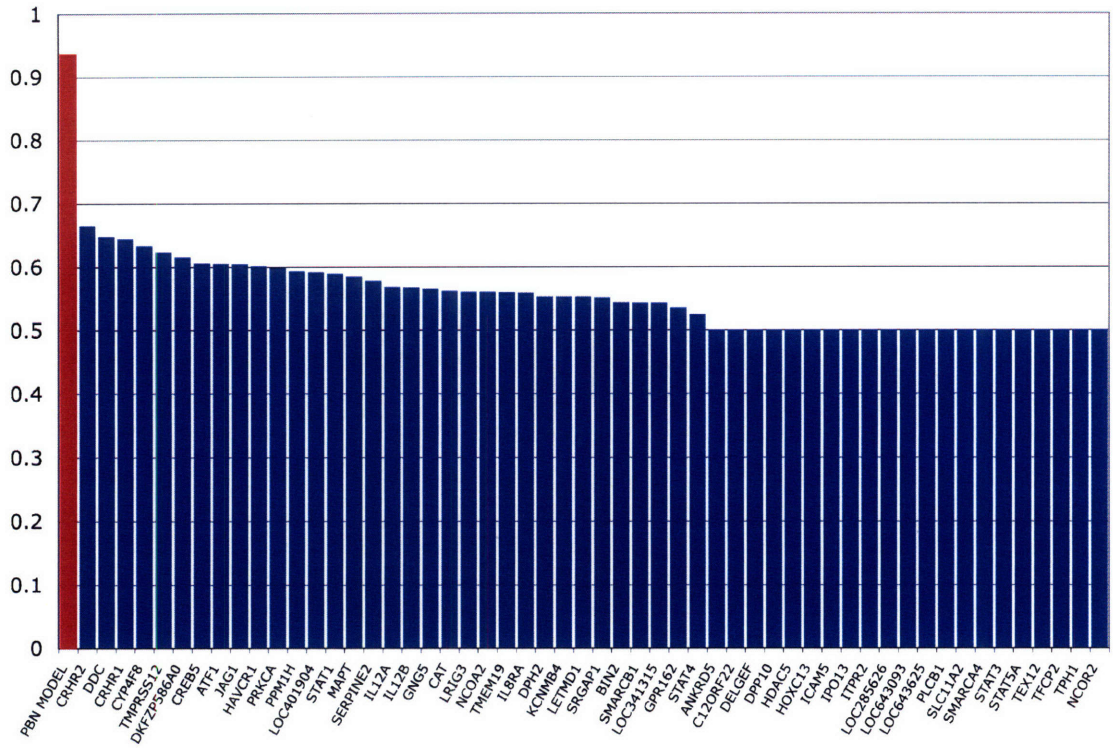


Figure 10-3: Predictive accuracy of single genes in the BDR PBN. The area under the ROC curve corresponding to the fitted values of a dataset consisting of single gene information shows that the full PBN BDR model (red) has far better predictive performance than single genes (blue).

changes [257, 243]. In addition to GPCR genes, several downstream effectors (PRKCA, PLCB1, ITPR2, ATF1, CREB5) are represented [222, 258]. Some of the downstream effectors are part of calcium signaling, which is significant because alterations in calcium homeostasis can increase airway smooth muscle contractile responses [259]. Overall, differences in the CRHR pathways could lead to asthma BDR differences through changes in glucocorticoid synthesis and inflammation.

Some steroid and beta-agonist pathway genes are represented in the BDR PBN [Figure 10-5]. Both of these pathways are targets for the most common asthma drugs: corticosteroids and beta-agonists [Section 7.2]. Intuitively, beta-agonist pathway involvement in BDR differences is expected. However, few of these genes are present in the BDR PBN. Steroid pathway genes are more highly present. As mentioned in Section 7.2, the glucocorticoid and beta-agonist pathways overlap [166, 167]. As glucocorticoid receptors become activated, they can increase translation of beta-2 adrenergic receptors, which can increase

Pathway Name	Represented Genes
G-protein coupled receptor (GPCR) signaling	CRHR1, CRHR2, PLCB1, ITPR2 PRKCA, CREB5, ATF1
Calcium signaling	PRKCA, PLCB1, ITPR2
Glucocorticoid	NCOA1, CREB5, STAT5A
Beta-agonist	DDC
JAK-STAT signaling pathway	STAT3, IL12B, STAT5A, STAT4, STAT1, IL12A
Phosphatidylinositol signaling system	ITPR2, PRKCA, PLCB1
TOLL-like receptor signaling pathway	IL12B, STAT1, IL12A
Long-term depression	CRHR1, PRKCA, ITPR2, PLCB1
Long-term potentiation	PRKCA, ITPR2, PLCB1
Gap junction	PRKCA, ITPR2, PLCB1
Tryptophan metabolism	TPH1, CYP4F8, DDC, CAT

Table 10.1: KEGG pathways that are represented by genes in the BDR PBN.

the effect of beta-2 agonists. Therefore, the represented glucocorticoid pathway genes, including the CRHR genes, could modulate BDR by influencing the beta-agonist pathway or by changing inflammatory response.

The JAK-STAT signaling pathway is highly represented in the BDR PBN, and unlike the pathways discussed above, was not present in the exacerbation PBN [Figure 10-6]. JAK-STAT signaling, which is used by immune cells, is involved in asthma by changing the phenotype of cells to be TH2-like [263]. Two cytokines (IL12A, IL12B) that activate cytokine receptors and four STATs (STAT1, STAT3, STAT4, STAT5A) that convey cytokine signals are in the BDR PBN. Changes in inflammation mediated by JAK-STAT signaling could alter BDR.

10.6 Traditional Association Test Results

Traditional single-SNP association was performed using the Cochran-Armitage trend test. This analysis found that 108 SNPs from 47 genes and 12 intergenic loci are individually associated to BDR at a $p < 0.05$ significance level [Table A.6]. Association results of this type, routinely reported in the literature, are verified by replication studies. Translation of this type of result to a clinical setting is slow because it does not useful for individual subject evaluation. Of the 47 genes that are significant in the trend test, 21 are part of the BDR PBN: CREB5, CRHR1, CRHR2, CYP4F8, DDC, DKFZP586A0, GPR162, HAVCR1, HDAC5, IL12A, IL12B, ITPR2, JAG1, MAPT, NCOA2, PPM1H, PRKCA, SERPINE2,

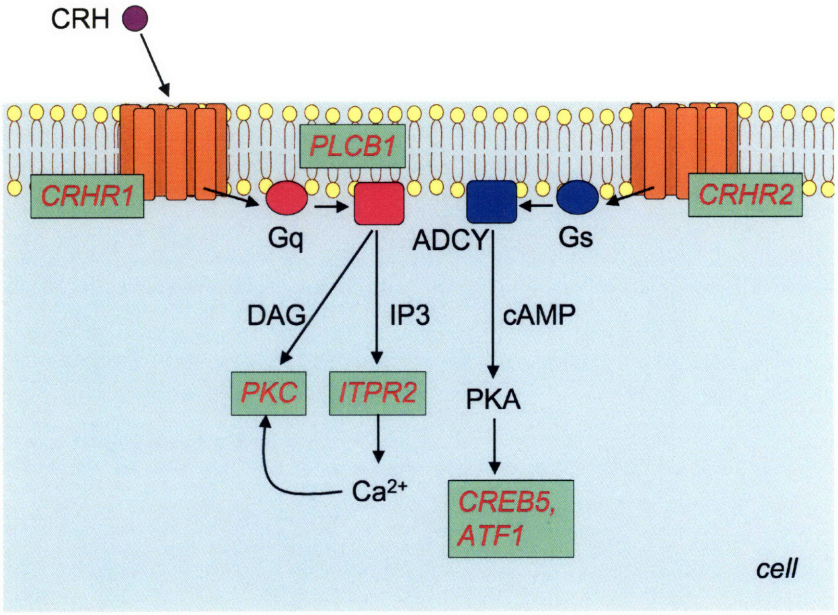


Figure 10-4: G-protein coupled receptor pathway genes in the BDR PBN. Genes that are represented in the PBN are in boxes.

SMARCB1, SRGAP1, STAT3. Of the 21 in common with the PBN, 18 are genes whose individual predictive accuracy is higher than 0.50 in Figure 10-3. Therefore, even the genes whose statistical association was significant according to the test, have low individual predictive accuracy. Using the combination of genes together in the PBN provides a more robust framework to differentiate BDR cases and controls.

10.7 Conclusion

A successful genetic predictive model of BDR was made using PBN. This model found that 163 out of 2444 SNPs from 55 out of 350 genes and 22 out or 402 intergenic loci are predictive of BDR. The predictive accuracy of the model is good, as established with fitted values (AUROC 0.94), and a 20-fold cross-validation (AUROC 0.80). The multivariate model was compared to a single-gene approach, and the superior predictive accuracy of the multivariate approach is demonstrated. Genes represented in the BDR PBN suggest that biological pathways that modulate BDR include those involved in glucocorticoid synthesis and action, which may lead to changes in inflammation and/or the beta-agonist pathway.

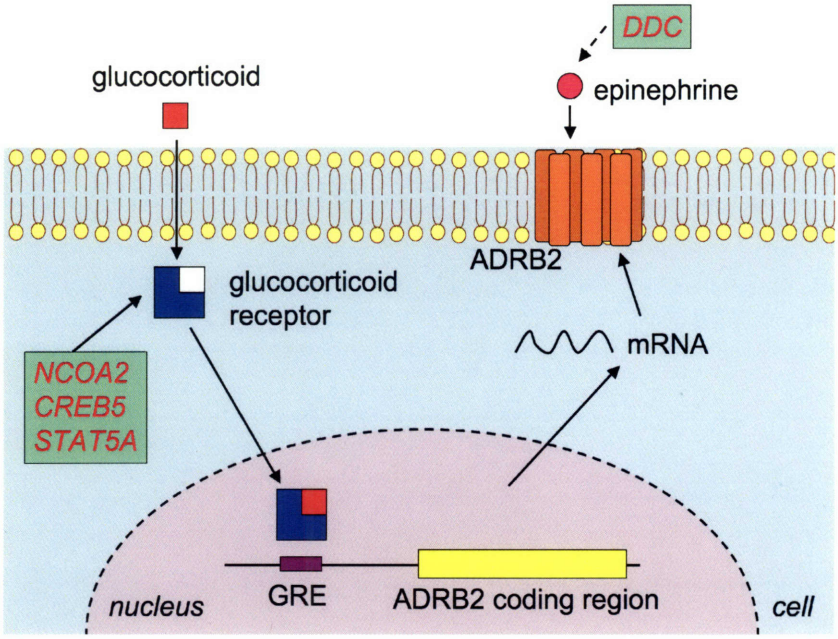


Figure 10-5: Glucocorticoid and beta-agonist pathway genes in BDR PBN. Genes that are represented in the PBN are in boxes.

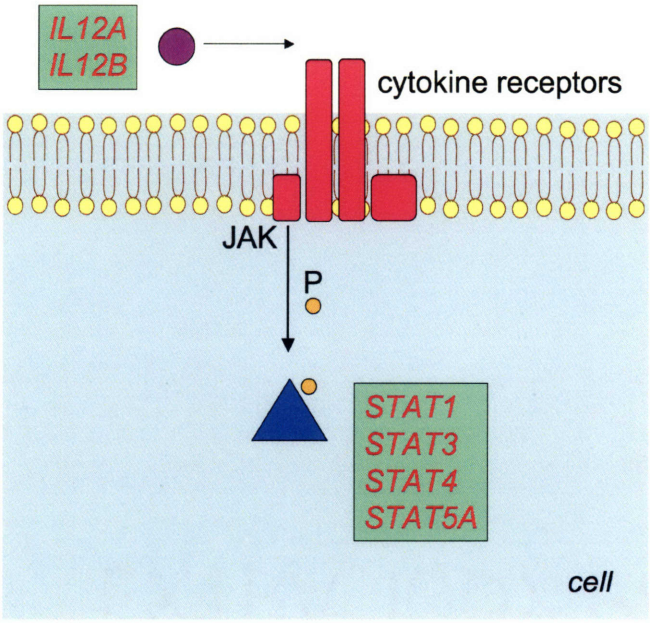


Figure 10-6: JAK-STAT signaling pathway genes in BDR PBN. Genes that are represented in the PBN are in boxes.

Chapter 11

Conclusion

11.1 Summary

In this thesis, the need for better analytic methods for the study of complex traits is addressed with the development of phenocentric Bayesian networks. The procedure is tailored for the discovery of multivariate models with large genomic datasets and can be used to predict outcomes of a phenotype of interest. It focuses on learning probabilistic relationships that best predict outcomes of a variable of interest, which is a suitable approach for gene association studies where the goal is to successfully predict a trait given a set of genetic markers (i.e. SNPs) The procedure is as powerful as other PGM to uncover complex dependencies among many variables, with the advantage that it can be used with large genetic datasets.

The utility of phenocentric Bayesian networks is demonstrated with the creation of predictive models for two complex traits related to asthma management: exacerbation and bronchodilator response. Successful genetic predictive models were built for each of these traits. The exacerbation model, utilizing 133 out of 2443 SNPs from 55 out of 350 genes and 28 out or 399 intergenic loci, has good predictive accuracy (fitted value AUROC 0.97, 20-fold cross-validation AUROC 0.84). The BDR model found that 164 out of 2444 SNPs from 55 out of 350 genes and 22 out or 402 intergenic loci are predictive of BDR. The predictive accuracy of this model is also good (fitted value AUROC 0.94, 20-fold cross-validation AUROC 0.80). The models obtained suggest biological pathways that could be involved in exacerbation and BDR. Both models are shown to be superior than single gene analysis, emphasizing the need for methods that consider complex dependencies among

variables and demonstrating that phenocentric Bayesian networks are a useful approach to study the genetic architecture of complex traits.

11.2 Future Directions

The exacerbation and bronchodilator response models created have been shown to have good predictive accuracy using fitted values and cross-validation. However, in order for the predictive models to be verified as generally applicable, they must be tested in independent populations. These models can be readily tested with new genetic data when it becomes available. Verification of their accuracy in other populations will provide certainty that the results are useful in a clinical setting. Further, verification will increase the likelihood that biological testing of hypotheses suggested by genes in the networks will provide valuable insights into the mechanisms underlying these traits.

The recent advent of genomewide association (GWA) studies with over 500,000 markers is promising for uncovering the genetic architecture of complex traits. However, without more powerful multivariate methods that can scale to these large datasets, whose size eclipses earlier candidate gene studies, the promise of fully understanding the genetic underpinnings of complex traits will likely not be fulfilled. PBNs are a promising approach for the analysis of these large datasets. Testing PBNs with GWA datasets will be necessary to verify their ability to create useful predictive models at a genomewide scale. If necessary, minor modifications to the algorithm's scoring procedure can easily be implemented to adjust the efficiency of the algorithm. Tailoring this method and other PGMs for the study of complex traits is the most promising approach for modeling traits that accounts for their complex genetic underpinnings and has a quantitative metric to assess their predictive accuracy for individuals.

Appendix A

Tables

Table A.1: **CAMP candidate genes**

Gene Name	SNPs Genotyped
AAA1	20
AACS	3
ACCN2	5
ACVR1B	1
ACVRL1	5
ADAM19	1
ADAM33	14
ADCY6	10
ADCY7	5
ADCY9	24
ADCYAP1	2
ADCYAP1R1	6
ADRB2	70
ALDH7A1	8
ALOX15	8
AMHR2	1
ANKRD33	1
ANKRD5	14
AOX1	20
APOB48R	1
AQP2	4
AQP6	5
ARG1	6
ARHGAP9	3
ARRB1	17
ARRB2	2
ATF1	6
ATF7	1
ATP2A2	4
ATP5G2	2

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
ATP6V0B	1
AVIL	6
BCL2L10	1
BDKRB1	6
BDKRB2	21
BIN2	7
BRD7	1
BXDC5	4
C11ORF72	1
C12ORF22	11
C14ORF166B	1
C1QL1	1
C4ORF9	2
C5	2
C5ORF16	1
C6ORF125	2
C9ORF26	11
CACNB3	2
CALCOCO1	1
CAPS2	2
CAT	7
CCDC93	1
CCDC97	2
CCL11	10
CCL17	11
CCR5	11
CD3E	1
CD4	10
CDK4	2
CEBPA	12
CENTG1	4
CHIT1	21
CHRM2	4
CHRM3	9
CLN3	1
COL2A1	11
CPAMD8	3
CPM	10
CPSF6	5
CREB1	8
CREB3L1	7
CREB3L2	1
CREB5	52
CREBBP	2
CREBL2	1

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
CREM	19
CRH	3
CRHBP	7
CRHR1	65
CRHR2	73
CRP	4
CSK	5
CSRP2	1
CST3	8
CTDSP2	4
CTLA4	4
CTSO	2
CX3CR1	3
CYP27B1	1
CYP2A13	7
CYP2C9	12
CYP2E1	12
CYP3A4	3
CYP3A5	4
CYP4F3	1
CYP4F8	3
DAZAP2	2
DBP	2
DCAL1	1
DCD	1
DCTN2	2
DDC	11
DDIT3	1
DEFB1	4
DELGEF	3
DERL3	3
DIAPH1	1
DKFZP547K0	1
DKFZP586A0	8
DKFZP586D0	2
DKFZP761L1	1
DNAJC14	1
DPH2	6
DPP10	18
E2F7	3
ECEL1	1
EGR1	4
ELA1	1
ELF2	6
ELF5	3

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
F2R	7
F2RL1	7
F2RL3	1
FAM19A2	1
FAM98C	2
FBN2	16
FCER2	18
FCHSD1	3
FKBP4	4
FKBP5	19
FLG	2
FLJ12355	4
FLJ20489	1
FLJ21125	8
FLJ21908	1
FLJ23436	2
FLJ45983	4
GAL	6
GALNT6	2
GATA3	4
GFRA4	7
GLS	4
GLS2	2
GLYCAM1	1
GMEB1	9
GNAI1	14
GNAI2	8
GNAI3	18
GNAS	8
GNAT2	3
GNB1	20
GNB1L	16
GNB2L1	5
GNB3	9
GNB5	24
GNG5	6
GNG7	8
GNS	4
GOSR2	1
GPR162	2
GRB2	2
GRIP1	1
GRK4	9
GRK5	13
GRK6	6

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
GRK7	6
GSTP1	5
HAT1	8
HAVCR1	20
HDAC1	1
HDAC2	6
HDAC3	1
HDAC5	2
HDAC7A	3
HELB	1
HLA-DQA1	1
HLA-G	3
HMGA2	1
HNMT	1
HOXC13	2
HS322B1A	1
HSP90AA1	4
HTR2A	22
HUMCYT2A	4
ICAM1	2
ICAM4	2
ICAM5	4
IFNG	17
IGFBP6	1
IKBKAP	2
IL10	6
IL12A	4
IL12B	12
IL12RB1	7
IL12RB2	22
IL13	6
IL18	10
IL18BP	7
IL22	2
IL26	2
IL27	1
IL4	7
IL4R	16
IL8	2
IL8RA	6
IL8RB	2
IMP5	2
INDO	7
INDOL1	2
INSIG2	1

Continued on Next Page...

Table A.1 - Continued

Gene Name	SNPs Genotyped
IPO13	15
IRAK3	7
ITFG2	1
ITGAL	17
ITGB7	3
ITPR1	36
ITPR2	51
ITPR3	25
JAG1	17
K5B	2
KCNC2	6
KCNH2	1
KCNMB4	2
KDR	2
KIAA0141	1
KIAA1755	1
KIF5A	1
KITLG	3
KRT18	1
KRT3	2
KRT4	3
KRT6IRS	1
KRT6L	1
KRT8	3
KRTHB5	1
KYNU	14
LAG3	1
LEMD3	1
LEPREL2	1
LETMD1	20
LGR5	4
LOC253264	1
LOC283400	2
LOC283403	1
LOC284890	4
LOC285626	8
LOC341315	2
LOC387894	4
LOC390338	1
LOC390342	3
LOC400050	2
LOC400568	2
LOC400620	1
LOC401904	2
LOC440098	1

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
LOC440399	1
LOC440552	1
LOC440591	1
LOC440806	1
LOC440935	4
LOC441121	12
LOC441346	1
LOC441641	6
LOC57228	2
LOC643093	4
LOC643231	1
LOC643489	1
LOC643559	2
LOC643625	3
LOC643645	1
LOC643788	13
LOC643865	2
LOC643878	2
LOC644733	1
LOC645253	1
LOC645495	1
LOC645507	10
LOC645623	2
LOC645738	2
LOC646067	2
LOC647071	16
LOC653518	1
LOX	4
LRIG3	1
LTA	1
LYZ	5
MAPK1	5
MAPK3	1
MAPT	2
MARS	2
MAST3	1
MBD6	2
MDM2	1
MED11	1
METTL1	1
MFNG	1
MGC4093	2
MGP	3
MIF	4
MMP12	16

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
MMP19	3
MRPL36	1
MS4A2	24
MSH3	1
MYO5C	2
NAB2	2
NADK	1
NAP1L1	3
NCOA1	26
NCOA2	33
NCOR2	4
NDFIP1	4
NFATC4	12
NONE	1
NOS3	1
NPFF	1
NPSR1	16
NR0B2	1
NR1I2	19
NR3C1	23
NUMA1	2
OR10P1	2
OR56B4	1
OR6C1	1
OR6C68	2
OS-9	7
OSBPL8	1
PCAF	27
PCDH12	10
PDGFB	7
PELP1	1
PERQ1	1
PHF11	10
PHLDA1	1
PIP5K2C	7
PLCB1	18
PLCB2	3
PLCB3	2
PLCB4	3
PLN	3
POMC	6
POU6F1	8
PPARG	4
PPM1H	5
PRKCA	42

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
PRR7	2
PTAFR	2
PTGDR	14
PTGES	16
PTGIR	2
PTPRB	2
PTPRR	2
QTRT1	1
RAB3IP	6
RAC2	4
RACGAP1	1
RAPGEF3	6
RARG	1
RASGRP4	3
RASSF3	3
RBMS2	1
RCBTB1	16
RGS12	2
RGS16	2
RHBDF2	5
ROBO1	1
RYR1	2
S100A10	9
SCARB1	4
SDPR	9
SEMA3B	1
SENP1	1
SERPINA6	3
SERPINE2	26
SETDB2	7
SGOL1	2
SLC11A2	59
SLC12A2	1
SLC16A7	2
SLC26A10	3
SLC4A8	2
SLC6A4	13
SMARCA4	17
SMARCB1	15
SMARCE1	2
SOAT2	1
SPATA1	1
SPHK2	2
SPINK5	1
SRGAP1	3

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
STAT1	27
STAT2	2
STAT3	20
STAT4	22
STAT5A	1
STAT5B	1
STAT6	4
SYCE1	1
TACR1	13
TBCA	1
TBK1	1
TBKBP1	1
TBX1	1
TBX21	16
TDO2	12
TEX12	2
TFCP2	25
TGFB1	3
TLR10	10
TLR4	5
TMEM106C	2
TMEM132B	7
TMEM142A	6
TMEM19	1
TMEM5	1
TMPRSS12	8
TNF	3
TPH1	19
TPH2	17
TRAPPC5	1
TRHDE	34
TRIM41	1
TRIM55	3
TSFM	2
TSPAN31	2
UBC	2
UBE2O	2
UCN3	4
URP2	1
USP5	1
USP52	2
VDR	7
VEGF	17
VMD2L3	4
WARS2	6

Continued on Next Page...

Table A.1 – Continued

Gene Name	SNPs Genotyped
WDFY3	1
XRCC5	3
XRCC6BP1	4
YEATS4	9
ZNF659	1
ZNF740	1

Table A.2: **KEGG pathways that are represented by CAMP candidate genes.**

Pathway Name	**Represented Genes**
Calcium signaling pathway	SPHK2, GNAS, PLN, PLCB2, PRKCA, F2R, BDKRB2, NOS3, HTR2A, CHRM2, CHRM3, BDKRB1, TACR1, ADCY7, PTAFR, PLCB1, ATP2A2, ITPR3, ADCY9, ITPR1, PLCB4, PLCB3, ADRB2, ITPR2, RYR1
Cytokine-Cytokine receptor interaction	CCR5, IL12B, IL4, CX3CR1, VEGF, IL4R, ACVR1B, IL10, KDR, LTA, PDGFB, IL8RA, IL12A, IL13, IL26, IFNG, IL12RB2, IL12RB1, CCL17, TNF, IL18, AMHR2, CCL11, KITLG, IL22, IL8RB, IL8, TGFB1
Fc Epsilon Ri signaling pathway	TNF, IL13, IL4, MS4A2, PRKCA, MAPK3, RAC2, MAPK1, GRB2
Gap junction	GNAI2, GNAI1, GNAI3, ADCY6, GNAS, PLCB2, PRKCA, MAPK3, PLCB1, MAPK1,ITPR3, ADCY9, ITPR1, PDGFB, HTR2A,PLCB4, PLCB3, ITPR2, ADCY7, GRB2
Inositol phosphate metabolism	GRK4, PIP5K2C, IRAK3, GRK6, GRK7, PLCB4, PLCB2, PLCB3, PLCB1, GRK5
JAK-STAT signaling pathway	IL13, IFNG, IL26, IL12B, IL12RB2, IL12RB1, IL4, STAT4, CREBBP, STAT5B, IL4R, STAT6, IL10, STAT3, STAT5A, IL22, STAT1, STAT2, IL12A, GRB2
Linoleic acid metabolism	CYP3A4, CYP2C9, CYP2E1, ALOX15, CYP3A5
Long-term depression	GNAI2, GNAI1, GNAI3, GNAS, CRHR1,PLCB2, PRKCA, PLCB1, MAPK3, MAPK1,ITPR3, CRH, NOS3, ITPR1, PLCB4,PLCB3, ITPR2, RYR1
Long-term potentiation	ITPR3, ITPR1, RAPGEF3, CREBBP, PLCB4, PLCB2, PRKCA, PLCB3, MAPK3, ITPR2, PLCB1, MAPK1
MAPK signaling pathway	MAPT, PTPRR, DDIT3, ARRB1, PRKCA, MAPK3, ACVR1B, RASGRP4, MAPK1,TNF, CACNB3,

Continued on Next Page...

Table A.2 – Continued

Pathway Name	Represented Genes
	PDGFB, NFATC4, RAC2, TGFB1, ARRB2, GRB2
Natural killer cell mediated cytotoxicity	TNF, IFNG, ITGAL, NFATC4, HLA-G, PRKCA, MAPK3, RAC2, MAPK1, GRB2, ICAM1
Neuroactive ligand-receptor interaction	POMC, PTAFR, UCN3, CRHR2, CRHR1, GAL, F2R, BDKRB2, NPFF, PTGDR, ADCYAP1, CRH, NR3C1, CHRM2, HTR2A, F2RL3, ADCYAP1R1, CHRM3, BDKRB1, TACR1, F2RL1, ADRB2, PTGIR
Nicotinate and nicotinamide metabolism	GRK4, IRAK3, GRK6, AOX1, GRK7, NADK, GRK5
Notch signaling pathway	HDAC2, MFNG, CREBBP, HDAC1, NCOR2, JAG1
Phosphatidylinositol signaling system	GRK4, IRAK3, GRK6, PLCB2, PRKCA, PLCB1, ITPR3, PIP5K2C, ITPR1, GRK7, PLCB4, PLCB3, ITPR2, GRK5
Regulation of actin cytoskeleton	ITGB7, ITGAL, MAPK3, F2R, BDKRB2, MAPK1, PIP5K2C, CSK, PDGFB, CHRM2, CHRM3, BDKRB1, RAC2, DIAPH1
T cell receptor signaling pathway	IL10, CD3E, TNF, CD4, IFNG, IL4, NFATC4, CTLA4, CDK4, GRB2
TGF-Beta signaling pathway	TNF, AMHR2, IFNG, ACVRL1, CREBBP, MAPK3, TGFB1, ACVR1B, MAPK1
TOLL-like receptor signaling pathway	TNF, IL12B, TBK1, IL8, TLR4, STAT1, RAC2, IL12A
Tryptophan metabolism	TDO2, WARS2, TPH1, CYP4F8, INDO, KYNU, AOX1, DDC, TPH2, ALDH7A1, CAT

Table A.3: **Exacerbation PBN genes.** The AUROC of fitted values predicting exacerbation using information for each of 55 genes used in Figure 9-3 is shown. An additional 28 intergenic SNPs are in the PBN.

Gene Name	SNPs in PBN	AUROC
AAA1	2	0.60
AACS	1	0.50
ADRB2	2	0.50
AOX1	1	0.47
AQP6	1	0.50
ARRB1	1	0.50
ATF1	4	0.60
BDKRB2	1	0.58
C12ORF22	1	0.50
C5ORF16	1	0.50
C6ORF125	1	0.56
CD3E	1	0.58
CRH	2	0.56
CRHR1	2	0.50
CYP2E1	1	0.50
DDC	1	0.57
DKFZP586A0	2	0.61
FCHSD1	3	0.57
GATA3	1	0.60
GNAI2	2	0.60
GNAS	5	0.71
GNB2L1	1	0.50
GRK6	2	0.50
HAVCR1	1	0.44
HDAC3	1	0.50
HTR2A	1	0.56
IFNG	3	0.50
IL18	2	0.50
INDOL1	2	0.50
ITGAL	1	0.50
ITPR3	2	0.50
LETMD1	5	0.54
LOC387894	1	0.50
LOC441121	2	0.61
MFNG	1	0.56
MIF	1	0.50
NCOA1	2	0.56
NR3C1	9	0.60
PHF11	1	0.50
PLCB1	1	0.50
PRKCA	3	0.53

Continued on Next Page...

Table A.3 – Continued

Gene Name	SNPs in PBN	AUROC
PTGES	1	0.50
RACGAP1	1	0.50
RASSF3	1	0.58
RCBTB1	5	0.61
SDPR	2	0.50
SLC11A2	2	0.50
SLC16A7	1	0.50
SRGAP1	1	0.50
TFCP2	1	0.55
TMPRSS12	6	0.60
TRHDE	1	0.50
TRIM55	1	0.55
UCN3	1	0.50
VEGF	2	0.57

Table A.4: **Exacerbation Trend Test.** SNPs associated to exacerbation at the $p < 0.05$ significance level are shown. SNPs are labeled with their gene name followed by a counter if more than one SNP from the same gene is in the table.

SNP	p – value
INTERGENIC.1	0.0002
ATF1	0.0018
RASSF3	0.0020
GNAS.1	0.0025
INTERGENIC.2	0.0029
NR3C1.1	0.0034
TMPRSS12.1	0.0036
INTERGENIC.3	0.0039
FCHSD1.1	0.0054
INTERGENIC.4	0.0054
NR3C1.2	0.0073
INTERGENIC.5	0.0079
FCHSD1.2	0.0079
BDKRB2	0.0081
INTERGENIC.6	0.0087
TMEM132B	0.0087
DKFZP586A0	0.0088
INTERGENIC.7	0.0090
GRK7	0.0102
INTERGENIC.8	0.0110
INTERGENIC.9	0.0122
GNAS.2	0.0131
IL12RB2	0.0141
GMEB1.1	0.0179
INTERGENIC.10	0.0179
CST3	0.0179
PRKCA	0.0179
PLCB1	0.0184
CREM.1	0.0188
NCOA1	0.0189
HTR2A	0.0190
ITPR1	0.0200
INTERGENIC.11	0.0201
KRTHB5	0.0208
CREM.2	0.0212
MFNG	0.0213
DPP10	0.0213
HAVCR1	0.0219
LETMD1.1	0.0220
GRK5.1	0.0221
LETMD1.2	0.0221

Continued on Next Page...

Table A.4 – Continued

SNP	p **– value**
GMEB1.2	0.0223
INTERGENIC.12	0.0223
INTERGENIC.13	0.0229
ARRB1	0.0233
WARS2	0.0233
INTERGENIC.14	0.0240
GMEB1.3	0.0255
C6ORF125	0.0256
GNB5.1	0.0262
INTERGENIC.15	0.0270
INTERGENIC.16	0.0270
ITPR3.1	0.0278
ITPR3.2	0.0278
TMPRSS12.2	0.0300
INTERGENIC.17	0.0303
ADCYAP1	0.0319
FCHSD1.3	0.0328
LOC645253	0.0332
CREM.3	0.0337
CRP	0.0357
LEMD3	0.0370
GRK5.2	0.0389
INTERGENIC.18	0.0390
TMPRSS12.3	0.0397
CAT.1	0.0397
GNB2L1	0.0399
GNB5.2	0.0410
LOC441121.1	0.0417
LOC441121.2	0.0429
CREM.4	0.0434
PTGDR	0.0434
CREM.5	0.0434
INTERGENIC.19	0.0436
INTERGENIC.20	0.0437
INTERGENIC.21	0.0442
DDC	0.0448
GNG7	0.0448
INTERGENIC.22	0.0458
PCAF	0.0471
HAT1	0.0481
AAA1	0.0483
INTERGENIC.23	0.0486
ITPR2	0.0497
CAT.2	0.0498

Table A.5: **BDR PBN genes.** The AUROC of fitted values predicting BDR using the information for each of 54 genes used in Figure 10-3 is shown. An additional 22 intergenic SNPs are in the PBN.

Gene Name	SNPs in PBN	AUROC
ANKRD5	1	0.50
ATF1	4	0.60
BIN2	1	0.54
C12ORF22	1	0.50
CAT	2	0.56
CREB5	2	0.61
CRHR1	15	0.64
CRHR2	12	0.67
CYP4F8	3	0.63
DDC	4	0.65
DELGEF	1	0.50
DKFZP586A0	3	0.62
DPH2	1	0.55
DPP10	1	0.50
GNG5	2	0.57
GPR162	1	0.53
HAVCR1	4	0.60
HDAC5	1	0.50
HOXC13	2	0.50
ICAM5	1	0.50
IL12A	2	0.57
IL12B	3	0.57
IL8RA	2	0.56
IPO13	2	0.50
ITPR2	2	0.50
JAG1	4	0.60
KCNMB4	1	0.55
LETMD1	2	0.55
LOC285626	1	0.50
LOC341315	1	0.54
LOC401904	2	0.59
LOC643093	2	0.50
LOC643625	2	0.50
LRIG3	1	0.56
MAPT	2	0.58
NCOA2	3	0.56
NCOR2	1	0.41
PLCB1	1	0.50
PPM1H	1	0.59
PRKCA	4	0.60
SERPINE2	2	0.58

Continued on Next Page...

Table A.5 – Continued

Gene Name	SNPs in PBN	AUROC
SLC11A2	3	0.50
SMARCA4	2	0.50
SMARCB1	7	0.54
SRGAP1	1	0.55
STAT1	5	0.59
STAT3	7	0.50
STAT4	4	0.52
STAT5A	1	0.50
TEX12	1	0.50
TFCP2	1	0.50
TMEM19	1	0.56
TMPRSS12	3	0.62
TPH1	2	0.50

Table A.6: **BDR Trend Test.** SNPs associated to exacerbation at the $p < 0.05$ significance level are shown. SNPs are labeled with their gene name followed by a counter if more than one SNP from the same gene is in the table.

SNP	$p-$value
LOC401904	0.000204
DDC	0.000407
CYP4F8.1	0.000536
JAG1.1	0.001969
PPM1H	0.002285
INTERGENIC.1	0.002575
CRHR1.1	0.00315
INTERGENIC.2	0.003622
CYP4F8.2	0.004672
CREB5	0.005198
GPR162	0.005335
TMEM19	0.005889
INTERGENIC.3	0.00589
INTERGENIC.4	0.00589
PRKCA.1	0.006161
CRHR2.1	0.006253
INTERGENIC.5	0.006998
PRKCA.2	0.00761
INTERGENIC.6	0.008065
CRHR1.2	0.008991
CRHR1.3	0.009568
CRHR1.4	0.009588
CRHR1.5	0.009716
CRHR2.2	0.009777
MAPT	0.010656
IL12B	0.010796
SMARCB1.1	0.011313
SMARCB1.2	0.011313
SMARCB1.3	0.011313
SMARCB1.4	0.011313
NAP1L1	0.011352
IL12A.1	0.012199
CRHR1.6	0.012337
GNAS.1	0.013494
CRHR1.7	0.013995
NCOA2.1	0.015033
CRHR1.8	0.015237
HAVCR1.1	0.015656
CRHR1.9	0.016026
PTGDR	0.016711
TPH2	0.016881

Continued on Next Page...

Table A.6 – Continued

SNP	p – value
CRHR1.10	0.017133
IL12A.2	0.019154
IL12RB2.1	0.019278
NPSR1	0.019583
CRHR1.11	0.020196
HAVCR1.2	0.020758
INTERGENIC.7	0.020763
SMARCB1.5	0.020848
SMARCB1.6	0.020848
SMARCB1.7	0.020848
GNAS.2	0.021429
ADCY9.1	0.023504
GNB5	0.023603
NCOA2.2	0.023865
PRKCA.3	0.025101
NCOA2.3	0.02546
CRHR1.12	0.026421
SRGAP1	0.027309
PRKCA.4	0.027927
HAVCR1.3	0.028029
SDPR	0.02805
INTERGENIC.8	0.028321
HAVCR1.4	0.028718
HTR2A	0.028814
ADCY9.2	0.029418
NCOA2.4	0.029457
NCOA2.5	0.029457
NCOA2.6	0.029457
CRHR1.13	0.030757
NCOA2.7	0.030925
SMARCB1.8	0.03143
TRHDE	0.031469
PRKCA.5	0.031724
INTERGENIC.9	0.032235
GSTP1	0.032252
IL12RB2.2	0.032258
ITPR2.1	0.032698
INTERGENIC.10	0.032954
DERL3	0.03548
SERPINE2	0.036034
DKFZP586A0	0.036261
NCOA2.8	0.036844
NCOA2.9	0.036844
NCOA2.10	0.036844
GNB1L	0.036863

Continued on Next Page...

Table A.6 – Continued

SNP	p **– value**
NCOA2.11	0.037023
IL4R	0.03754
GALNT6	0.037825
HAVCR1.5	0.038089
ITPR2.2	0.038416
OS	0.039674
IMP5.1	0.039929
NCOA2.12	0.039988
GNAI1	0.041676
ARRB1	0.041879
HDAC5	0.042981
IMP5.2	0.043874
INTERGENIC.11	0.044381
INTERGENIC.12	0.044804
IL18BP	0.046656
NCOA2.13	0.046987
STAT3	0.047067
JAG1.2	0.048085
CYP2A13	0.048344
LOC645253	0.049131
LOC390338	0.049455
PRKCA.6	0.049689

Bibliography

[1] Lander, ES, Schork, NJ. Genetic dissection of complex traits. *Science*, 265(5181):2037–48, 1994.

[2] Ioannidis, JP, Ntzani, EE, Trikalinos, TA, Contopoulos-Ioannidis, DG. Replication validity of genetic association studies. *Nat Genet*, 29(3):306–9, 2001.

[3] Dahlman, I, Eaves, IA, Kosoy, R, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet*, 30(2):149–50, 2002.

[4] Freely associating. *Nat Genet*, 22(1):1–2, 1999.

[5] Strohman, R. Maneuvering in the complex path from genotype to phenotype. *Science*, 296(5568):701–3, 2002.

[6] Holtzman, NA. Putting the search for genes in perspective. *Int J Health Serv*, 31(2):445–61, 2001.

[7] Farrall, M, Morris, AP. Gearing up for genome-wide gene-association studies. *Hum Mol Genet*, 14 Spec No. 2:R157–62, 2005.

[8] Verzilli, CJ, Stallard, N, Whittaker, JC. Bayesian graphical models for genomewide association studies. *Am J Hum Genet*, 79(1):100–12, 2006.

[9] Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

[10] Jansen, R, Yu, H, Greenbaum, D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–53, 2003.

[11] Lauritzen, S, Sheehan, N. Graphical models for genetic analysis. *Stat Sci*, 18:489–514, 2003.

[12] Sebastiani, P, Ramoni, MF, Nolan, V, Baldwin, CT, Steinberg, MH. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet*, 37(4):435–40, 2005.

[13] American Lung Association. Trends in asthma morbidity and mortality. Technical report, Epidemiology and Statistics Unit, Research and Program Services, American Lung Association, 2006.

[14] Global Initiative for Asthma Management and Prevention. NHLBI/WHO workshop report. Technical report, US Department of Health and Human Services. National Institutes of Health, Bethesda, 1995.

[15] Mannino, DM, Homa, DM, Pertowski, CA, et al. Surveillance for asthma–United States, 1960-1995. *MMWR CDC Surveill Summ*, 47(1):1–27, 1998.

[16] Eder, W, Ege, MJ, von Mutius, E. The asthma epidemic. *N Engl J Med*, 355(21):2226–35, 2006.

[17] Osborne, ML, Vollmer, WM, Pedula, KL, Wilkins, J, Buist, AS, O'Hollaren, M. Lack of correlation of symptoms with specialist-assessed long-term asthma severity. *Chest*, 115(1):85–91, 1999.

[18] Ng, TP. Validity of symptom and clinical measures of asthma severity for primary outpatient assessment of adult asthma. *Br J Gen Pract*, 50(450):7–12, 2000.

[19] Nieminen, MM, Kaprio, J, Koskenvuo, M. A population-based study of bronchial asthma in adult twin pairs. *Chest*, 100(1):70–5, 1991.

[20] Laitinen, T, Rasanen, M, Kaprio, J, Koskenvuo, M, Laitinen, LA. Importance of genetic factors in adolescent asthma: a population-based twin-family study. *Am J Respir Crit Care Med*, 157(4 Pt 1):1073–8, 1998.

[21] Duffy, DL, Martin, NG, Battistutta, D, Hopper, JL, Mathews, JD. Genetics of asthma and hay fever in Australian twins. *Am Rev Respir Dis*, 142(6 Pt 1):1351–8, 1990.

[22] Harris, JR, Magnus, P, Samuelsen, SO, Tambs, K. No evidence for effects of family environment on asthma. a retrospective study of Norwegian twins. *Am J Respir Crit Care Med*, 156(1):43–9, 1997.

[23] Lichtenstein, P, Svartengren, M. Genes, environments, and sex: factors of importance in atopic diseases in 7-9-year-old Swedish twins. *Allergy*, 52(11):1079–86, 1997.

[24] Koeppen-Schomerus, G, Stevenson, J, Plomin, R. Genes and environment in asthma: a study of 4 year old twins. *Arch Dis Child*, 85(5):398–400, 2001.

[25] Skadhauge, LR, Christensen, K, Kyvik, KO, Sigsgaard, T. Genetic and environmental influence on asthma: a population-based study of 11,688 Danish twin pairs. *Eur Respir J*, 13(1):8–14, 1999.

[26] Ober, C, Hoffjan, S. Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun*, 7(2):95–100, 2006.

[27] Lane, S, Molina, J, Plusa, T. An international observational prospective study to determine the cost of asthma exacerbations (COAX). *Respir Med*, 100(3):434–50, 2006.

[28] Andersson, F, Borg, S, Stahl, E. The impact of exacerbations on the asthmatic patient's preference scores. *J Asthma*, 40(6):615–23, 2003.

[29] Skrepnek, GH, Skrepnek, SV. Epidemiology, clinical and economic burden, and natural history of chronic obstructive pulmonary disease and asthma. *Am J Manag Care*, 10(5 Suppl):S129–38, 2004.

[30] Nelson, HS. Beta-adrenergic bronchodilators. *N Engl J Med*, 333(8):499–506, 1995.

[31] Botstein, D, Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 33 Suppl:228–37, 2003.

[32] Ott, J. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, 3rd edition, 1999.

[33] Risch, N, Merikangas, K. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 1996.

[34] Cardon, LR, Bell, JI. Association study designs for complex diseases. *Nat Rev Genet*, 2(2):91–9, 2001.

[35] Lander, ES, Linton, LM, Birren, B, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[36] Venter, JC, Adams, MD, Myers, EW, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[37] National Asthma Education Program Expert Panel. guidelines for the diagnosis and management of asthma: Update on selected topics 2002. Technical report, National Heart, Lung, and Blood Institute, 2003.

[38] Weiss, KM, Terwilliger, JD. How many diseases does it take to map a gene with SNPs? *Nat Genet*, 26(2):151–7, 2000.

[39] Colhoun, HM, McKeigue, PM, Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet*, 361(9360):865–72, 2003.

[40] Morton, NE, Collins, A. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A*, 95(19):11389–93, 1998.

[41] Sasieni, PD. From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–61, 1997.

[42] Balding, DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–91, 2006.

[43] Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386, 1955.

[44] Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[45] Schwarz, G. Estimating the dimension of a model. *Ann Stat*, 6:461–464, 1978.

[46] Peduzzi, P, Concato, J, Kemper, E, Holford, TR, Feinstein, AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*, 49(12):1373–9, 1996.

[47] Lohmueller, KE, Pearce, CL, Pike, M, Lander, ES, Hirschhorn, JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33(2):177–82, 2003.

[48] Sillanpaa, MJ, Auranen, K. Replication in genetic studies of complex traits. *Ann Hum Genet*, 68(Pt 6):646–57, 2004.

[49] Zondervan, KT, Cardon, LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet*, 5(2):89–100, 2004.

[50] Setakis, E, Stirnadel, H, Balding, DJ. Logistic regression protects against population structure in genetic association studies. *Genome Res*, 16(2):290–6, 2006.

[51] Price, AL, Patterson, NJ, Plenge, RM, Weinblatt, ME, Shadick, NA, Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006.

[52] Egger, M, Davey Smith, G, Schneider, M, Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–34, 1997.

[53] Reich, DE, Lander, ES. On the allelic spectrum of human disease. *Trends Genet*, 17(9):502–10, 2001.

[54] Pritchard, JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–37, 2001.

[55] Cohen, JC, Kiss, RS, Pertsemlidis, A, Marcel, YL, McPherson, R, Hobbs, HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869–72, 2004.

[56] Moore, JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered*, 56(1-3):73–82, 2003.

[57] Carlborg, O, Haley, CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet*, 5(8):618–25, 2004.

[58] Lauritzen, S, Spiegelhalter, D. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J Roy Stat Soc Ser B*, 50:157–224, 1988.

[59] Pearl, J. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[60] Jordan, M, editor. *Learning in Graphical Models.* MIT Press, Cambridge, MA, 1998.

[61] Cowell, R, Dawid, A, Lauritzen, S, Spiegelhalter, D. *Probabilistic Networks and Expert Systems.* Springer, New York, NY, 1999.

[62] Jensen, F. *Bayesian Networks and Decision Graphs.* Springer, New York, NY, 2001.

[63] Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 77(2):257–286, 1989.

[64] Durbin, R, Eddy, S, Krogh, A, Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK, 1998.

[65] Churchill, GA. Stochastic models for heterogeneous DNA sequences. *Bull Math Biol*, 51(1):79–94, 1989.

[66] Churchill, GA, Lazareva, B. Bayesian restoration of a hidden Markov chain with applications to DNA sequencing. *J Comput Biol*, 6(2):261–77, 1999.

[67] Fan, R, Knapp, M. Genome association studies of complex diseases by case-control designs. *Am J Hum Genet*, 72(4):850–68, 2003.

[68] Schaid, DJ. Evaluating associations of haplotypes with traits. *Genet Epidemiol*, 27(4):348–64, 2004.

[69] Daly, MJ, Rioux, JD, Schaffner, SF, Hudson, TJ, Lander, ES. High-resolution haplotype structure in the human genome. *Nat Genet*, 29(2):229–32, 2001.

[70] The International HapMap Project. *Nature*, 426(6968):789–96, 2003.

[71] Clark, AG. The role of haplotypes in candidate gene studies. *Genet Epidemiol*, 27(4):321–33, 2004.

[72] Greenspan, G, Geiger, D. High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics*, 20 Suppl 1:I137–I144, 2004.

[73] McPeek, MS, Strahs, A. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet*, 65(3):858–75, 1999.

[74] Morris, AP, Whittaker, JC, Balding, DJ. Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am J Hum Genet*, 67(1):155–69, 2000.

[75] Browning, SR. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet*, 78(6):903–13, 2006.

[76] Thomas, A, Camp, NJ. Graphical modeling of the joint distribution of alleles at associated loci. *Am J Hum Genet*, 74(6):1088–101, 2004.

[77] Thomas, A. Characterizing allelic associations from unphased diploid data by graphical modeling. *Genet Epidemiol*, 29(1):23–35, 2005.

[78] Eronen, L, Geerts, F, Toivonen, H. A Markov chain approach to reconstruction of long haplotypes. *Pac Symp Biocomput*, pages 104–15, 2004.

[79] Stephens, M, Smith, NJ, Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, 2001.

[80] Zhang, Y, Niu, T, Liu, JS. A coalescence-guided hierarchical Bayesian method for haplotype inference. *Am J Hum Genet*, 79(2):313–22, 2006.

[81] Geiger, D, Heckerman, D. A characterization of the dirichlet distribution through global and local parameter independence. *Ann Stat*, 25(3):1344–1369, 1997.

[82] Wilks, S. *Mathematical Statistics*. John Wiley & Sons, Inc., New York, 1963.

[83] Cooper, GF, Herskovits, E. A bayesian method for the induction of probabilistic networks from data. *Mach Learn*, 9(4):309–347, 1992.

[84] Yannakakis, M. Computing the minimum fill-in is NP-complete. *SIAM J Alg Disc Meth*, 2:77–79, 1981.

[85] Kjaerulff, U. Optimal decomposition of probabilistic networks by simulated annealing. *Stat Comp*, 2:7–17, 1992.

[86] Amestoy, P, Davis, T, Duff, I. An approximate minimum degree ordering algorithm. *SIAM J Matrix Anal Appl*, 17:886–905, 1996.

[87] Dawid, A. Applications of a general propagation algorithm for probabilistic expert systems. *Statist Comput*, 2:25–36, 1992.

[88] Beaumont, MA, Rannala, B. The Bayesian revolution in genetics. *Nat Rev Genet*, 5(4):251–61, 2004.

[89] Rodin, AS, Boerwinkle, E. Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). *Bioinformatics*, 21(15):3273–8, 2005.

[90] Hoh, J, Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet*, 4(9):701–9, 2003.

[91] Risch, NJ. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–56, 2000.

[92] Jewell, R. *Statistics for Epidemiology*. CRC/Chapman and Hall, Boca Raton, 2003.

[93] Maraganore, DM, de Andrade, M, Lesnick, TG, et al. High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet*, 77(5):685–93, 2005.

[94] Klein, RJ, Zeiss, C, Chew, EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–9, 2005.

[95] Frayling, TM, Timpson, NJ, Weedon, MN, et al. A common variant in the FTO gene is associated with Body Mass Index and predisposes to childhood and adult obesity. *Science*, 2007.

[96] Masoli, M, Fabian, D, Holt, S, Beasley, R. The global burden of asthma: executive summary of the GINA Dissemination Committee report. *Allergy*, 59(5):469–78, 2004.

[97] Weiss, KB, Gergen, PJ, Wagener, DK. Breathing better or wheezing worse? the changing epidemiology of asthma morbidity and mortality. *Annu Rev Public Health*, 14:491–513, 1993.

[98] Weiss, KB, Gergen, PJ, Crain, EF. Inner-city asthma. the epidemiology of an emerging US public health concern. *Chest*, 101(6 Suppl):362S–367S, 1992.

[99] Mannino, DM, Homa, DM, Akinbami, LJ, Moorman, JE, Gwynn, C, Redd, SC. Surveillance for asthma–United States, 1980-1999. *MMWR Surveill Summ*, 51(1):1–13, 2002.

[100] Braman, SS, Kaemmerlen, JT, Davis, SM. Asthma in the elderly. a comparison between patients with recently acquired and long-standing disease. *Am Rev Respir Dis*, 143(2):336–40, 1991.

[101] Barbee, RA, Dodge, R, Lebowitz, ML, Burrows, B. The epidemiology of asthma. *Chest*, 87(1 Suppl):21S–25S, 1985.

[102] Fergusson, DM, Horwood, LJ, Shannon, FT. Parental asthma, parental eczema and asthma and eczema in early childhood. *J Chronic Dis*, 36(7):517–24, 1983.

[103] Peat, JK, Britton, WJ, Salome, CM, Woolcock, AJ. Bronchial hyperresponsiveness in two populations of Australian schoolchildren. ii. relative importance of associated factors. *Clin Allergy*, 17(4):283–90, 1987.

[104] Herxheimer, H, Schaefer, O. Letter: Asthma in Canadian Eskimos. *N Engl J Med*, 291(26):1419, 1974.

[105] Kromann, N, Green, A. Epidemiological studies in the Upernavik district, Greenland. incidence of some chronic diseases 1950-1974. *Acta Med Scand*, 208(5):401–6, 1980.

[106] Crapo, RO. Pulmonary-function testing. *N Engl J Med*, 331(1):25–30, 1994.

[107] Knudson, RJ, Lebowitz, MD, Holberg, CJ, Burrows, B. Changes in the normal maximal expiratory flow-volume curve with growth and aging. *Am Rev Respir Dis*, 127(6):725–34, 1983.

[108] Enright, PL, Lebowitz, MD, Cockroft, DW. Physiologic measures: pulmonary function tests. asthma outcome. *Am J Respir Crit Care Med*, 149(2 Pt 2):S9–18; discussion S19–20, 1994.

[109] Platts-Mills, TA. How environment affects patients with allergic disease: indoor allergens and asthma. *Ann Allergy*, 72(4):381–4, 1994.

[110] Lau, S, Illi, S, Sommerfeld, C, et al. Early exposure to house-dust mite and cat allergens and development of childhood asthma: a cohort study. Multicentre Allergy Study Group. *Lancet*, 356(9239):1392–7, 2000.

[111] Cullinan, P, MacNeill, SJ, Harris, JM, et al. Early allergen exposure, skin prick responses, and atopic wheeze at age 5 in English children: a cohort study. *Thorax*, 59(10):855–61, 2004.

[112] Magnussen, H, Jorres, R, Nowak, D. Effect of air pollution on the prevalence of asthma and allergy: lessons from the German reunification. *Thorax*, 48(9):879–81, 1993.

[113] von Mutius, E, Martinez, FD, Fritzsch, C, Nicolai, T, Roell, G, Thiemann, HH. Prevalence of asthma and atopy in two areas of West and East Germany. *Am J Respir Crit Care Med*, 149(2 Pt 1):358–64, 1994.

[114] Dockery, DW, Speizer, FE, Stram, DO, Ware, JH, Spengler, JD, Ferris, B. G., J. Effects of inhalable particles on respiratory health of children. *Am Rev Respir Dis*, 139(3):587–94, 1989.

[115] Tatum, AJ, Shapiro, GG. The effects of outdoor air pollution and tobacco smoke on asthma. *Immunol Allergy Clin North Am*, 25(1):15–30, 2005.

[116] Wardlaw, AJ. The role of air pollution in asthma. *Clin Exp Allergy*, 23(2):81–96, 1993.

[117] Barnes, PJ. Air pollution and asthma. *Postgrad Med J*, 70(823):319–25, 1994.

[118] Nicholson, KG, Kent, J, Ireland, DC. Respiratory viruses and exacerbations of asthma in adults. *BMJ*, 307(6910):982–6, 1993.

[119] Johnston, SL, Pattemore, PK, Sanderson, G, et al. Community study of role of viral infections in exacerbations of asthma in 9-11 year old children. *BMJ*, 310(6989):1225–9, 1995.

[120] Corne, JM, Marshall, C, Smith, S, et al. Frequency, severity, and duration of rhinovirus infections in asthmatic and non-asthmatic individuals: a longitudinal cohort study. *Lancet*, 359(9309):831–4, 2002.

[121] Illi, S, von Mutius, E, Lau, S, et al. Early childhood infectious diseases and the development of asthma up to school age: a birth cohort study. *BMJ*, 322(7283):390–5, 2001.

[122] Martinez, FD. Role of viral infections in the inception of asthma and allergies during childhood: could they be protective? *Thorax*, 49(12):1189–91, 1994.

[123] Strachan, DP, Butland, BK, Anderson, HR. Incidence and prognosis of asthma and wheezing illness from early childhood to age 33 in a national British cohort. *BMJ*, 312(7040):1195–9, 1996.

[124] Leuenberger, P, Schwartz, J, Ackermann-Liebrich, U, et al. Passive smoking exposure in adults and chronic respiratory symptoms (sapaldia study). Swiss Study on Air Pollution and Lung Diseases in Adults, SAPALDIA Team. *Am J Respir Crit Care Med*, 150(5 Pt 1):1222–8, 1994.

[125] Weitzman, M, Gortmaker, S, Walker, DK, Sobol, A. Maternal smoking and childhood asthma. *Pediatrics*, 85(4):505–11, 1990.

[126] Martinez, FD, Cline, M, Burrows, B. Increased incidence of asthma in children of smoking mothers. *Pediatrics*, 89(1):21–6, 1992.

[127] Cunningham, J, O'Connor, GT, Dockery, DW, Speizer, FE. Environmental tobacco smoke, wheezing, and asthma in children in 24 communities. *Am J Respir Crit Care Med*, 153(1):218–24, 1996.

[128] Ehrlich, RI, Du Toit, D, Jordaan, E, et al. Risk factors for childhood asthma and wheezing. importance of maternal and household smoking. *Am J Respir Crit Care Med*, 154(3 Pt 1):681–8, 1996.

[129] Laitinen, A, Laitinen, LA. Cellular infiltrates in asthma and in chronic obstructive pulmonary disease. *Am Rev Respir Dis*, 143(5 Pt 1):1159–60; discussion 1161, 1991.

[130] National Asthma Education Program Expert Panel. guidelines for the diagnosis and management of asthma. Technical report, US Department of Health and Human Services, 1997.

[131] Haley, KJ, Sunday, ME, Wiggs, BR, et al. Inflammatory cell distribution within and along asthmatic airways. *Am J Respir Crit Care Med*, 158(2):565–72, 1998.

[132] Bousquet, J, Chanez, P, Lacoste, JY, et al. Eosinophilic inflammation in asthma. *N Engl J Med*, 323(15):1033–9, 1990.

[133] Crimi, E, Spanevello, A, Neri, M, Ind, PW, Rossi, GA, Brusasco, V. Dissociation between airway inflammation and airway hyperresponsiveness in allergic asthma. *Am J Respir Crit Care Med*, 157(1):4–9, 1998.

[134] Robinson, DS, Hamid, Q, Ying, S, et al. Predominant TH2-like bronchoalveolar T-lymphocyte population in atopic asthma. *N Engl J Med*, 326(5):298–304, 1992.

[135] Sallusto, F, Mackay, CR, Lanzavecchia, A. Selective expression of the eotaxin receptor CCR3 by human T helper 2 cells. *Science*, 277(5334):2005–7, 1997.

[136] Sutton, BJ, Gould, HJ. The human IgE network. *Nature*, 366(6454):421–8, 1993.

[137] Ravetch, JV, Kinet, JP. Fc receptors. *Annu Rev Immunol*, 9:457–92, 1991.

[138] Sampath, D, Castro, M, Look, DC, Holtzman, MJ. Constitutive activation of an epithelial signal transducer and activator of transcription (STAT) pathway in asthma. *J Clin Invest*, 103(9):1353–61, 1999.

[139] Wenzel, SE, Szefler, SJ, Leung, DY, Sloan, SI, Rex, MD, Martin, RJ. Bronchoscopic evaluation of severe asthma. persistent inflammation associated with high dose glucocorticoids. *Am J Respir Crit Care Med*, 156(3 Pt 1):737–43, 1997.

[140] Galli, SJ. New concepts about the mast cell. *N Engl J Med*, 328(4):257–65, 1993.

[141] Kuwano, K, Bosken, CH, Pare, PD, Bai, TR, Wiggs, BR, Hogg, JC. Small airways dimensions in asthma and in chronic obstructive pulmonary disease. *Am Rev Respir Dis*, 148(5):1220–5, 1993.

[142] Hogg, JC. Pathology of asthma. *J Allergy Clin Immunol*, 92(1 Pt 1):1–5, 1993.

[143] Roche, WR, Beasley, R, Williams, JH, Holgate, ST. Subepithelial fibrosis in the bronchi of asthmatics. *Lancet*, 1(8637):520–4, 1989.

[144] Pare, PD, Wiggs, BR, James, A, Hogg, JC, Bosken, C. The comparative mechanics and morphology of airways in asthma and in chronic obstructive pulmonary disease. *Am Rev Respir Dis*, 143(5 Pt 1):1189–93, 1991.

[145] Lambert, RK, Wiggs, BR, Kuwano, K, Hogg, JC, Pare, PD. Functional significance of increased airway smooth muscle in asthma and COPD. *J Appl Physiol*, 74(6):2771–81, 1993.

[146] Leckie, MJ, ten Brinke, A, Khan, J, et al. Effects of an interleukin-5 blocking monoclonal antibody on eosinophils, airway hyper-responsiveness, and the late asthmatic response. *Lancet*, 356(9248):2144–8, 2000.

[147] Bryan, SA, O'Connor, BJ, Matti, S, et al. Effects of recombinant human interleukin-12 on eosinophils, airway hyper-responsiveness, and the late asthmatic response. *Lancet*, 356(9248):2149–53, 2000.

[148] Doull, IJ, Lawrence, S, Watson, M, et al. Allelic association of gene markers on chromosomes 5q and 11q with atopy and bronchial hyperresponsiveness. *Am J Respir Crit Care Med*, 153(4 Pt 1):1280–4, 1996.

[149] Van Eerdewegh, P, Little, RD, Dupuis, J, et al. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature*, 418(6896):426–30, 2002.

[150] Allen, M, Heinzmann, A, Noguchi, E, et al. Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nat Genet*, 35(3):258–63, 2003.

[151] Zhang, Y, Leaves, NI, Anderson, GG, et al. Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nat Genet*, 34(2):181–6, 2003.

[152] Laitinen, T, Polvi, A, Rydman, P, et al. Characterization of a common susceptibility locus for asthma-related traits. *Science*, 304(5668):300–4, 2004.

[153] Kobilka, BK, Dixon, RA, Frielle, T, et al. cDNA for the human beta 2-adrenergic receptor: a protein with multiple membrane-spanning domains and encoded by a gene whose chromosomal location is shared with that of the receptor for platelet-derived growth factor. *Proc Natl Acad Sci U S A*, 84(1):46–50, 1987.

[154] Barnes, KC. Atopy and asthma genes–where do we stand? *Allergy*, 55(9):803–17, 2000.

[155] Dewar, JC, Wilkinson, J, Wheatley, A, et al. The glutamine 27 beta2-adrenoceptor polymorphism is associated with elevated IgE levels in asthmatic families. *J Allergy Clin Immunol*, 100(2):261–5, 1997.

[156] Turki, J, Pak, J, Green, SA, Martin, RJ, Liggett, SB. Genetic polymorphisms of the beta 2-adrenergic receptor in nocturnal and nonnocturnal asthma. evidence that Gly16 correlates with the nocturnal phenotype. *J Clin Invest*, 95(4):1635–41, 1995.

[157] Hershey, GK, Friedrich, MF, Esswein, LA, Thomas, ML, Chatila, TA. The association of atopy with a gain-of-function mutation in the alpha subunit of the interleukin-4 receptor. *N Engl J Med*, 337(24):1720–5, 1997.

[158] Fanta, C, Fletcher, S. An overview of asthma management. In Rose, B, editor, *UpToDate*. UpToDate, Waltham, MA, 2007.

[159] Castro, M, Zimmermann, NA, Crocker, S, Bradley, J, Leven, C, Schechtman, KB. Asthma intervention program prevents readmissions in high healthcare users. *Am J Respir Crit Care Med*, 168(9):1095–9, 2003.

[160] Gibson, PG, Coughlan, J, Wilson, AJ, et al. Self-management education and regular practitioner review for adults with asthma. *Cochrane Database Syst Rev*, (2), 2000.

[161] Fanta, C. Treatment of acute exacerbations of asthma in adults. In Rose, B, editor, *UpToDate*. UpToDate, Waltham, MA, 2007.

[162] Braun, CM, Huang, SK, Bashian, GG, Kagey-Sobotka, A, Lichtenstein, LM, Essayan, DM. Corticosteroid modulation of human, antigen-specific Th1 and Th2 responses. *J Allergy Clin Immunol*, 100(3):400–7, 1997.

[163] Tseng, YT, Wadhawan, R, Stabila, JP, McGonnigal, BG, Padbury, JF. Molecular interactions between glucocorticoid and catecholamine signaling pathways. *J Allergy Clin Immunol*, 110(6 Suppl):S247–54, 2002.

[164] Adcock, IM, Maneechotesuwan, K, Usmani, O. Molecular interactions between glucocorticoids and long-acting beta2-agonists. *J Allergy Clin Immunol*, 110(6 Suppl):S261–8, 2002.

[165] Barnes, PJ. Scientific rationale for inhaled combination therapy with long-acting beta2-agonists and corticosteroids. *Eur Respir J*, 19(1):182–91, 2002.

[166] Adcock, IM, Stevens, DA, Barnes, PJ. Interactions of glucocorticoids and beta 2-agonists. *Eur Respir J*, 9(1):160–8, 1996.

[167] Roth, M, Johnson, PR, Rudiger, JJ, et al. Interaction between glucocorticoids and beta2 agonists on bronchial airway smooth muscle cells through synchronised cellular signalling. *Lancet*, 360(9342):1293–9, 2002.

[168] Usmani, OS, Ito, K, Maneechotesuwan, K, et al. Glucocorticoid receptor nuclear translocation in airway cells after inhaled combination therapy. *Am J Respir Crit Care Med*, 172(6):704–12, 2005.

[169] Barnes, PJ. Clinical outcome of adding long-acting beta-agonists to inhaled corticosteroids. *Respir Med*, 95 Suppl B:S12–6, 2001.

[170] van der Molen, T, Postma, DS, Turner, MO, et al. Effects of the long acting beta agonist formoterol on asthma control in asthmatic patients using inhaled corticosteroids. the Netherlands and Canadian Formoterol Study Investigators. *Thorax*, 52(6):535–9, 1997.

[171] Henderson, Jr, WR. The role of leukotrienes in inflammation. *Ann Intern Med*, 121(9):684–97, 1994.

[172] Henderson, Jr, WR. Role of leukotrienes in asthma. *Ann Allergy*, 72(3):272–8, 1994.

[173] Spector, SL, Farr, RS. The heterogeneity of asthmatic patients–an individualized approach to diagnosis and treatment. *J Allergy Clin Immunol*, 57(5):499–511, 1976.

[174] Malmstrom, K, Rodriguez-Gomez, G, Guerra, J, et al. Oral montelukast, inhaled beclomethasone, and placebo for chronic asthma. a randomized, controlled trial. Montelukast/Beclomethasone Study Group. *Ann Intern Med*, 130(6):487–95, 1999.

[175] Drazen, JM, Silverman, EK, Lee, TH. Heterogeneity of therapeutic responses in asthma. *Br Med Bull*, 56(4):1054–70, 2000.

[176] Szefler, SJ, Martin, RJ, King, TS, et al. Significant variability in response to inhaled corticosteroids for persistent asthma. *J Allergy Clin Immunol*, 109(3):410–8, 2002.

[177] Liggett, SB. Pharmacogenetic applications of the Human Genome project. *Nat Med*, 7(3):281–3, 2001.

[178] Silverman, E, Hjoberg, J, Palmer, L, Tantisira, K, Weiss, S, Drazen, J. Application of pharmacogenetics to the therapeutics of asthma. In Eissa, N. Huston, D, editors, *Therapeutic Targets of Airway Inflammation*, volume 177, pages 823–838. Marcel Dekker, Inc, New York, NY, 2003.

[179] Raofi, S. Medication therapy in ambulatory medical care: United States, 2003-04. Technical Report Vital Health Stat 13(163), National Center for Health Statistics., 2006.

[180] Health, United States, 2006. Technical report, National Center for Health Statistics, 2006.

[181] Johnston, NW, Sears, MR. Asthma exacerbations . 1: epidemiology. *Thorax*, 61(8):722–8, 2006.

[182] Tan, WC. Viruses in asthma exacerbations. *Curr Opin Pulm Med*, 11(1):21–6, 2005.

[183] Friedlander, SL, Busse, WW. The role of rhinovirus in asthma exacerbations. *J Allergy Clin Immunol*, 116(2):267–73, 2005.

[184] Gern, JE. Rhinovirus respiratory infections and asthma. *Am J Med*, 112 Suppl 6A:19S–27S, 2002.

[185] Schaller, M, Hogaboam, CM, Lukacs, N, Kunkel, SL. Respiratory viral infections drive chemokine expression and exacerbate the asthmatic response. *J Allergy Clin Immunol*, 118(2):295–302; quiz 303–4, 2006.

[186] Grunberg, K, Timmers, MC, Smits, HH, et al. Effect of experimental rhinovirus 16 colds on airway hyperresponsiveness to histamine and interleukin-8 in nasal lavage in asthmatic subjects in vivo. *Clin Exp Allergy*, 27(1):36–45, 1997.

[187] Atkinson, RW, Strachan, DP. Role of outdoor aeroallergens in asthma exacerbations: epidemiological evidence. *Thorax*, 59(4):277–8, 2004.

[188] Murray, CS, Poletti, G, Kebadze, T, et al. Study of modifiable risk factors for asthma exacerbations: virus infection and allergen exposure increase the risk of asthma hospital admissions in children. *Thorax*, 61(5):376–82, 2006.

[189] Heaney, LG, Robinson, DS. Severe asthma treatment: need for characterising patients. *Lancet*, 365(9463):974–6, 2005.

[190] Alvarez, GG, Schulzer, M, Jung, D, Fitzgerald, JM. A systematic review of risk factors associated with near-fatal and fatal asthma. *Can Respir J*, 12(5):265–70, 2005.

[191] Koga, T, Oshita, Y, Kamimura, T, Koga, H, Aizawa, H. Characterisation of patients with frequent exacerbation of asthma. *Respir Med*, 100(2):273–8, 2006.

[192] ten Brinke, A, Sterk, PJ, Masclee, AA, et al. Risk factors of frequent exacerbations in difficult-to-treat asthma. *Eur Respir J*, 26(5):812–8, 2005.

[193] Lung function testing: selection of reference values and interpretative strategies. American Thoracic Society. *Am Rev Respir Dis*, 144(5):1202–18, 1991.

[194] Pellegrino, R, Viegi, G, Brusasco, V, et al. Interpretative strategies for lung function tests. *Eur Respir J*, 26(5):948–68, 2005.

[195] Guyatt, GH, Townsend, M, Nogradi, S, Pugsley, SO, Keller, JL, Newhouse, MT. Acute response to bronchodilator. an imperfect guide for bronchodilator therapy in chronic airflow limitation. *Arch Intern Med*, 148(9):1949–52, 1988.

[196] Waalkens, HJ, Merkus, PJ, van Essen-Zandvliet, EE, et al. Assessment of bronchodilator response in children with asthma. Dutch CNSLD Study Group. *Eur Respir J*, 6(5):645–51, 1993.

[197] Brand, PL, Quanjer, PH, Postma, DS, et al. Interpretation of bronchodilator response in patients with obstructive airways disease. the Dutch Chronic Non-Specific Lung Disease (CNSLD) Study Group. *Thorax*, 47(6):429–36, 1992.

[198] Dompeling, E, van Schayck, CP, Molema, J, et al. A comparison of six different ways of expressing the bronchodilating response in asthma and COPD; reproducibility and dependence of prebronchodilator FEV1. *Eur Respir J*, 5(8):975–81, 1992.

[199] Irvin, C. Use of pulmonary function testing in the diagnosis of asthma. In Rose, B, editor, *UpToDate*. UpToDate, Waltham, MA, 2007.

[200] Dales, RE, Spitzer, WO, Tousignant, P, Schechter, M, Suissa, S. Clinical interpretation of airway response to a bronchodilator. epidemiologic considerations. *Am Rev Respir Dis*, 138(2):317–20, 1988.

[201] Spitzer, WO, Suissa, S, Ernst, P, et al. The use of beta-agonists and the risk of death and near death from asthma. *N Engl J Med*, 326(8):501–6, 1992.

[202] Sears, MR, Taylor, DR, Print, CG, et al. Regular inhaled beta-agonist treatment in bronchial asthma. *Lancet*, 336(8728):1391–6, 1990.

[203] Suissa, S, Ernst, P, Boivin, JF, et al. A cohort analysis of excess mortality in asthma and the use of inhaled beta-agonists. *Am J Respir Crit Care Med*, 149(3 Pt 1):604–10, 1994.

[204] Suissa, S, Hemmelgarn, B, Blais, L, Ernst, P. Bronchodilators and acute cardiac death. *Am J Respir Crit Care Med*, 154(6 Pt 1):1598–602, 1996.

[205] Mullen, M, Mullen, B, Carey, M. The association between beta-agonist use and death from asthma. a meta-analytic integration of case-control studies. *JAMA*, 270(15):1842–5, 1993.

[206] Anderson, HR, Ayres, JG, Sturdy, PM, et al. Bronchodilator treatment and deaths from asthma: case-control study. *BMJ*, 330(7483):117, 2005.

[207] Drazen, JM, Israel, E, Boushey, HA, et al. Comparison of regularly scheduled with as-needed use of albuterol in mild asthma. Asthma Clinical Research Network. *N Engl J Med*, 335(12):841–7, 1996.

[208] Dennis, SM, Sharp, SJ, Vickers, MR, et al. Regular inhaled salbutamol and asthma control: the TRUST randomised trial. Therapy Working Group of the National Asthma Task Force and the MRC General Practice Research Framework. *Lancet*, 355(9216):1675–9, 2000.

[209] Haahtela, T, Jarvinen, M, Kava, T, et al. Comparison of a beta 2-agonist, terbutaline, with an inhaled corticosteroid, budesonide, in newly detected asthma. *N Engl J Med*, 325(6):388–92, 1991.

[210] Simons, FE. A comparison of beclomethasone, salmeterol, and placebo in children with asthma. Canadian Beclomethasone Dipropionate-Salmeterol Xinafoate Study Group. *N Engl J Med*, 337(23):1659–65, 1997.

[211] Greenstone, IR, Ni Chroinin, MN, Masse, V, et al. Combination of inhaled long-acting beta2-agonists and inhaled steroids versus higher dose of inhaled steroids in children and adults with persistent asthma. *Cochrane Database Syst Rev*, (4):CD005533, 2005.

[212] Ni Chroinin, M, Greenstone, IR, Danish, A, et al. Long-acting beta2-agonists versus placebo in addition to inhaled corticosteroids in children and adults with chronic asthma. *Cochrane Database Syst Rev*, (4), 2005.

[213] Niu, T, Rogus, JJ, Chen, C, et al. Familial aggregation of bronchodilator response: a community-based study. *Am J Respir Crit Care Med*, 162(5):1833–7, 2000.

[214] Martinez, FD, Graves, PE, Baldini, M, Solomon, S, Erickson, R. Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. *J Clin Invest*, 100(12):3184–8, 1997.

[215] Tan, S, Hall, IP, Dewar, J, Dow, E, Lipworth, B. Association between beta 2-adrenoceptor polymorphism and susceptibility to bronchodilator desensitisation in moderately severe stable asthmatics. *Lancet*, 350(9083):995–9, 1997.

[216] Lima, JJ, Thomason, DB, Mohamed, MH, Eberle, LV, Self, TH, Johnson, JA. Impact of genetic polymorphisms of the beta2-adrenergic receptor on albuterol bronchodilator pharmacodynamics. *Clin Pharmacol Ther*, 65(5):519–25, 1999.

[217] Drysdale, CM, McGraw, DW, Stack, CB, et al. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci U S A*, 97(19):10483–8, 2000.

[218] Israel, E, Drazen, JM, Liggett, SB, et al. The effect of polymorphisms of the beta(2)-adrenergic receptor on the response to regular use of albuterol in asthma. *Am J Respir Crit Care Med*, 162(1):75–80, 2000.

[219] Taylor, DR, Drazen, JM, Herbison, GP, Yandava, CN, Hancox, RJ, Town, GI. Asthma exacerbations during long term beta agonist use: influence of beta(2) adrenoceptor polymorphism. *Thorax*, 55(9):762–7, 2000.

[220] The Childhood Asthma Management Program (CAMP): design, rationale, and methods. Childhood Asthma Management Program Research Group. *Control Clin Trials*, 20(1):91–120, 1999.

[221] Coultas, DB, Howard, CA, Skipper, BJ, Samet, JM. Spirometric prediction equations for Hispanic children and adults in New Mexico. *Am Rev Respir Dis*, 138(6):1386–92, 1988.

[222] Kanehisa, M, Goto, S, Kawashima, S, Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30(1):42–6, 2002.

[223] Hein, H, Schluter, C, Kulke, R, Christophers, E, Schroder, JM, Bartels, J. Genomic organization, sequence, and transcriptional regulation of the human eotaxin gene. *Biochem Biophys Res Commun*, 237(3):537–42, 1997.

[224] Patel, VP, Kreider, BL, Li, Y, et al. Molecular and functional characterization of two novel human C-C chemokines as inhibitors of two distinct classes of myeloid progenitors. *J Exp Med*, 185(7):1163–72, 1997.

[225] Shinkai, A, Yoshisue, H, Koike, M, et al. A novel human CC chemokine, eotaxin-3, which is expressed in IL-4-stimulated vascular endothelial cells, exhibits potent activity toward eosinophils. *J Immunol*, 163(3):1602–10, 1999.

[226] Ponath, PD, Qin, S, Ringler, DJ, et al. Cloning of the human eosinophil chemoattractant, eotaxin. expression, receptor binding, and functional properties suggest a mechanism for the selective recruitment of eosinophils. *J Clin Invest*, 97(3):604–12, 1996.

[227] Rothenberg, ME. Eotaxin. an essential mediator of eosinophil trafficking into mucosal tissues. *Am J Respir Cell Mol Biol*, 21(3):291–5, 1999.

[228] Gonzalo, JA, Lloyd, CM, Wen, D, et al. The coordinated action of CC chemokines in the lung orchestrates allergic inflammation and airway hyperresponsiveness. *J Exp Med*, 188(1):157–67, 1998.

[229] Moore, KW, de Waal Malefyt, R, Coffman, RL, O'Garra, A. Interleukin-10 and the interleukin-10 receptor. *Annu Rev Immunol*, 19:683–765, 2001.

[230] John, M, Lim, S, Seybold, J, et al. Inhaled corticosteroids increase interleukin-10 but reduce macrophage inflammatory protein-1alpha, granulocyte-macrophage colony-stimulating factor, and interferon-gamma release from alveolar macrophages in asthma. *Am J Respir Crit Care Med*, 157(1):256–62, 1998.

[231] Lim, S, Crawley, E, Woo, P, Barnes, PJ. Haplotype associated with low interleukin-10 production in patients with severe asthma. *Lancet*, 352(9122):113, 1998.

[232] Rosenwasser, LJ, Borish, L. Genetics of atopy and asthma: the rationale behind promoter-based candidate gene studies (IL-4 and IL-10). *Am J Respir Crit Care Med*, 156(4 Pt 2):S152–5, 1997.

[233] Lyon, H, Lange, C, Lake, S, et al. IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genet Epidemiol*, 26(2):155–65, 2004.

[234] Jeannin, P, Lecoanet, S, Delneste, Y, Gauchat, JF, Bonnefoy, JY. IgE versus IgG4 production can be differentially regulated by IL-10. *J Immunol*, 160(7):3555–61, 1998.

[235] John, M, Au, BT, Jose, PJ, et al. Expression and release of interleukin-8 by human airway smooth muscle cells: inhibition by Th-2 cytokines and corticosteroids. *Am J Respir Cell Mol Biol*, 18(1):84–90, 1998.

[236] Kadowaki, N, Ho, S, Antonenko, S, et al. Subsets of human dendritic cell precursors express different toll-like receptors and respond to different microbial antigens. *J Exp Med*, 194(6):863–9, 2001.

[237] Means, TK, Golenbock, DT, Fenton, MJ. The biology of Toll-like receptors. *Cytokine Growth Factor Rev*, 11(3):219–32, 2000.

[238] Lazarus, R, Raby, BA, Lange, C, et al. TOLL-like receptor 10 genetic variation is associated with asthma in two independent samples. *Am J Respir Crit Care Med*, 170(6):594–600, 2004.

[239] Israel, E, Chinchilli, VM, Ford, JG, et al. Use of regularly scheduled albuterol treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. *Lancet*, 364(9444):1505–12, 2004.

[240] Vale, W, Spiess, J, Rivier, C, Rivier, J. Characterization of a 41-residue ovine hypothalamic peptide that stimulates secretion of corticotropin and beta-endorphin. *Science*, 213(4514):1394–7, 1981.

[241] Chrousos, GP. The hypothalamic-pituitary-adrenal axis and immune-mediated inflammation. *N Engl J Med*, 332(20):1351–62, 1995.

[242] Webster, EL, Torpy, DJ, Elenkov, IJ, Chrousos, GP. Corticotropin-releasing hormone and inflammation. *Ann N Y Acad Sci*, 840:21–32, 1998.

[243] Silverman, ES, Breault, DT, Vallone, J, et al. Corticotropin-releasing hormone deficiency increases allergen-induced airway inflammation in a mouse model of asthma. *J Allergy Clin Immunol*, 114(4):747–54, 2004.

[244] Tantisira, KG, Lake, S, Silverman, ES, et al. Corticosteroid pharmacogenetics: association of sequence variants in CRHR1 with improved lung function in asthmatics treated with inhaled corticosteroids. *Hum Mol Genet*, 13(13):1353–9, 2004.

[245] Johnson, GC, Esposito, L, Barratt, BJ, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29(2):233–7, 2001.

[246] Sebastiani, P, Lazarus, R, Weiss, ST, Kunkel, LM, Kohane, IS, Ramoni, MF. Minimal haplotype tagging. *Proc Natl Acad Sci U S A*, 100(17):9900–5, 2003.

[247] Walt, DR. Techview: molecular biology. bead-based fiber-optic arrays. *Science*, 287(5452):451–2, 2000.

[248] Oliphant, A, Barker, DL, Stuelpnagel, JR, Chee, MS. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques*, Suppl:56–8, 60–1, 2002.

[249] Gunderson, KL, Kruglyak, S, Graige, MS, et al. Decoding randomly ordered DNA arrays. *Genome Res*, 14(5):870–7, 2004.

[250] Wigginton, JE, Cutler, DJ, Abecasis, GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*, 76(5):887–93, 2005.

[251] SAS Institute, Inc. SAS software, version 9.1 of the SAS System for Windows., 2002-2003.

[252] Fawcett, T. ROC graphs: Notes and practical considerations for researchers, technical report, MS 1143. Technical report, HP Laboratories, 2004.

[253] Barber, CB, David, PD, Hannu, H. The quickhull algorithm for convex hulls. *ACM Trans Math Softw*, 22(4):469–483, 1996.

[254] Szarek, JL, Zhang, JZ, Gruetter, CA. Mechanisms of 5-hydroxytryptamine-induced contraction of isolated rat intrapulmonary bronchi. *Pulm Pharmacol*, 8(6):273–81, 1995. Hl35711/hl/nhlbi Hl41548/hl/nhlbi Journal Article Research Support, U.S. Gov't, P.H.S. England.

[255] Coyle, AJ, Ackerman, SJ, Burch, R, Proud, D, Irvin, CG. Human eosinophil-granule major basic protein and synthetic polycations induce airway hyperresponsiveness in vivo dependent on bradykinin generation. *J Clin Invest*, 95(4):1735–40, 1995. Ai-22660/ai/niaid Ai-25230/ai/niaid P0i-37665/phs Journal Article Research Support, U.S. Gov't, P.H.S. United states.

[256] Ichinose, M, Barnes, PJ. Bradykinin-induced airway microvascular leakage and bronchoconstriction are mediated via a bradykinin B2 receptor. *Am Rev Respir Dis*, 142(5):1104–7, 1990. Journal Article United states.

[257] Smith, GW, Aubry, JM, Dellu, F, et al. Corticotropin releasing factor receptor 1-deficient mice display decreased anxiety, impaired stress response, and aberrant neuroendocrine development. *Neuron*, 20(6):1093–102, 1998. Dk-26741/dk/niddk Dk09551/dk/niddk Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United states.

[258] Hoffmann, R, Valencia, A. A gene network for navigating the literature. *Nat Genet*, 36(7):664, 2004. Letter United States.

[259] Amrani, Y, Panettieri Jr, RA. Modulation of calcium homeostasis as a mechanism for altering smooth muscle responsiveness in asthma. *Curr Opin Allergy Clin Immunol*, 2(1):39–45, 2002. Journal Article Review United States.

[260] DeLong, E, DeLong, D, Clarke-Pearson, D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845, 1988.

[261] Bacharier, LB, Dawson, C, Bloomberg, GR, Bender, B, Wilson, L, Strunk, RC. Hospitalization for asthma: atopic, pulmonary function, and psychological correlates among participants in the Childhood Asthma Management Program. *Pediatrics*, 112(2):e85–92, 2003.

[262] Fuhlbrigge, AL, Weiss, ST, Kuntz, KM, Paltiel, AD. Forced expiratory volume in 1 second percentage improves the classification of severity among children with asthma. *Pediatrics*, 118(2):e347–55, 2006.

[263] Pernis, AB, Rothman, PB. JAK-STAT signaling in asthma. *J Clin Invest*, 109(10):1279–83, 2002. Journal Article Review United States.