An Application of the Latent p Value Method to Assess Linkage in Asthma Pedigrees

Craig C. Teerlink  Alun Thomas

Introduction

The latent p value methodology attempts to identify the null distribution of a test statistic that is a function of latent, or unobservable, variables in order to conduct meaningful hypothesis tests. Linkage analysis is an ideal setting for this method because it relies on latent variables in the form of inheritance patterns among related individuals in a pedigree that are not directly observable. Linkage analysis uses observable genetic marker data to infer such inheritance patterns at or between marker loci. The latent p value approach to linkage is an attempt to sample directly from the distribution of possible inheritance patterns in order to assess significance of findings. Here, we demonstrate that the latent p value method can be successfully implemented with existing software and provides greater computational efficiency compared to a conventional p value calculation.

Abstract

Objective: The latent p value is a recently proposed empirical method for assessing evidence against a null hypothesis in a stochastic system involving latent, unobservable variables. It is particularly applicable to genome-wide genetic linkage analysis for test statistics with poorly defined analytical distributions. Methods: We describe an implementation of the latent p value method and its application to a linkage analysis of asthma in 81 extended pedigrees containing 1,858 people genotyped at 533 microsatellite markers. We compare the performance of the latent p value method to a more conventional p value calculation. We also compare the performance of various linkage statistics within this pedigree resource. Results: Using a novel linkage score referred to as the C-link statistic, our analysis provides strong evidence for a recessive gene influencing asthma on chromosome 5q13 (median latent p value = 0.03). We also demonstrate remarkable improvement in computational requirements compared to a more conventional empirical p value calculation. Conclusions: The latent p value method is indeed feasible and provides a computationally efficient means to evaluate evidence for linkage regardless of the choice of linkage statistic.

Key Words
Linkage analysis · Extended pedigrees · Genetic heterogeneity
tistic to the p value scale. The resulting distribution is called the latent p value for the test statistic. The latent p value expresses both the strength of the evidence against the null hypothesis and the level of uncertainty in that expression [2].

A primary advantage of the latent p value method is the ability to convert any arbitrary test statistic to the p value scale. We take advantage of this flexibility in two ways. First, it is well known that conventional linkage statistics such as the two-point and multi-point lod scores have reduced power to detect linkage in the presence of interfamilial heterogeneity. Alternative linkage statistics that have been proposed to account for the presence of interfamilial heterogeneity may lack clear interpretability since their theoretical distributions may not be available [3]. The latent p value method allows us to include such statistics in our analysis by providing a means to clearly interpret results obtained from them, independent of specific knowledge about their asymptotic behavior. Second, we use the latent p value method to interpret globally maximized test statistics. The fact that the latent p value method allows all test statistics to be converted to the p value scale conforms to the notion advanced by Kurbasic and Hossjer [4] that the p value is a more appropriate measure of performance than the actual value of a maximal test statistic itself.

A primary motivation of our implementation is to test for the presence of linkage in an extended pedigree resource ascertained for asthma. Although the precise etiology of asthma is unknown, previous evidence suggests a heritable contribution may exist [5]. Furthermore, analysis of extended pedigrees provides a statistically powerful approach for mapping rare and highly penetrant genetic factors [6, 7]. We previously reported the results of a genome-wide scan for asthma in 81 extended pedigrees genotyped on 533 short tandem repeat markers with an average spacing of 6 cM. In that analysis, we reported a significant linkage on chromosome 5q under a general recessive model using the het-TLOD linkage statistic [9]. We interpreted the significance of this finding using a published threshold for reporting results of genome-wide scans, where it is suggested that LOD scores exceeding 3.2 can be considered as significant [10]. However, since the threshold was derived from the theoretic distribution of a two-point LOD score linkage statistic, it may not be generally applicable to other linkage statistics with potentially different asymptotic properties, such as the het-TLOD statistic we used in our previous analysis. Our implementation of the latent p value approach is in part intended to test the reliability of interpreting test statistics using thresholds based on asymptotic statistics, such as those appearing in Lander and Kruglyak [10], at least in this resource. Hence, we followed an analytical strategy similar to that of our previous report to see if the more rigorous latent p value approach will provide the same conclusion with respect to significance as our previous asymptotic threshold approach.

### Materials and Methods

#### Asthma Pedigree Resource

We analyzed 1,874 individuals, including 742 affected, in 81 pedigrees ascertained for asthma that ranged from 3 to 6 generations in size. Individuals were genotyped on a genome-wide set of 533 short tandem repeat markers with an average spacing of 6 cM. A more detailed description of the study cohort and phenotype determination is given elsewhere [7].

#### Dealing with Linkage Heterogeneity

Goring and Terwilliger [11] proposed the TLOD statistic as a solution to the sensitivity of the multipoint linkage statistic to model misspecification. The TLOD statistic uses multipoint information yet also includes optimization over the recombination fraction as a mechanism to absorb potential model misspecification. The TLOD statistic can therefore be considered as a more informative two-point linkage statistic and has been implemented in MCLINK software [9]. In defining the TLOD statistic, we first consider the standard two-point linkage statistic which is based on the likelihood function, for each marker locus $i = 1, ..., M$,

\[
L(\theta_i; D, g_i) = P(D, g_i | \theta_i)
\]

where $D$ represents the disease status of pedigree members, $g_i$ represents the observed genotypes at the i-th locus, and $\theta_i$ represents the recombination fraction, or the estimated genetic distance between the putative disease locus and the i-th marker locus. It should be noted that the likelihood function is also conditional on additional parameters that describe the mode of inheritance of the disease at a linked locus. We have omitted these parameters from our notation since we assume a fixed set of parameters for all analyses. If we define $x_i$ as the inheritances of the i-th locus, the likelihood function can be written as

\[
L(\theta_i; D, g_i) = \sum_{x_i} P(D, g_i, x_i | \theta_i)
\]

\[
= \sum_{x_i} P(D | x_i, g_i, \theta_i) P(x_i | g_i) P(g_i | \theta_i)
\]

\[
= \sum_{x_i} P(D | x_i, \theta_i) P(x_i | g_i) P(g_i),
\]

\[
\propto \sum_{x_i} P(D | x_i, \theta_i) P(x_i | g_i).
\]

In equation 2, the second and third lines are equal because $P(D | x_i, g_i, \theta_i)$ does not depend on $g_i$ – the $x_i$ are sufficient statistics for linkage. Further, in the terms $P(x_i | g_i, \theta_i)$ and $P(g_i | \theta_i)$, the $x_i$ and $g_i$ do not depend on $\theta_i$ because $\theta_i$ are parameters concerning the disease locus only. The third and fourth lines of equation 2 are proportional to each other because $P(g_i)$ is a constant, since the observed data $g_i$ is fixed.
The standard two-point LOD score at the \( i \)-th locus is

\[
LOD_i = \max_{\theta_i} \log \frac{L(\theta_i; D, g_i)}{L(\frac{1}{2}; D, g_i)} = \max_{\theta_i} \log \frac{\sum_{x_i} P(D|x_i, \theta_i) P(x_i|g_i)}{\sum_{x_i} P(D|x_i, \frac{1}{2}) P(x_i|g_i)}. \tag{3}
\]

For the TLOD, we replace \( P(x_i|g_i) \) with \( P(x_i|G) \), where \( G \) denotes the genotypes observed for all loci similar to the multipoint linkage statistic. The definition of the TLOD at the \( i \)-th locus is then

\[
TLOD_i = \max_{\theta_i} \log \frac{\sum_{x_i} P(D|x_i, \theta_i) P(x_i|G)}{\sum_{x_i} P(D|x_i, \frac{1}{2}) P(x_i|G)}. \tag{4}
\]

Summing over all possible inheritance configurations in a multipoint linkage analysis of an extended pedigree resource may become an intractable problem since the number of possible inheritance configurations increases exponentially with pedigree size and the number of markers. However, we can use Monte Carlo Markov Chain (MCMC) techniques to approximate the likelihood function from a sufficient sample of possible inheritance configurations. We used MCLINK software to generate samples of inheritance states in our analysis. MCLINK uses a blocked Gibbs sampling method to sequentially generate inheritance states based on multipoint information [12]. For a sample of \( n \) inheritance states, at the \( i \)-th locus we approximate the likelihood function as

\[
L(\theta_i; D, G) \approx \frac{1}{n} \sum_{1 \leq i \leq n} P(D|X_{ij}, \theta_i). \tag{5}
\]

where each \( X_{ij} \sim P(x_i|G) \). We approximate the TLOD as

\[
TLOD_i \approx \max_{\theta_i} \log \frac{\sum_{k=1}^{n} P(D|x_{ij}, \theta_i)}{\sum_{k=1}^{n} P(D|x_{ij}, \frac{1}{2})} = \max_{\theta_i} \log \frac{\sum_{k=1}^{n} P(D|x_{ij}, \theta_i)}{\sum_{k=1}^{n} P(D|x_{ij}, \frac{1}{2})}, \tag{6}
\]

where each \( X_{ij} \sim P(x_i|G) \).

While the TLOD statistic may provide some benefit to a general modeling parametric analysis in terms of overcoming sensitivity to potential model misspecification, the statistic does not account for interfamilial heterogeneity and may suffer from low power if this is a feature of the data. Smith [13] and later, Ott [14] proposed the heterogeneity LOD score (het-LOD) to account for the presence of interfamilial heterogeneity. The het-LOD statistic estimates the proportion of linked pedigrees (denoted \( \alpha \)) and unlinked pedigrees, and retains the score that corresponds to the optimal estimate of \( \alpha \). In our analysis, we incorporated the parameter \( \alpha \) with the TLOD statistic, which we refer to as the het-TLOD. We define the het-TLOD for the \( i \)-th locus as

\[
\text{het-TLOD}_i \approx \max_{\alpha, \theta_i} \sum_{k=1}^{Z} \log (\alpha \Lambda_k(\theta_i) + (1 - \alpha)), \tag{7}
\]

where \( \Lambda_k \) denotes the likelihood ratio of the \( k \)-th pedigree, for \( k = 1, \ldots, Z \).

While the het-LOD statistic has become the standard approach for dealing with interfamilial heterogeneity, other statistics have also been proposed. One alternative is the C-statistic proposed by MacLean et al. [3]. The C-statistic is a sum of individual pedigree lod scores, each maximized with respect to a unique recombination fraction. The difference between the TLOD and the C-statistic is the order in which maximization with respect to the recombination fraction and the summation of likelihood ratios occurs. For instance, the TLOD statistic given in equation 10 can be rewritten as

\[
TLOD_i \approx \max_{\theta_i} \sum_{k=1}^{Z} \log \frac{\sum_{j} P(D|X_{ijk}; \theta_i)}{\sum_{j} P(D|X_{ijk}, \frac{1}{2})}, \tag{8}
\]

where \( X_{ijk} \sim P(x_i|G) \) and \( k \) denotes individual pedigrees. In contrast, the C-statistic is defined as

\[
C_i \approx \max_{\theta_i} \sum_{k=1}^{Z} \log \frac{\sum_{j} P(D|X_{ijk}, \theta_{ik})}{\sum_{j} P(D|X_{ijk}, \frac{1}{2})}, \tag{9}
\]

where \( X_{ijk} \sim P(x_i|G) \). Valid interpretation of the C-statistic requires empirical assessment since the statistic is a function of the number of pedigrees and hence a potentially large number of free parameters, \( \theta_{ik} \).

Recently, Christensen et al. [15] proposed the sumlink statistic to account for the presence of linkage heterogeneity. The sumlink statistic is the summation over pedigrees of classic multipoint LOD scores (\( \theta = 0 \)) at each locus that exceed a threshold of 0.588, which corresponds to a nominal p value of 0.05. For our analysis, we combined the threshold requirement of the sumlink statistic with the C-statistic, which we refer to as the C-link statistic. More precisely, we define the C-link statistic as

\[
C_{\text{link}} \approx \sum_{k=1}^{Z} \max_{\theta_i} \left\{ \log \frac{\sum_{j} P(D|X_{ijk}, \theta_{ik})}{\sum_{j} P(D|X_{ijk}, \frac{1}{2})} \geq 0.588 \right\}, \tag{10}
\]

where \( X_{ijk} \sim P(x_i|G) \) and where \( \max \{ \cdot \} \) equals the expression on the left of the inequality if the term in the brackets is true and zero otherwise. We included the C-link, TLOD, and het-TLOD statistics in our analysis so that we could compare the relative power of these three statistics in this resource. We used a general recessive model identical to that used in our previously reported genomewide scan, where we assumed a disease allele frequency of 0.05 and a penetrance of 50% for gene carriers and 0.5% for non-gene carriers.

**The Latent \( p \) Value Approach**

Using the TLOD statistic as an example, to get the latent \( p \) value, we first find

\[
S_q = \max_{1 \leq i \leq M} \left\{ \max_{\theta_i} \log \frac{P(D|X_{iq}, \theta_i)}{P(D|X_{iq}, \frac{1}{2})} \right\}, \tag{11}
\]

where \( X_{iq} \sim P(x_i|G) \) for \( q = 1, \ldots, \) by MCMC. Each \( S_q \) is a genomewide maximum test statistic calculated from a genomewide sample of inheritance states. We refer to the set of observations \( X_{1q}, \ldots, X_{Mq} \) as the conditional simulations and to the set of observations \( S_1, \ldots, S_j \) as the conditional test statistics, since they are sampled conditionally on the observed marker data. For each meiosis, the latent inheritance variable at a locus depends on the state of the previous locus according to the recombination fraction between them, assuming no interference. Hence, the simulations are generated so as to be consistent with the specified recombination fractions.

Then, for \( t = 1, \ldots, M \), we simulate \( X_{it} \) from \( P(x_i) \) at each locus \( i = 1, \ldots, M \) by Monte Carlo (MC) and find

\[
R_t = \max_{1 \leq i \leq M} \left( \max_{\theta_i} \log \frac{P(D|X_{it}, \theta_i)}{P(D|X_{it}, \frac{1}{2})} \right). \tag{12}
\]
Each $R_i$ is also a genome-wide maximum test statistic calculated from a genome-wide sample of inheritance states. However, each $X_m$ is generated unconditionally of observed marker data and is conditional only on pedigree structure and affection status. We refer to the set of observations $X_{11}, \ldots, X_{Mm}$ as the unconditional simulations and to the set of observations $R_0, \ldots, R_m$ as the unconditional test statistics. Since the unconditional samples are generated independently of the observed marker data, there are no constraints to the MCMC algorithm in the form of allele assignments, effectively reducing the MCMC algorithm to simple MC. The unconditional simulations reflect random assignment of allele inheritance states based on allele frequencies and an assumed rate of recombination. The reduced number of constraints reduces the time required to generate the unconditional samples.

The latent p value is the empirical distribution of $\{p_d\}$ where

$$p_d = \frac{1}{m} \sum_{1 \leq l \leq m} I\{R_l \geq S_d\},$$

and where $I\{\cdot\}$ equals one if the expression in the braces is true, and zero otherwise. The number of required simulations is $l$ MCMC plus $m$ MC. The latent p value for the het-TLOD and the C-link statistics are computed similarly.

The distribution of the latent p value provides a means to assess the strength of the evidence against the null hypothesis. The median of the distribution provides a point estimate of the p value for a test of the null hypothesis of linkage. Since all such point estimates are on the p value scale, a point estimate of, say, 0.05 indicates that the associated randomized test rejects at level 0.05 half the time [2]. For our purposes, a point estimate lower than 0.05 will be regarded as a significant result. Note that our application of the latent p value method relies on an omnibus test statistic, the genome-wide maximum test statistic. The latent p value method for omnibus test statistics presents a fair test for linkage if the same strategy for selection over markers occurs in both the conditional and unconditional sets of simulations.

A Conventional Empirical p Value Approach

We are interested in comparing the latent p value approach to a more conventional empirical p value assessment. To compute a conventional empirical p value we first calculate a genome-wide test statistic from the observed data. Using the TLOD as an example, we define the genome-wide test statistic as

$$T = \max_{1 \leq l \leq M} \left( \max_{\theta_l} \log \sum_{j=1}^n P(D|X_{ij}, \theta_l) / \sum_{j=1}^n P(D|X_{ij}, \frac{1}{2}) \right),$$

where $X_{ij} \sim P(x|G)$. $T$ is the maximum LOD score over all loci. We then generate $f$ replicate data sets denoted $g^d$ by random gene drop for each $d = 1, \ldots, f$ [16]. For each locus $1, \ldots, M$, we generate $X_{1j}^d, \ldots, X_{Mj}^d$ from $P(x|g^d)$ for $j = 1, \ldots, n$ using MCMC. Hence, the empirical null distribution for a genome-wide test statistic is

$$T^d = \max_{1 \leq l \leq M} \left( \max_{\theta_l} \log \sum_{j=1}^n P(D|X_{ij}^d, \theta_l) / \sum_{j=1}^n P(D|X_{ij}^d, \frac{1}{2}) \right),$$

for $d = 1, \ldots, f$. To get the conventional empirical p value, we compare $T$ with the distribution of $\{T^d\}$ such that

$$p = \frac{1}{f} \sum_d I\{T^d \geq T\},$$

where $I\{\cdot\}$ is one if the condition in the braces is true, and zero otherwise. The MCMC simulations required for the calculation are $(f + 1)n$. The conventional empirical p value for the het-TLOD and the C-link statistic are computed similarly.

Implementation

The extended structure of the pedigree resource provides a large sampling space that may require many samples to adequately approximate the distribution of inheritance patterns, and where it may be possible for the MCMC algorithm to remain confined to a local maximum for a large number of simulations. It is therefore important to demonstrate that the MCMC algorithm has adequate mixing properties for this data set, and to use an appropriate number of conditional samples to overcome any effect that the mixing properties of the algorithm may introduce. To identify how many simulations are sufficient to obtain adequate mixing, we generated 100 independent MCMC samples conditional on observed marker data using a randomly selected marker and our largest pedigree consisting of 97 individuals spanning six generations. To generate these simulations we started the MCMC algorithm with 100 distinct random seeds and collected 10,000 sequential inheritance states. We then calculated the TLOD statistic for each sampled state. We assessed the convergence of the distributions of the 100 independent samples by plotting the cumulative average at each consecutive inheritance state for each independent sample. Figure 1 shows a plot of the cumulative average of the 100 independent, sequentially generated samples. We noted that 2,000 sequential observations should be sufficient to overcome any effect of initial conditions, and that 10,000 sequential observations are sufficient to obtain convergence to a representative distribution.

Computer memory constraints during the calculation of test statistics created an effective upper limit of about 20,000 samples for this data set. Recognizing that a sample of 20,000 states is more than sufficient to achieve convergence to a representative distribution, we chose to use this upper limit of 20,000 to define $l$, the number of conditional inheritance states, and $m$, the number of unconditional inheritance states, for each marker and pedigree. We also included a burn of 2,000 states for the conditional observations to overcome any effects of initial conditions.

We recognized that sequentially generated samples may exhibit significant auto-correlation and thereby reduce our ability to obtain a representative sample of inheritance states. In order to guage the effect of auto-correlation in our sample, we determined the approximate number of independent inheritance states that would be equivalent to our sample of 20,000 sequentially generated inheritance states. First, using the largest pedigree at a randomly selected marker, we obtained a set of 20,000 sequentially generated inheritance states and calculated the TLOD statistic for each state. Second, we selected the 10,000th inheritance state from 100 sequentially generated samples, each starting from a distinct random seed. We considered these 100 observations as independent. We then calculated the TLOD statistic for each of the 100 independent states. We calculated the relative efficiency of the sequential samples as the ratio of the estimated variance of the independent samples to the estimated variance of the sequential samples. The relative efficiency of the sequential samples was 0.26, indicating that about 4 sequential MCMC samples are approximately equivalent to one independent sample, and that a sample of 20,000 sequential inheritance states from the MCMC
An Implementation of the Latent p Value Method

algorithm is approximately equivalent to 5,000 independent samples. Therefore, we concluded that a sample of 20,000 sequentially generated inheritance states is sufficient to overcome any effect of auto-correlation between states.

For the calculation of the conventional empirical p value, we chose to generate \( f = 100 \) replicate data sets by randomized gene drop and sample \( n = 1,000 \) inheritance states after a burn of 1,000.

**Results**

We calculated the distribution of latent p-values under a general recessive model using the TLOD, het-TLOD and C-link statistics. Figure 2 shows the cumulative distribution functions (CDFs) of the latent p value for each of the statistics. The CDF of the C-link statistic was skewed to the left of the diagonal and had a median value of 0.03. Furthermore, 60% of the distribution of p values were less than 0.05 and the mode was zero, indicating that linked loci are detectable in the resource. However, the shape of the CDF also suggests that not all of the 20,000 test statistics were extreme values with respect to the distribution of unconditional observations, indicating some variability in the distribution of conditional test statistics for the C-link statistic. The fact that the distribution was centered on zero yet exhibited some variability from this value indicates that expansion of the resource has at least the potential to increase evidence for linkage.

In contrast to the C-link statistic, figure 2 shows that the distributions of the latent p values for the TLOD (median = 0.85) and het-TLOD (median = 0.12) do not deviate substantially from the diagonal. The fact that the distribution of the TLOD statistic is centered on 0.85 indi-
icates that interfamilial heterogeneity has a significant effect in this resource, making the TLOD statistic much less powerful to detect linkage. It is also apparent from figure 2 that the distribution of the latent p value for the het-TLOD statistic has greater power to detect linkage in the presence of interfamilial heterogeneity compared to the TLOD statistic, yet has less power than the C-link statistic, at least in this resource.

Figure 3 shows the location of the marker where the maximum C-link statistic occurred for each of the 20,000 genome-wide sets of sampled inheritance states for each combination of model and conditional or unconditional sampled inheritance states. For the conditional set of samples under the recessive model, the location of the maxima shows very little variability, where a few adjacent markers taken together provided the maximum test statistic over a genome-wide set in almost every attempt. Under the dominant model, some sharing is evident at different loci, but none of these are indicative of evidence for linkage given the low performance of the statistic under this model, as observed in the CDF of the latent p value for this model. The location of the maximum score under either model for the unconditional realizations generally follows the distribution of a uniform random variable, which is anticipated for a set of null observations. The locations of the maxima for the other test statistics that we analyzed are not shown, but exhibited similar features.

For the conventional empirical p value calculation, the estimated p value was less than 0.01 for the C-link statistic, consistent with the results obtained from our latent p value analysis.

Discussion

Our implementation of the latent p value method demonstrates the feasibility of the method for assessing the presence of linkage. One of the primary strengths of the latent p value method is decreased computational time relative to a conventional empirical assessment. Decreased computational time occurs in two ways. First, when sampling unconditional inheritance states, the removal of constraints to the MCMC algorithm in the form of allele calls reduces the MCMC algorithm to MC. The reduced algorithm produces unconditional inheritance states in approximately 10% of the time required of the constrained algorithm, at least for this resource and choice of algorithm. A conventional empirical assessment does not take advantage of this form of time reduction. Second, we demonstrated that the latent p value method is able to reliably achieve the null distribution of a test statistic with fewer simulations than are required by a conventional empirical approach. For example, in our implementation of the two methods, the latent p value method used 42,000 total sim-
ulations per pedigree and locus, whereas the conventional assessment used 202,000 total simulations per pedigree and locus (2,000 simulations per 100 replicate data sets plus the observed dataset). The reduction in simulations is a result of the fact that the latent p value method uses only one MCMC run plus one MC run, whereas a conventional p value calculation requires f MCMC runs. This outcome is all the more striking when considering that in a typical genome-wide scan relying solely on a conventional empirical assessment, 10,000 replicate data sets would be required to produce an empirical p value with four significant digits. Furthermore, the current trend toward ever-denser marker maps could incur an implicit limitation for a conventional empirical approach in the form of file storage (depending upon the implementation), whereas the latent p value approach does not possess this particular limitation.

Another strength of the latent p value approach is that it provides a means to capture the null distribution of test statistics with unknown asymptotic distributions by converting any particular test statistic to the p value scale. Hence, the method can provide a useful framework for comparing properties of different linkage statistics. For instance, in our analysis we were able to compare the distributions of three linkage statistics on the same scale. The strategy of converting test statistics to the p value scale should also make the latent p value method extendible to more complex types of analyses, such as those involving imprinting, multi-locus interactions, or optimizations of various modeling parameters.

The comparison of the three statistics that we included in our analysis led to several insights. First, the TLOD statistic, unadjusted for interfamilial heterogeneity, is not a reasonable option for analysis of complex diseases, at least when using more than one pedigree, given the high probability of encountering interfamilial heterogeneity in such settings. Second, the het-TLOD statistic does provide better power to detect linkage in the presence of interfamilial heterogeneity than the unadjusted TLOD. However, more powerful statistics are available. Our interest in implementing the latent p value method was partly oriented towards further interpretation of some results from a previous genome-wide scan of this resource that made use of published thresholds for interpreting the significance of linkage results. Our previous analysis provided statistically significant evidence for linkage to a region on chromosome 5q31 under a general recessive model with approximately 20% of the resource linked to this locus [8]. The maximum het-TLOD score occurring in our previous analysis exceeded the established threshold for claiming significant evidence for linkage, after adjusting for multiple tests. This threshold was based on the asymptotic distribution of a two-point LOD score, which our previous analysis assumed was also appropriate for the het-TLOD statistic [10]. The results of our current analysis, where the latent p value for the het-TLOD statistic had a median value of only 0.12, show that the conclusions of our previous analysis were incorrect and that the selected thresholds we used in that analysis were inappropriate for interpreting the het-TLOD statistic. Third, we discovered that the C-link statistic, which logically requires some evidence for sharing among affected individuals in order for a pedigree to contribute to the LOD score, was more powerful than the frequently encountered het-LOD score approach. The het-LOD statistic has been in wide use because it has a more tractable theoretical distribution (i.e., chi-squared with 1.5 degrees of freedom) than some of the other statistics proposed to account for interfamilial heterogeneity. For instance, the asymptotic distribution of the C-link statistic we propose here is resource-specific, being a function of the number and sizes of pedigrees being analyzed, and therefore requires empirical assessment. Despite this limitation, our results provide a clear example where the C-link statistic is more suitable for detecting linkage than a statistic based on the heterogeneity LOD score approach.

Despite the failure to replicate significance with the het-TLOD statistic, our results do provide evidence for linkage in this resource. The distribution of the latent p value that we observed for the C-link statistic indicates significant linkage under a general recessive model at approximately chromosome 5q13, as evidenced by the majority of observations being less than 0.05 on the p value scale, almost all of which occurred at this locus. Some variation in the distribution is present, which could likely be reduced by further expansion of the pedigree resource. In such a case, we would not expect the distribution to shift, but rather to center more closely on zero as variability in the distribution of conditional observations decreases.

**Conclusion**

We were able to demonstrate the feasibility of implementing the latent p value method for linkage and show that it has computational advantages over more conventional means of empirically assessing the presence of linkage. Additionally, we were able to demonstrate that the method provides a means to more clearly interpret findings from test statistics designed to account for link-
age heterogeneity. Of further benefit, the application of the latent p value approach to this data set provides strong evidence of linkage to asthma at chromosome 5q13, and therefore encourages further expansion of this resource.

Acknowledgements

C.T. was supported by a National Library of Medicine Training Grant to the University of Utah Department of Biomedical Informatics, 1T15LM007124. This work was supported in part by grants NIH R01 GM81417, and DOD W81XWH-07-01-0483 to A.T. The original pedigree collection, phenotyping, and genotyping activities were funded by a grant (to Lisa Cannon-Albright) through Intermountain Healthcare from Myriad Genetics (1995–2000). An allocation of computer time from the Center for High Performance Computing at the University of Utah is gratefully acknowledged. The computational resources for this project have been provided by the National Institutes of Health (grant no. NCRR 1 S10 RR17214-01) on the Arches Metacluster, administered by the University of Utah Center for High Performance Computing.

References

15 Christensen GB, Knight SS, Camp NJ: The sumLINK statistic for genetic linkage analysis in the presence of heterogeneity. Genet Epidemiol 2008;advance online publication.