

Linkage Analysis of a Cluster-Based Quantitative Phenotype Constructed from Pulmonary Function Test Data in 27 Multigenerational Families with Multiple Asthmatic Members

Cavan Reilly^a Michael B. Miller^b Yuhong Liu^b William S. Oetting^{c, d}
Richard King^{d, e} Malcolm Blumenthal^e

Divisions of ^aBiostatistics and ^bEpidemiology and Community Health, School of Public Health, ^cCollege of Pharmacy, ^dInstitute of Human Genetics, and ^eDepartment of Medicine, University of Minnesota, Minneapolis, Minn., USA

Key Words

Asthma · Linkage analysis · Locus heterogeneity · Quantitative trait locus

Abstract

Objective: To identify genes involved in phenotypes that increase one's risk for developing asthma, a complex disease that is likely genetically heterogeneous. Unlike other approaches to locus discovery in the presence of heterogeneity, this method seeks loci that segregate in all or most ascertained families while recognizing that other genes and environmental factors that modify the action of the common gene may vary across families. **Methods:** The method is based on seeking groups of families that differ, between groups, in the way affected individuals express the genotype. Then we use the distance of each individual to the cluster center for his family to define a quantitative trait. This quantitative trait is then subjected to a genome scan using variance components methods. **Results:** The method is applied to a data set of 27 multigenerational families with asthma, and a novel locus at 2q33 (at 210 cM) is identified. **Conclusions:** The proposed method has the potential to identify

loci near genes that increase risk for asthma related phenotypes. The method could be used for other complex disorders that exhibit locus heterogeneity.

Copyright © 2007 S. Karger AG, Basel

Introduction

Inadequate phenotype development is a major obstacle to finding susceptibility genes for complex diseases (i.e. diseases involving multiple loci and environmental factors). Disease phenotypes typically are defined by clinical status, which is sensible when the goal is to treat the disorder, but other methods may be more productive when the goal is to determine the genetic etiology of the disorder. Here we consider the problem of identifying susceptibility loci for a complex disease that may exhibit locus heterogeneity for at least one of the genes involved in the disorder [see 1 for more on heterogeneity in the context of quantitative traits]. We focus on quantitative traits as this provides a natural framework for studying complex diseases. The method was developed in the context of seeking genes that increase susceptibility to asth-

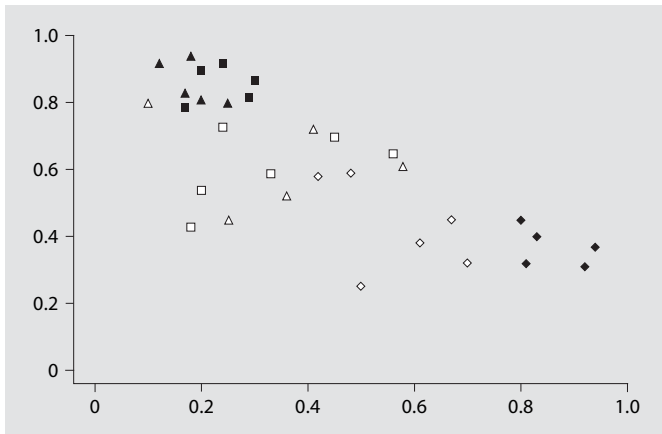


Fig. 1. An idealized example of a plot of 2 quantitative variables that are under the control of a major gene common to all affected but modified by factors that vary with family. In the figure the 3 different shapes correspond to 3 different families and a filled in shape indicates an affected.

ma, hence we also discuss the results from applying the method to this widespread condition. It is widely accepted that asthma and its associated quantitative traits are genetically complex [2] and genetically heterogeneous [3].

The goal here is to identify a major gene involved in a quantitative phenotype that is modified by other genes and environmental factors that are heterogeneous. We would expect that there would be some familial environmental sources and polygenes that vary from family to family, but we would want to localize the common major genes since identification of such genes would allow detection and/or treatment of most with the disorder. We note that in almost all applications, these genetic sources will be confounded with environmental factors that differ between families, but since interest is on identifying the common major gene rather than the polygenes, the distinction is not so important. Rather than develop methods for analyzing genotypic data, we here address the problem by attempting to define quantitative phenotypes based on a collection of variables that are measured for each subject. The constructed phenotypes will then be subjected to a genome scan using variance components linkage methods.

To accomplish these goals, we investigate the use of family information, although not genotypic information, for the construction of phenotypes. Below we present a model that allows incorporation of this information. The basic idea behind the model is best illustrated with an

idealized example (fig. 1). This figure represents data for affected (represented with black symbols) and unaffected (represented with white symbols) for 3 families (the different symbols indicate family membership) for 2 continuous variables (the 2 coordinate axes in the figure). Here we suppose that there is a single major gene that is consistent across families, but there are 2 varieties of polygenes that modify the action of this gene (the family represented by diamonds differs from the other two families). No linear transformation of these variables will result in a variable which will allow separation of the affected and unaffected, but if we measure the distance of each individual to the center of the cluster of the affected in his or her family, we obtain a variable which has the property that small values are associated with the disease allele for the major gene.

Previous approaches to heterogeneity take a different approach than the one proposed here. These other approaches assume that the major gene varies with family and attempt to localize this gene by fitting a mixture model to the recombination fraction. Ekstrøm and Dalgaard [1] have extended the usual approach to mixture modeling for detecting linkage of binary traits to the QTL case. The most important difference between their approach and the approach proposed here is that they assume there is a single phenotype that is measured through a single response variable, and all families either have or do not have linkage at any given locus. In contrast, the method proposed here uses data on multiple measurements and recognizes that there are multiple, slightly varying phenotypes that should likely be classified as affected. In addition, the method proposed here attempts to detect loci that are common to all or most families.

Phenotype Development in Asthma

There have been attempts to refine phenotypes based on clinical understanding. In the context of asthma, the Tuscon Children's Respiratory Cohort Study found distinct wheezing phenotypes in children under the age of 6 [4, 5]. More recently, Kurukulaaratchy et al. [6] have used data on wheezing, atopy, lung function and bronchial hyper-responsiveness to differentiate between 2 groups of subjects they label transient wheezers and persistent wheezers. They have demonstrated statistically significant differences between these groups in terms of morbidity in the first 10 years of life, but they did not investigate the possibility of genetic differences between these phenotypes.

Several investigations have been undertaken to define phenotypes relating to asthma and examine the extent to

which these phenotypes are genetically independent. Palmer et al. [7] concluded that total and specific serum IgE levels, blood eosinophil counts and airway responsiveness to an inhaled agonist are useful as phenotypes since the narrow sense heritability was sufficiently high. In contrast, this work found the heritability of FEV1 was quite low (6.1). This group also found evidence for differing genetic determinants of IgE levels and airway responsiveness [8]. There have also been attempts to conduct segregation analysis using quantitative traits related to asthma [9]. This work is closer to our approach in that it sought the existence of major genes that influence two or more of the following trait values: total IgE levels, blood eosinophil counts and the dose-response slope of methacholine challenge. They found evidence for a single major locus at which a recessive allele increased or decreased each of these phenotypes. Further modeling indicated the 3 traits do not share a common gene, hence this was interpreted as evidence for the existence of at least 3 distinct genetic pathways involving major genes.

Materials and Methods

Data Sources

Our data were originally collected as part of the Collaborative Study on the Genetics of Asthma [10]. The present study used only the 27 multigenerational Caucasian families that were collected in Minnesota. These families had 169 asthmatic members (as defined below), 347 who were not asthmatic and 129 for whom the diagnosis was unavailable. Pulmonary function data were available on 619 individuals, but only the 456 phenotyped subjects who also had genotype data were used in the linkage analysis (there were not genotypic data available for all 619 subjects due to decisions regarding which matings were potentially informative for the phenotype used in the original study).

For the CSGA, families were ascertained through two asthmatic siblings. The families were 'expanded' to permit recruitment of other relatives either by (a) extending the families through asthmatic relatives or (b) including no more than one unaffected relative to permit a lineage to incorporate other relatives with asthma. The inclusion criteria for each family consisted of each of the two asthmatic siblings having met the following criteria for the proband, namely (1) being at least 6 years of age; (2) having either bronchial hyper-responsiveness, defined as a fall from baseline FEV1 greater than 20% in one second after inhalation of 25 mg/ml or less of methacholine, or reversibility, defined as a 15% or greater increase from baseline FEV1 after inhaled bronchodilator (albuterol) for those with reduced baseline FEV1; (3) having the presence of two or more of the symptoms of coughing, wheezing and shortness of breath; (4) having less than three pack-years of cigarette smoking, and (5) having a physician's diagnosis of asthma with no conflicting pulmonary disease. The Institutional Review Board of each institution approved the CSGA protocol, and informed consent for all diagnostic procedures was

obtained from all subjects. All family members underwent a standardized CSGA protocol consisting of an interviewer administered questionnaire, pulmonary function studies including a methacholine challenge and/or reversibility studies, blood drawing for serum IgE levels and skin prick testing using standardized allergens. Additional details of the study design can be found in an earlier publication [10].

Our basic data for the construction of phenotypes here was to use the logarithm of the percent predicted of the following variables: volume exhaled during the first second of a forced expiratory maneuver (FEV1), forced expiratory vital capacity (FVC), maximum expiratory flow when half of the FVC has been exhaled (FEF50) and forced expiratory flow rate over the middle half of FVC (FEF25). While other transformations were considered, the logarithm led to the greatest amount of symmetry in the marginal distributions. Percent predicted refers to an observed value as a percentage of the predicted value given height, sex, and age; thus our analysis controls for sets of confounders that have an established influence on lung function measurements. These variables were used since they were obtained as part of the CSGA protocol and are potentially informative about asthmatics since asthma is a disorder of the respiratory system. These variables were selected since they reflect potentially different aspects of lung function. Other variables relating to lung function that were included in the data set (such as variables based on response to a methacholine challenge, such as PD20) had marginal distributions that departed strongly from normality, hence we were reluctant to include these in a cluster analysis with variables that were roughly normally distributed. Other variables that are frequently used in studies of asthma (such as serum IgE) were excluded in this analysis because we were focusing on lung function. We defined an individual's PFT profile to be the set of these 4 variables.

The Model for Phenotype Construction

Here we describe a probability model for the trait values corresponding to the mechanism described in the introduction. The model uses a 2 level mixture structure. Suppose there are K possible genetic sources of some disease and we have N families with data on individuals. A genetic source is a high-risk genotype involving one or more loci. Furthermore, assume an affected individual has only one of the genetic sources, and all affected individuals in a family have the same genetic source. Our model supposes that the trait values for an individual arise from a family specific mixture distribution with 2 components: a diseased component and a healthy component. Supposing the distribution for those with the disease in family j has mean μ_j , we further suppose that these μ_j arise from another mixture model, reflecting the different genetic sources of the disorder. Note that we assume the number of genetic sources is less than the number of families. If one suspects that each family has a distinct genetic source, one could just conduct a distinct linkage analysis for each family. In some situations, we may have further data on the subjects than just the trait values used for constructing the quantitative phenotype, as in the asthma data set investigated here, hence we allow the mixing proportions in the family specific mixture model to be subject specific, depending on a set of other variables.

In general, if y_{ij} represents the p vector of quantitative variables for subject i in family j , x_{ij} is a set of covariates for this subject indicating if the subject is likely affected, β is a vector of regression

coefficients giving the effect of each element of $x_{ij}, f_j(\mu_j)$ and $h_k(\eta_k)$ are both densities on a Euclidean space of dimension p with means μ_j and η_k respectively, g_j is another density on the same space, λ_{ij} is an indicator variable which is one when subject i in family j is asthmatic, and weights ϕ_k satisfying $\sum_{k=1}^K \phi_k = 1$, then we can write the above model in the following form:

$$y_{ij} \sim \lambda_{ij} f_j(\mu_j) + (1 - \lambda_{ij}) g_j$$

$$\mu_j \sim \sum_{k=1}^K \phi_k h_k(\eta_k)$$

$$\lambda_{ij} \sim \text{Ber}(x'_{ij} \beta).$$

The above model describes the marginal distribution for a single individual. A full likelihood based treatment (which is not pursued here) needs to also take account of correlation in the y_{ij} that is induced by familial relationships. This would be accomplished by using the kinship information for related subjects (unrelated subjects are modeled as independent).

Given this framework, we define our quantitative variable as

$$z_{ij} = \| y_{ij} - \eta_k \|_2^2,$$

where k is the index of the mixture component to which family j belongs (and $\|x\|_2$ is the Euclidean norm of the vector x). We use this as our variable for a genome screen because we want to measure how far an individual is from the typical affected in families that have the same genetic source as the family to which this individual belongs. The Euclidean norm is a convenient method for measuring distance in Euclidean spaces. In practice, we take the natural log of these values and standardize them before conducting the linkage analysis (to improve the approximation to normality).

An Algorithm for Phenotype Construction

Note that the phenotypes z_{ij} depend on a set of parameters (namely η_k for $k = 1, \dots, K$ and cluster membership for all subjects), hence we must estimate these parameters to determine the phenotypic values. A full likelihood based approach would involve many assumptions about the densities in the previous section, and even if one would provisionally assume normality, issues of identifiability would likely arise given the full generality of the parameterization there (for example, would each family have its own covariance matrix describing the covariance of y_{ij} for a common j). Even then computation is daunting (due to the 2 level mixture model combined with estimation of β). Thus we propose an algorithm that is based on the model which combines method of moments estimators with heuristic clustering algorithms.

Here we suppose information is available about likely cases (but this information alone is not sufficient to detect linkage). First, we use this information about likely cases to identify 'affected'. Then we compute the mean of all p variables used for phenotype definition across all 'affected' in a family. The resulting set of mean values (one set of p means for each family) can be thought of as characterizing the typical affected for each family (the typical family level affected). Denote these averages \bar{x}_j for $j = 1, \dots, N$, where \bar{x}_j is a p vector. Next we conduct a cluster analysis on these typical family level affected, \bar{x}_j . If this suggests the presence of clusters, then, for each subject in the data set, we determine family membership and measure the distance of the p vector

of data for that subject, y_{ij} , to the average of the \bar{x}_j 's in the cluster containing that subject's family. That is

$$z_{ij} = \| y_{ij} - \bar{x}_k \|_2^2$$

where k is the index of the cluster to which the affected in family j belongs, and \bar{x}_k is the average of all \bar{x}_j in that cluster. This provides us with a quantitative score for each subject in the study.

This algorithm can be interpreted as an approximate solution to a special case of the general model. Our information about likely affected is used to estimate the set of λ_{ij} , then given this set of estimates, \bar{x}_j provides a consistent estimate of μ_j . Clustering of the typical family level affected corresponds to fitting a mixture model [see 11 for more on this correspondence], and choices regarding the cluster algorithm correspond to models for the distribution of the vectors μ_j . Thus a cluster algorithm implies a choice for h_k in the notation of the probability model, and the selection of a specific cluster algorithm would imply the use of a particular distribution for h_k . In our experience, simple methods, such as k -means clustering [12] or Ward's method [13], are sufficient to identify clusters well supported by the data and typically show substantial agreement with one another [14]. Hence we use k -means clustering to identify clusters of families that may have similar genetic sources of the complex disorder. Since k -means clustering seeks spherically symmetric clusters, we are supposing that the distribution of points drawn from h_k is spherically symmetric, i.e. uncorrelated. In fact it has been reported that k -means does relatively well even when the data depart from this assumption [14]. The number of clusters was selected by examining the change in the within cluster sum of squares as a function of the number of clusters, and choosing the smallest number of clusters so that the drop in this quantity is negligible as the number of clusters increases (this is a commonly used heuristic for selecting the number of clusters, see for example [15]). While many methods for determining the number of clusters have been proposed, none enjoys universal support, so here we qualitatively assess the clusters found using a given number of clusters. If we choose an inappropriate number of clusters, then we would not find linkage using the constructed phenotypes (in particular, our algorithm does not employ the strategy of attempting to use a varying number of clusters and selecting the number of clusters with reference to the results from the linkage analysis as this would distort our type I error).

All computations were done using the statistical software S-plus (Version 3.4 Release 1 for Sun SPARC). The kmeans function was used to find a clustering solution given a set of initial centers. For all cluster analyses presented below, 100 initial sets of centers are used and the solution that gives the lowest pooled average distance to the cluster center is retained as the solution (solutions that have only one item in a cluster are excluded). Initial centers were selected by randomly selecting items to be cluster centers (this prevents the algorithm from breaking down due to the presence of an empty cluster).

Analysis of Pedigrees and Computation of LOD Scores

To ensure that familial relationships had been specified correctly, we ran all pedigree data through the Graphical Representation of Relationships (GRR) program [16]. This program allowed rapid confirmation of most relationships and easy detection of DNA sample mixups between or within families. We detected no errors. We then checked for Mendelian inconsistencies using Ped-

check [17] and we corrected errors by setting appropriate genotypes to missing. Finally, we ran MERLIN [18, 19] with the 'error' option to identify further errors that appeared only as low-probability double recombinants and we again corrected likely errors by setting appropriate genotypes to missing.

Nonparametric variance components linkage analyses were conducted in SOLAR [20] but we used multipoint identity-by-descent (IBD) probabilities computed in MERLIN and Loki [21, 22] to avoid the problems in SOLAR's implementation of the Fulker algorithm [20, 23, 24]. We used marker allele frequencies computed by MERLIN from founders in our pedigree data. When families were small enough (fewer than 27 bits) we used MERLIN's implementation of the precise Lander-Green algorithm to compute multipoint IBD probabilities. For larger families, we used the Markov chain Monte Carlo (MCMC) algorithms implemented in Loki with 1,000,000 iterations and a 50-50 mix of L and M steps. This produced 44 files of IBD probabilities (large and small families for 22 chromosomes) that were converted to the SOLAR 'mibd' format using the program MER2SOL [25]. When kurtosis exceeded 0.5 in absolute value, we computed robust LOD scores based on the multivariate *t* distribution [26, 27]. Finally, we computed empirical singlepoint *p* values by computer simulation using SOLAR's 'lodadj' procedure which assumes complete marker information at a single locus.

Note that since the phenotypes are constructed without reference to any genetic data, the behavior of any statistic under the null hypothesis is the same as in any other use of QTL methods. If there is really no QTL acting to affect some trait, then no matter how we define the phenotype, the nominal significance level will be retained when empirical *p*-values are used.

Results

Results of the Cluster Analysis

For this data set, we have several pieces of information that are informative about the mixing proportions in the top level mixture distribution. These include, a questionnaire response to the question: 'Have you ever had asthma', in addition to several other questionnaire items inquiring about coughing, wheezing and shortness of breath (often a positive response to 2 of 3 of these questions is taken as an indication of asthma). Here we report the use of the question: 'Have you ever had asthma?' to select likely affected.

The algorithm described above was used to define a quantitative trait based on the PFT profiles of the subjects. We used *k*-means clustering with increasing numbers of clusters. Examination of the results from the cluster analysis suggested the presence of 4 clusters. Sensitivity of the results to the number of clusters is presented below.

Cluster Identities

In order to characterize the different varieties of asthmatics delineated in the cluster analysis, we can examine

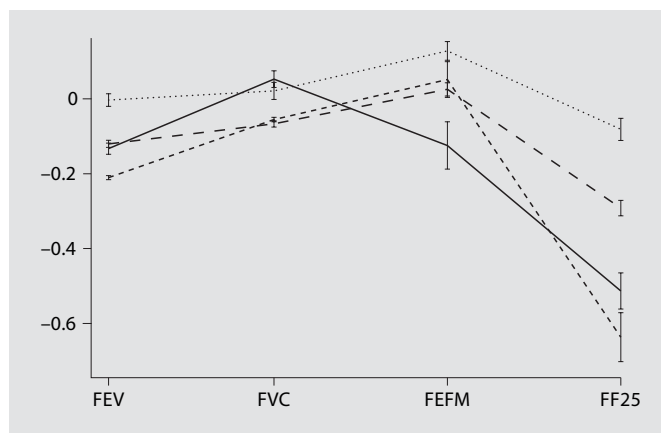


Fig. 2. A graphical representation of the 4 identified clusters. Each line corresponds to a distinct cluster and the vertical lines show the size of the standard errors for each variable within each cluster.

the cluster centers for each variety of asthmatic. If we compare the asthmatics to one another on the basis of the 4 variables used, the 4 types can be described as: (i) high FVC, with low FEFM and FF25; (ii) high FEV1, FEFM and FF25; (iii) low FEV1 and FF25, and (iv) low FVC. For each type, if a variable is not mentioned then that group of asthmatics was moderate for that variable, and high or low is relative to the mean value for asthmatics in this data set. Figure 2 summarizes these results: the solid line is cluster (i), the dotted line is cluster (ii), the line with short dashes is cluster (iii) and the line with the long dashes is cluster (iv). Asthmatics typically have a low ratio of FEV1 to FVC and have low values for FEFM and FF25, hence cluster (i) is closest to this view among the affected. Clusters (ii) and (iii) have similar PFT profiles but differ in the levels of all 4 variables, hence these 2 may represent a distinct variety of asthmatic with the clusters differing in terms of severity. Finally cluster (iv) is similar to the typical asthmatic in terms of the PFT profile except for high levels of FEFM.

Results from the Genome Scans

The results of the SOLAR genome scan for the derived quantitative phenotype for all chromosomes are shown in figure 3. Initially, a peak LOD score of 3.17 was observed on chromosome 2 at marker D2S2944 which was located at 210 cM on our map. The peak was approximately 15 cM wide and was flanked by marker D2S1384 at 200 cM and marker D2S434 at 215 cM. The addition of four microsatellite markers in that region (D2S1782, D2S1369, D2S1345 and D2S1371) increased the peak LOD

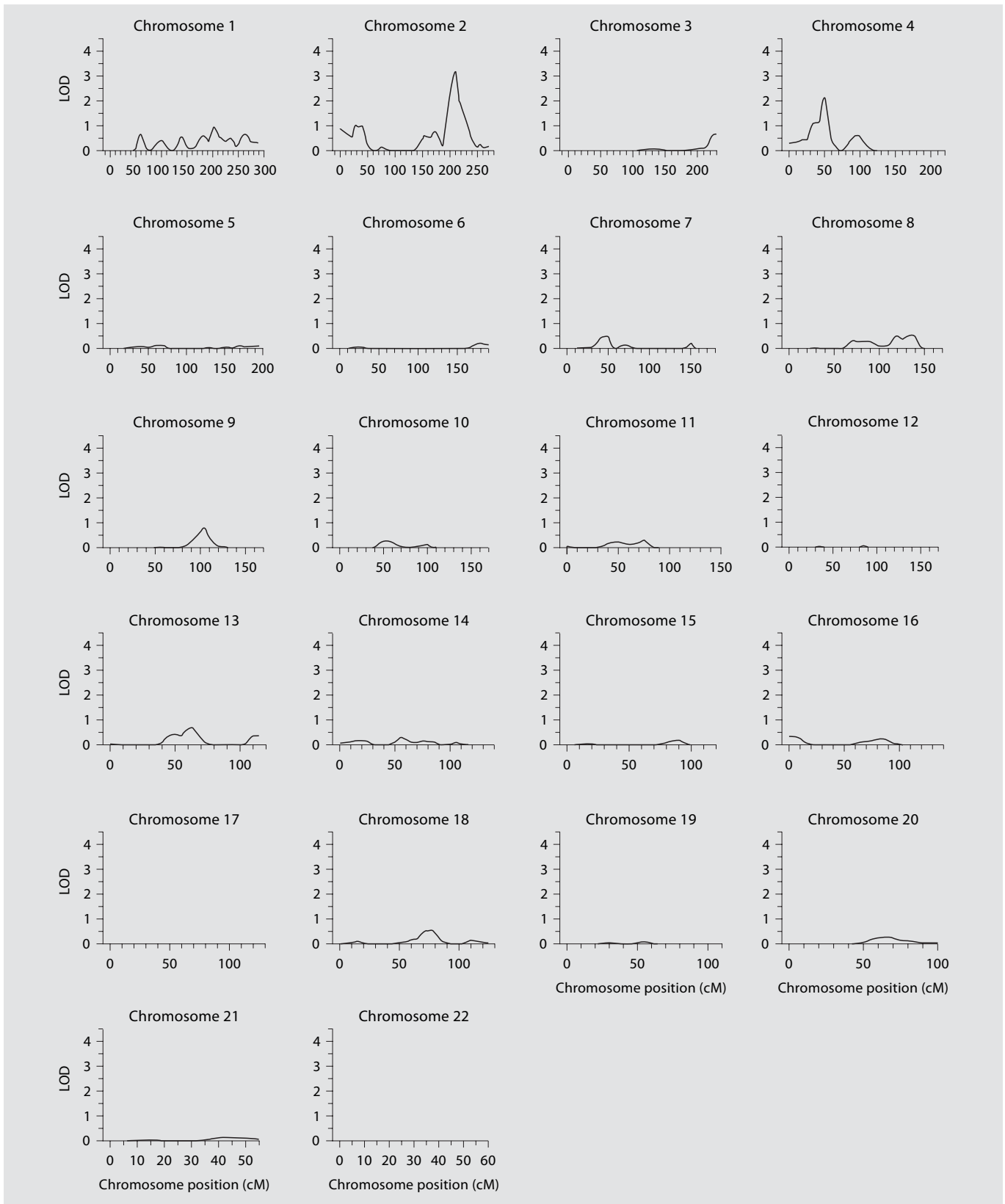


Fig. 3. LOD scores for all markers included in the present study.

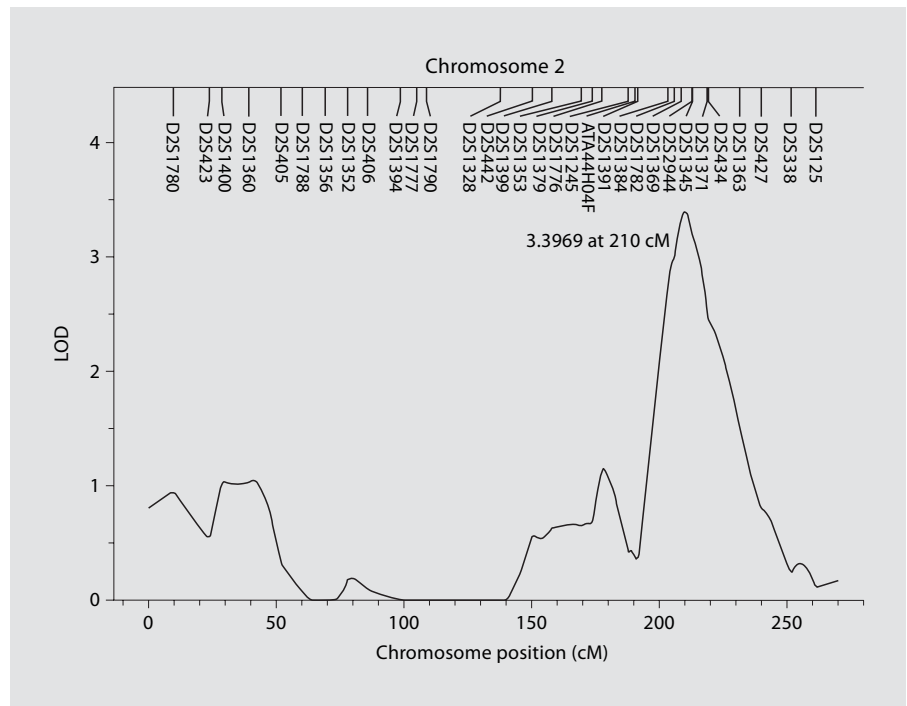


Fig. 4. LOD scores for chromosome 2 with the addition of 4 markers.

score to 3.40 at the same location. Figure 4 shows the LOD scores after the addition of these markers. High LOD scores in model free variance-components QTL linkage analyses can be due to false positive errors when the trait is non-normal with a large kurtosis. Our cluster phenotype had a kurtosis of only 0.18, which is not a concern. It is still wise to do further testing of the robustness of the finding by computer simulation. Holding phenotypes constant, we performed a gene-dropping simulation where genotypes for a single, fully-informative marker locus were transmitted according to Mendel's laws under the null model of no QTL effect. We repeated this procedure 100,000 times and computed a LOD score for every repetition. The largest of these 100,000 null-model LOD scores was 3.26, which is less than the observed LOD of 3.40. Therefore, we have an empirical p value of less than 0.00001 for our LOD of 3.40. By contrast, a LOD score of 3.0, when the normality assumption holds, corresponds to a p value of 0.0001, an order of magnitude larger than our empirical p value.

We looked at the contributions of individual families to the total peak LOD on chromosome 2. We found that 16 families contributed positive LOD scores and 11 families contributed negative LOD scores. The highest family LOD scores were 1.15, 1.09, 0.69, 0.55 and 0.53. This is an encouraging result because it indicates that the

linkage evidence was not due to the extraordinary influence of a single unusual family. The family LOD scores were essentially the same when the robust method was used.

We also conducted full genome scans in SOLAR for all four of the variables used to construct the derived phenotype. None of the four showed evidence for linkage anywhere in the genome. The highest LOD score among all four across all chromosomes was 0.72, and no trait had a substantial peak on chromosome 2. Thus the proposed method used a nonlinear mapping from a 4-vector to a scalar where there was no linkage signal from any component of the 4-vector and yet there was for the scalar. In addition the distance of each subject to the center of all affected was used as a variable (to assess the role of clustering of family level affected), and no significant linkage signal was found (the highest LOD was 0.3458).

To investigate the possibility of locus heterogeneity within the context of our overall approach, we also constructed four phenotypes as follows. We followed the algorithm for phenotype construction except, rather than computing the distance of each subject to the cluster center for his or her family, for each subject, we computed the distance to each of the four cluster centers. If there were four different types of asthmatics each with a distinct set of genes predisposing to that variety of asthma,

then one would expect that this approach should provide evidence for linkage. But the results of the genome screens found no such evidence: The highest LOD score for all four variables was 2.73. Moreover, when we used the robust variance components method, the peak LOD dropped to 0.97 and moved to a different chromosome which indicates that the high LOD score was most likely attributable to outliers. In contrast, the highest LOD score observed for our phenotype was 3.17 (with the original 10 cM map) and it still exceeded 3.0 when we used the robust method.

The 2q33 region has been identified in other genome scans [10, 28], but the peak at 210 cM has not been specifically identified in previous studies. Most other studies concentrate on the region of 2q33 where CD28 and CTLA4 are located, which is at approximately 200 cM. An investigation of known genes in this region reveals several promising candidates, some of which have shown evidence for linkage to phenotypes associated with asthma: inducible T-cell co-stimulator, cytotoxic T-lymphocyte-associated protein 4 (CTLA4), CD28 antigen and the interleukin 8 receptors alpha and beta. Van Oosterhout et al. [29] noted that CTLA4 is expressed only in activated T cells and is a powerful negative regulator of T cell activation. Yang et al. [30] reported that a polymorphism in CTLA4 is associated in females with elevated serum levels of total IgE and allergic rhinitis. SNPs in CTLA4 have been reported to be associated with serum IgE, asthma severity, airway responsiveness, and asthma [3]. Both IL-8 and its receptors have been shown to play a role in the inflammatory process of lung disease [31]. Thus the present study finds further support for previously reported genes (CTLA4) and suggests others which have been implicated in inflammatory lung disease but have not shown evidence of linkage to asthma related phenotypes previously.

Sensitivity Analysis

Here we consider the effects of perturbations of the methodology on the results we find. In general, the location of the highest LOD score is quite robust with regard to variations in the methodology while the actual LOD score is less so. In general, the composition of the data set, in terms of variables and families, is rather important, whereas the number of clusters is not so important. This is encouraging as determining the number of clusters in a data set is a rather difficult problem, while on the other hand, a strength of this investigation is our large data set.

First, we consider varying the set of variables used in the cluster analysis. While there are many ways one could do this, here we report the results from clustering a set of only 3 (rather than 4) variables. If we exclude the variable FF25 then the location of the maximum LOD score shifts to another location and is only 0.8138. If we exclude the variable FEFM then the maximum LOD score is in the same location as our identified locus but only has the value 0.6162. In contrast, if either of the other 2 variables are excluded from the analysis then the location of the maximum LOD score shifts 1 cM and is 4.2606 when FVC is excluded and 3.4499 when FEV1 is excluded.

Next we consider the effect of using a different number of clusters. First, if we use just one cluster and measure all subjects from the center of the affected in this cluster than there is no evidence for linkage to any locus. Given that 4 clusters were selected based on changes in the within cluster sum of squares and the composition of the clusters, we consider the effect of perturbing the number of clusters by 1. In both cases the highest LOD score is observed at locus previously identified. When there are 3 clusters the LOD score at this locus is 3.2017 and when there are 5 clusters the LOD score is 2.6262.

In addition, rather than use the self report of whether a subject is asthmatic, we also consider the effect of using a confirmed doctor's diagnosis to identify asthmatics. The effect of this perturbation is to move the maximal LOD score to another location on the genome, but this LOD score is only 0.7537 which is not statistically significant. A possible reason for the loss of signal in this case is that 16% of the subjects identified as asthmatics using self-report are not identified as asthmatic using this more strict criterion.

Finally, in order to gain an understanding of the effect of individual families on the cluster analysis and resulting phenotype definition, we dropped each family from the analysis and then conducted the cluster analysis (always using 4 clusters). We next determined what cluster the excluded family should be classified as by comparing typical affected for the dropped family to the cluster centers identified in the cluster analysis. We then recomputed the phenotype for all subjects in the data set (including the family excluded from the cluster analysis) and computed LOD scores. The results of this analysis were that the largest LOD score was always at our previously identified locus or 1 cM away and ranged from 0.4143 to 4.1600 (the quartiles are 0.7109 and 1.789).

Discussion

Here we have proposed a novel method for defining quantitative disease phenotypes for complex disorders that may exhibit certain forms of heterogeneity. This work differs from other approaches that account for heterogeneity because it seeks loci that are common among most affected while recognizing that other genes likely modify the action of any common loci. The method was applied to a large data set of asthmatics and their extended families, and a locus was identified that includes several genes that are logical candidates for predisposition to asthma. Evidence was found for linkage to some genes that have previously been reported in the literature, and some evidence was found for linkage to genes that have not been previously been reported but for which other studies have suggested a role in inflammatory lung disease. Fine mapping of this region is currently underway.

While the proposed methods have been successful in identifying a locus of potential interest here, there are a number of difficult unanswered questions that remain. First, it is possible that for some forms of the disorder,

many genes are involved and each only has a very small marginal effect. This would make identification of the genes involved difficult with this method and perhaps a method that searches for multiple genes simultaneously (or their interaction) could be used in conjunction with the proposed methodology. Nor is the effect of the ascertainment scheme obvious. This method clearly needs data for many families each with a substantial number of affected to be useful, and the ascertainment scheme for the CSGA data set has this property. Finally, there is not universal agreement on how to interpret PFT profiles (as defined here). One of our clusters corresponds to the typical PFT profile of an asthmatic, yet the others appear to be distinct from this profile. This further corroborates the view that asthma is a trait that exhibits locus heterogeneity.

Acknowledgements

This work was supported in part by NHLBI grant number 5RO1-HL09609-1. This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute.

References

- 1 Ekström C, Dalgaard P: Linkage analysis of Quantitative Trait Loci in the presence of heterogeneity. *Hum Hered* 2003;55:16–26.
- 2 Daniels SE, Bhattacharrya S, James A, Leaves NI, Young A, Hill MR, Faux JA, Ryan GF, le Söuef PN, Lathrop GM, Musk AW, Cookson W: A genome-wide search for quantitative trait loci underlying asthma. *Nature* 1996;383:247–250.
- 3 Hoffjan S, Nicolae D, Ober C: Association studies for asthma and atopic diseases: A comprehensive review of the literature. *Resp Res* 2003;4:14.
- 4 Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ: Asthma and wheezing in the first six years of life. *N Eng J Med* 1995;332:133–138.
- 5 Stein RT, Holberg CJ, Morgan WJ, Wright AL, Lombardi E, Taussig L, Martinez FD: Peak flow variability, methacholine responsiveness and atopy as markers for detecting different wheezing phenotypes in childhood. *Thorax* 1997;52:946–952.
- 6 Kurukulaaratchy RJ, Fenn MH, Waterhouse LM, Matthews SM, Holgate ST, Arshad SH: Characterization of wheezing phenotypes in the first 10 years of life. *Clin and Exp Allergy* 2003;133:573–578.
- 7 Palmer LJ, Burton PR, James AL, Musk AW, Cookson WO: Familial aggregation and heritability of asthma-associated quantitative traits in a population-based sample of nuclear families. *Eur J Hum Genet* 2000;8:853–860.
- 8 Palmer LJ, Burton PR, Faux JA, James AL, Musk AW, Cookson WO: Independent inheritance of serum immunoglobulin E concentrations and airway responsiveness. *Am J Resp Crit Care Med* 2000;161:1836–1843.
- 9 Palmer LJ, Cookson WO, James AL, Musk AW, Burton PR: Gibbs sampling-based segregation analysis of asthma-associated quantitative traits in a population-based sample of nuclear families. *Genet Epidemiol* 2001;20:356–372.
- 10 The Collaborative Study on the Genetics of Asthma (CSGA): A genome-wide search for asthma susceptibility loci in ethnically diverse populations. *Nat Genet* 1997;15:389–392.
- 11 Banfield J, Raftery A: Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993;40:1089–1094.
- 12 MacQueen JB: Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1*, University of California Press, Berkeley, CA, 281–297, 1967.
- 13 Ward JH: Hierarchical groupings to optimize an objective function. *J Am Stat Assoc* 1963;58:236–244.
- 14 Reilly C, Wang C, Rutherford M: A rapid method for the comparison of cluster analyses. *Statistica Sinica* 2005;15:19–33.
- 15 Tibshirani R, Walther G, Hastie T: Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc B* 2001;63:411–423.
- 16 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: GRR: graphical representation of relationship errors. *Bioinformatics* 2001;17:742–743.
- 17 O'Connell JR, Weeks DE: PedCheck: A program for identifying genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998;63:259–266.
- 18 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30:97–101.
- 19 Nicolae D, Cox NJ: MERLIN... and the geneticist's stone? *Nat Genet* 2002;30:3–4.
- 20 Almasy L, Blangero J: Multipoint quantitative trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998;62:1198–211.
- 21 Heath SC: Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 1997;61:748–760.

- 22 Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM: MCMC segregation and linkage analysis. *Genet Epidemiol* 1997;14:1011–1015.
- 23 Fulker DW, Cherny SS, Cardon LR: Multi-point interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet* 1995;56:1224–1233.
- 24 Fulker DW, Cherny SS: An improved multi-point sib-pair analysis of quantitative traits. *Behav Genet* 1996;26:527–532.
- 25 Miller, MB MER2SOL: Translating MERLIN or Loki IBD data to SOLAR format. *Genet Epidemiol* 2003;25:261–262.
- 26 Lange KL, Little RJA, Taylor JMG: Robust statistical modeling using the t distribution. *J Am Stat Assoc* 1989;84:881–896.
- 27 Blangero J, Williams JT, Almasy L: Robust LOD scores for variance component-based linkage analysis. *Genet Epidemiol* 2000;19:S8–14.
- 28 Lee JK, Park C, Kimm K, Rutherford MS: Genome-wide multilocus analysis for immune-mediated complex diseases. *Biochem Biophys Res Commun* 2002;295:771–773.
- 29 van Oosterhout AJM, Deurloo DT, Groot PC: Cytotoxic T lymphocyte antigen 4 polymorphisms and allergic asthma. *Clin Exp Allergy* 2004;34:4–8.
- 30 Yang KD, Liu C-A, Chang J-C: Polymorphism of the immune-braking gene CTLA-4 (+49) involved in gender discrepancy of serum total IgE levels and allergic diseases. *Clin Exp Allergy* 2004;34:32–37.
- 31 Pease JE, Sabroe I: The role of interleukin-8 and its receptors in inflammatory lung disease: Implications for therapy. *Am J Respir Med* 2005;280:4808–4816.