Unifying Ideas for Non-Parametric Linkage Analysis

Aaron G. Day-Williams^a John Blangero^c Thomas D. Dyer^c Kenneth Lange^{a, b}
Eric M. Sobel^a

Departments of ^aHuman Genetics, and ^bBiomathematics, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, Calif., and ^cDepartment of Genetics, Texas Biomedical Research Institute, San Antonio, Tex., USA

Key Words

Non-parametric linkage • NPL • QTL • Kong and Cox adjustment • Perfect data

Abstract

Objectives: Non-parametric linkage analysis (NPL) exploits marker allele sharing among affected relatives to map genes influencing complex traits. Computational barriers force approximate analysis on large pedigrees and the adoption of a questionable perfect data assumption (PDA) in assigning p values. To improve NPL significance testing on large pedigrees, we examine the adverse consequences of missing data and PDA. We also introduce a novel statistic, Q-NPL, appropriate for NPL analysis of quantitative traits. **Methods:** Using simulated and real data sets with qualitative traits, we compare NPL analysis results for four testing procedures and various degrees of missing data. The simulated data sets vary from all nuclear families, to all large pedigrees, to a mix of pedigrees of different sizes. We implemented the Kong and Cox linear adjustment of p values in the software packages Mendel and SimWalk. We perform similar analysis with Q-NPL on quantitative traits of various heritabilities. **Results:** The Kong and Cox extension for significance testing is robust to realistic missing data patterns, greatly improves p values in approximate analyses, and works equally well for qualita-

tive and quantitative traits and small and large pedigrees. The Q-NPL statistic is robust to missing data and shows good power to detect linkage for quantitative traits with a wide spectrum of heritabilities. **Conclusions:** The Kong and Cox extension should be a standard tool for calculating NPL p values. It allows the combination of exact and estimated analyses into a single significance score. Q-NPL should be a standard statistic for NPL analysis of quantitative traits. The new statistics are implemented in Mendel and SimWalk.

Copyright © 2011 S. Karger AG, Basel

Introduction

Despite the recent successes of genome-wide association studies, most of the variation of the complex traits investigated remains unexplained [1]. A common hypothesis is that some of the missing genetic effects are from the collective action of numerous rare mutations in proximity. Since each rare mutation acts on relatively few individuals, these mutations will remain hidden to association studies for now, but their location might be exposed by well-designed linkage designs. For complex traits, linkage analysis based on allele sharing among affected relatives is still an attractive option. Whittemore and Halpern proposed a class of allele-sharing statistics

that do not require the specification of an inheritance model [2, 3]. Their ideas have been successfully implemented and currently constitute the core of non-parametric linkage (NPL) analysis [4–7]. In addition to the original Whittemore and Halpern omnibus NPL statistics, NPL_{Pairs} and NPL_{All} , other statistics are now available that may be more effective in mapping traits with common inheritance patterns [8, 9].

All of the traditional NPL statistics assume a dichotomous qualitative trait. The statistics operate by comparing the estimated and expected numbers of alleles two affected relatives share who are identical by descent (IBD) at a marker. Excess of allele sharing suggests linkage between the marker and the trait. The two major barriers to implementing NPL analysis are IBD imputation and determination of statistical significance. Marker-specific IBD scores can be determined exactly on small pedigrees. For large pedigrees, which often contain the majority of information in linkage studies, the computational problems are too difficult to solve exactly, and researchers resort to Markov chain Monte Carlo (MCMC) sampling [7, 10]. Traditionally, approximation of p values has invoked a perfect data assumption (PDA) of complete information (i.e. no missing genotypes and complete phase information). From its original definition, it has been known that the PDA simplification results in conservative p values in the presence of missing information [5]. When analyzing complex traits with late-onset phenotypes, many individuals at the top of multi-generation pedigrees are not available for genotyping, introducing missing data which impacts the determination of the IBD process. Many researchers have shown the limitations and loss of power in using PDA to determine significance [7, 11–16]. All of these investigations involved the analysis of PDA in exact analysis methods, but no one has shown the effects in the MCMC-based analysis of large pedigrees.

In 1997, Kong and Cox (K&C) proposed two one-parameter extensions (linear and exponential) to traditional NPL statistics that compensate for missing information. The K&C extensions were originally developed for exact NPL analysis, but we show that MCMC-based methods can accurately implement the K&C linear model. In large pedigrees with a considerable fraction of missing genotypes, the improvements in power are impressive. These successes have prompted us to make the K&C linear adjustment a standard option in SimWalk, our comprehensive package for statistical analysis of large pedigrees [7].

The original NPL method was limited to qualitative traits. However, quantitative traits contain more information and should be easier to map. We propose a new

test statistic, Q-NPL, applicable to quantitative traits. Although Q-NPL is motivated by standard variance component models, it avoids parameter estimation and dubious normality assumptions. We show that Q-NPL has a good type I error rate and reasonable power to map traits with a range of heritabilities.

The K&C linear extension serves to unify NPL analysis on small and large pedigrees. It is suitable for both the traditional qualitative NPL statistics and Q-NPL. In the subsequent sections, we illustrate the weaknesses of the PDA in MCMC-based methods, how the K&C linear extension with estimated parameters is robust to missing data, the favorable properties of the Q-NPL statistic for quantitative traits, and the ability of the K&C linear framework to unify exact and approximate NPL analyses for pedigrees of all sizes and shapes. Our supporting evidence comes from both simulated and real data.

Methods

Pedigree Configurations

For the sake of convenience, we use the same pedigree types A through E featured in Sobel et al.'s paper on IBD computation [17]. Our first data set consists of 51 type B pedigrees with 2 or 3 affected people per pedigree; each pedigree is a nuclear family with a total of 5 people. Our second data set consists of 14 type D pedigrees with 2, 3, or 4 affected people per pedigree; each pedigree contains 15 people spanning 3 generations. Our third data set consists of 17 type A pedigrees with between 2 and 10 affected people per pedigree; each pedigree contains 45 people spanning 4 generations. Our fourth data set consists of a mixture of various types: 1 pedigree of type A, 10 pedigrees of type B, 30 pedigrees of type C, 7 pedigrees of type D, and 2 inbred pedigrees of type E. The same data sets were used for the qualitative and quantitative investigations. The novel aspects of the qualitative NPL results are shown in data sets 3 and 4, and therefore they are the only results discussed. For the quantitative NPL discussion, we will discuss results for data sets 1 and 4.

Data Simulation

We simulated a 9-marker microsatellite map with 10 cM between markers, a typical density for genome-wide linkage scans. The allele frequencies for the 9 markers replicate the frequencies from 9 microsatellites on chromosome 16 used in a previous mapping study [18]. The genetic map begins with Marker1 at 0 cM and ends with Marker9 at 80 cM (table 1). We simulated a disease locus situated halfway between Marker4 and Marker5 at 35 cM. The disease locus has two alleles, allele 1 with frequency 0.8905 and allele 2 with frequency 0.1095.

For the qualitative NPL investigations, the assigned penetrances

$$\begin{array}{l|l} \text{Pr(affected)} & 1/1) = 0.01 \\ \text{Pr(affected)} & 1/2) = 0.45 \\ \text{Pr(affected)} & 2/2) = 0.90 \end{array}$$

Table 1. Simulated genetic map

Real marker	Simulated marker	Map position cM	Alleles n
D16S518	Marker1	0	7
D16S3096	Marker2	10	10
D16S516	Marker3	20	9
D16S3040	Marker4	30	7
	TRAIT	35	2
D16S507	Marker5	40	9
D16S505	Marker6	50	8
D16S3091	Marker7	60	10
D16S402	Marker8	70	12
D16S3061	Marker9	80	4

are nearly additive. Genotypes were simulated by gene dropping, assuming Hardy-Weinberg and linkage equilibrium. Individuals were randomly assigned an affection status based upon their genotype at the trait locus and the above model. We rejected pedigrees with fewer than 2 affected people or affected spouses external to the main body of the pedigree, and we continued sampling until the desired quota for each pedigree type was met. To make the simulation as real as possible, we deleted all marker genotypes for some pedigree members. We also deleted all genotypes at some markers. The missing data patterns conform to those seen in the 73 families of the previous study [18]. This sampling procedure generated a variety of realistic relationships among the affected.

For the quantitative NPL investigations, we modeled the quantitative trait as a mixture of three normal distributions with means -2.0 (genotype 1/1), 0.0 (genotype 1/2), and 2.0 (genotype 2/2). This gives the population mean

$$\mu = f_1^2\mu_{1/1} + 2f_1f_2\mu_{1/2} + f_2^2\mu_{2/2} = 1.56$$

and the additive genetic variance

$$\sigma_a^2 = f_1^2(\mu_{1/1} - \mu)^2 + 2f_1f_2(\mu_{1/2} - \mu)^2 + f_2^2(\mu_{2/2} - \mu)^2 = 0.7831.$$

The definition of trait heritability

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

allows one to solve for the environmental variance σ_e^2 . We analyzed simulated data sets with heritability equal to 10, 25 and 50%. For each data set and for each heritability, we assigned trait means to individuals in our simulated pedigrees based on their genotypes at the trait locus. To these means we added independent Gaussian errors centered at 0 with variance σ_e^2 . We introduced missing marker genotypes exactly as in our qualitative NPL simulations. Finally, we computed the Q-NPL statistic from equ. 1 for each replicate and stored its p value. These p values were tabulated under the null hypothesis of no linkage to determine the bias and type I error rate of the Q-NPL test.

Simulation of Pedigree-Specific Null Distributions

For the data sets containing only small pedigrees, we approximated the null distributions of the qualitative and quantitative

NPL scores by simulation. This entailed resampling each study pedigree repeatedly by gene dropping. Sampling continued until 10,000 independent replicates exactly reproduced the affection states of the original pedigree. This kind of conditional sampling is prohibitively expensive for large pedigrees. NPL statistics were calculated at each location along the marker map for each of the 10,000 replicates using the Mendel software package [6]. The resulting approximate null distributions were then used to generate p values for the NPL statistics across all pedigrees.

Variance Component Models

Variance component methods are powerful tools for mapping quantitative trait loci (QTL) [19–21]. In parametric QTL mapping, the value of a quantitative trait y_i for individual i is decomposed as the sum

$$y_i = \mu + \beta'z_i + a_i + g_i + e_i,$$

where μ is the population mean of the trait, z_i is a vector of environmental predictors, a_i is the polygenic effect, g_i is the major gene effect, and e_i is the residual error [21]. The variance of y_i is

$$\text{Var}(y_i) = \sigma_a^2 + \sigma_g^2 + \sigma_e^2,$$

and the covariance between the trait values y_i and y_j of two relatives i and j is

$$\text{Cov}(y_i, y_j) = 2\Phi_{ij}\sigma_a^2 + 2\hat{\Phi}_{ij}\sigma_g^2.$$

In these equations, σ_a^2 is the additive genetic variance, σ_g^2 is the major gene variance, and σ_e^2 is the environmental variance. For the pair i and j , Φ_{ij} is their theoretical kinship coefficient and $\hat{\Phi}_{ij}$ is their conditional kinship coefficient at the location of the major gene. This model assumes that all random contributions are independent, that dominance can be neglected, and that the trait vector $y = (y_i)$ over a pedigree is normally distributed. If we collect the variance and covariances into a single matrix Ω , then the various assumptions allow us to write the full log-likelihood as

$$S = -\frac{1}{2} \ln \det \Omega - \frac{1}{2} (y - \mu)' \Omega^{-1} (y - \mu) \quad (1)$$

and test the null hypothesis $\sigma_g^2 = 0$ by a likelihood ratio statistic.

One can raise several objections to the model. We will focus on the normality assumption. Many quantitative traits are far from normally distributed. For some traits, a transformation such as logarithm or square root may create a new trait that more closely resembles normality. For some other traits, their distribution may be just far enough from normal that the use of a variance components analysis based on the t distribution is justified. However, for many traits such as coronary calcification that are partly discrete and partly continuous, no transformation of trait values can be close to normally distributed. This dilemma has prompted us to rethink the problem of QTL mapping to build a non-parametric approach which is not dependent on even approximate normality.

Definition of the Q-NPL Statistic

Often parametric test statistics form the basis of hypothesis testing in non-parametric statistics. What changes is how p values are computed under the null hypothesis. This is often done by simulation techniques such as permutation or gene dropping. Be-

cause it takes many random samples to achieve much accuracy in approximating low p values, the modified test statistic should be easy to evaluate. The parametric test statistic (equ. 1) is amenable to this type of modification. To avoid some of the computational expense of maximum likelihood estimation, we regress trait values y_i on non-genetic predictors prior to analysis. Thus, our first step is to substitute regression-based standardized residuals for the entries of y and omit the trait mean μ from the model.

Using a fixed value of Ω is a less appealing simplification because a single Ω cannot possibly reflect gradations in marker allele sharing among relatives in different chromosome regions. The solution to this dilemma is to employ a different Ω for each descent graph at the disease locus. A descent graph, or inheritance vector, determines the gene flow in a pedigree by specifying a grand maternal or grand paternal source for every gamete passed. The probabilities of the different descent graphs at a given genome location can be computed conditional on the marker genotypes observed in the pedigree [7]. These conditional probabilities can in turn be used to compute the conditional expectation of an IBD scoring function. Finally, the conditional expectations from different pedigrees can be combined into a single NPL test statistic [8, 21].

The question now becomes one of proposing a simple random effects model for the inheritance of the trait conditional on a given descent graph. Such a model has to balance computational efficiency with power to detect linkage. Computational efficiency argues for a random effects model because we know only the pattern of gene flow, not what disease allele flows along each descent path. For simplicity, the random effects should also act additively. If there are f unrelated founders, then under an autosomal model there are $2f$ ancestral genes and consequently $2f$ random effects. Under Hardy-Weinberg and linkage equilibrium, these random effects are uncorrelated. These considerations prompt the model

$$\Omega = \sigma_a^2 \sum_{s=1}^{2f} u_s u_s' + \sigma_e^2 I, \quad (2)$$

where I is the identity matrix, σ_a^2 is the additive variance, σ_e^2 is the environmental variance, and the i -th component u_{si} of the vector u_s has value 0, $1/\sqrt{2}$, or $\sqrt{2}$ depending on whether person i inherits 0, 1, or 2 genes from founder source s .

Several assumptions are implicit in this model. One is the incorporation of random environment via the term $\sigma_e^2 I$. A second is that a non-inbred person has total variance $\sigma_a^2 + \sigma_e^2$. Because y is standardized, it is natural to assume $\sigma_a^2 + \sigma_e^2 = 1$ and define the heritability $h^2 = \sigma_a^2$. This convention allows us to rewrite equ. 2 as

$$\Omega = h^2 \sum_{s=1}^{2f} u_s u_s' + (1 - h^2) I. \quad (3)$$

A third assumption is the neglect of additive polygenic effects. This sounds like a serious omission, but it is worth recalling that we are engaged in hypothesis testing not parameter estimation. Although it is true that additive polygenic and major gene contributions will be confounded, we are looking for excessive contributions by a major gene. The average value of Ω in the absence of marker data captures polygenic inheritance. If a major gene is present and marker coverage is good, then the conditional probabilities of one or just a few descent graphs will dominate. For these descent graphs, we want a variance decomposition that favors contributions by the major locus. This is exactly what model 3 achieves.

Computational speed is another reason for preferring the model. Inspection of equ. 3 shows that Ω is a rank $2f$ perturbation of a diagonal matrix. This fact facilitates quick evaluation of $\det \Omega$ and Ω^{-1} . Ordinarily evaluation of these entities requires arithmetic operations on the order of n^3 when Ω is $n \times n$. However, the Sherman-Morrison algorithm brings these operation counts down to a multiple of n^2 [22]. In practice, we apply each of the two identities

$$(M + ww')^{-1} = M^{-1} - \frac{1}{1 + w' M^{-1} w} M^{-1} w w' M^{-1}$$

$$\det(M + ww') = \det M \det(1 + w' M^{-1} w)$$

$2f$ times, starting from $M = (1 - h^2)I$ and $\det M = (1 - h^2)^n$.

We have another reason for preferring the Q-NPL statistic, that is, equ. 1 under the decomposition (equ. 3). Suppose we consider this statistic with a generic variance matrix Σ substituted for Ω . Then on average, conditional on the given descent graph, the statistic has value

$$\begin{aligned} E(S) &= -\frac{1}{2} \ln \det \Sigma - \frac{1}{2} E \left[(y - \mu)' \Sigma^{-1} (y - \mu) \right] \\ &= -\frac{1}{2} \ln \det \Sigma - \frac{1}{2} \text{tr} \left\{ \Sigma^{-1} E \left[(y - \mu)(y - \mu)' \right] \right\} \\ &= -\frac{1}{2} \ln \det \Sigma - \frac{1}{2} \text{tr} (\Sigma^{-1} \Omega). \end{aligned}$$

Using standard results from multivariate normal theory, $E(S)$ is maximized by taking $\Sigma = \Omega$. Finally, we anticipate that the Q-NPL statistic derived from equ. 1 will perform well because it is closely related to the Mahalanobis statistic $(y - \mu)' \Omega^{-1} (y - \mu)$ prominent in hypothesis testing and outlier detection. Q-NPL is now implemented exactly in Mendel for small pedigrees and approximately in SimWalk for large pedigrees.

K&C NPL Extensions

For each NPL statistic S , the standardized score on pedigree i is defined by

$$\bar{Z}_i = \frac{\bar{S}_i - \mu_i}{\sigma_i},$$

where μ_i and σ_i^2 are the mean and variance of S on pedigree i under the null hypothesis of no linkage, and $\bar{S}_i$ is the conditional expectation $E(S_i | \text{data})$. The traditional test statistic over m pedigrees is then taken to be the sum $\sum_{i=1}^m \gamma_i \bar{Z}_i$, with weight γ_i assigned to pedigree i . To adjust for missing data, Kong and Cox (in 1997) constructed a class of empirical models depending on a single free parameter δ for the likelihood of the gene flow in each pedigree. In their linear model, the log-likelihood across all pedigrees is written as

$$L(\delta) = \sum_{i=1}^m \ln [1 + \delta \gamma_i \bar{Z}_i]. \quad (4)$$

Kong and Cox stress that this is the exact log-likelihood under any missing data pattern, including the case of no missing data. The latter assumption obviously makes all computations much easier. In testing for linkage, they restrict δ to be non-negative because a deficit of sharing is less remarkable than an excess of sharing. Under this restriction, they advocate estimating δ by maximizing

$L(\delta)$. The resulting test is robust to missing data. The K&C method can be motivated as an attempt to adjust for the over-estimation of the variances σ_i^2 under PDA.

For this approach to work, δ must be bounded above by $b = \min_i b_i$, where

$$b_i = \frac{\sigma_i}{\gamma_i(\mu_i - a_i)}$$

and a_i is the minimum value possible for S_i . The values for a_i , and thus b_i , can be determined exactly for small pedigrees by examining all possible gene-flow patterns in the pedigree. These gene-flow patterns are also called inheritance vectors or descent graphs. For large pedigrees, a_i and b_i can be estimated by randomly sampling a large number of the possible gene flow patterns. Once the maximum likelihood estimate $\hat{\delta}$ has been found, define the constant $c = \sqrt{2[L(\hat{\delta}) - L(0)]}$. When the number of pedigrees is large, the p value of the test $\delta = 0$ can be approximated by $1 - \Phi(c)$, where Φ is the standard normal distribution function. The K&C linear model for exact NPL calculations on small pedigrees is implemented in Merlin [4] and now in Mendel as well [6]. The K&C linear model is now implemented for approximate NPL analysis on large pedigrees in SimWalk [7].

The linear model has limitations when the number of pedigrees is small and the excess of allele sharing is high. To account for these situations, Kong and Cox developed the exponential model [14, 15, 23]. The exponential model also finds the maximum likelihood estimate $\hat{\delta}$, and evaluates the constant $c = \sqrt{2[L(\hat{\delta}) - L(0)]}$. But unlike the linear model, δ no longer has an upper bound and $L(\hat{\delta}) - L(0)$ is defined as:

$$L(\hat{\delta}) - L(0) = \ln \prod_i r_i(\delta) E_{H_0} [\exp(\delta Z_i) | \text{marker data}],$$

where $r_i(\delta)^{-1}$ is a normalization constant satisfying

$$r_i(\delta)^{-1} = E_{H_0} [\exp(\delta Z_i)].$$

The exponential model is computationally more intensive than the linear model because with missing data $L(\delta)$ cannot be written down solely on the basis of the conditional expectations $\bar{Z}_i$, like the linear model, instead requiring the entire conditional distributions of Z_i . Since investigating a small number of pedigrees with excessive sharing is not the expected study design or sharing expectation when investigating complex disease, we have based our unification strategy upon the linear model. Additionally, when the sample size is large and the excess of sharing (δ) is small, the two models are very similar [14, 23]. The exponential model is a non-default option in Merlin, and we use its results to support the fact that the exponential model is only better in the case of small numbers of pedigrees [4]. For a more detailed description of the K&C methods, we refer the reader to the following papers [14, 15].

p Value and Method Comparisons

We have described three methods for determining the significance of exact NPL scores in small pedigrees. The first method builds a highly accurate approximation to the null distribution of the score for each pedigree; earlier we mentioned replicating each pedigree 10,000 times. With much less effort, one can obtain nearly the same level of accuracy with many fewer replicates per pedigree. This is the replicate pool method implemented in Men-

del [8, 24, 25]. The second and even faster method employs the PDA. The PDA options in Genehunter, Mendel, and Merlin yield identical results [4–6]. (Mendel refers to this method as the full enumeration method.) Finally, the third method for small pedigrees invokes the K&C linear model. The K&C options implemented in Merlin and Mendel also give identical results.

For large pedigrees where approximation of NPL scores is necessary, p values can be estimated by either the PDA method or the K&C model. Both options are implemented in SimWalk. The replicate pool method is generally infeasible for large pedigrees. Another difference between the major software packages is pedigree weighting. The default weighting in Mendel and SimWalk for a given pedigree is $1/\sqrt{a}$, where a is the number of affected people in the pedigree. Other packages do not use pedigree weights.

For the qualitative simulated data sets containing a mixture of pedigree structures, we compared seven methods for estimating p values: (1) simulated null with 10,000 replicates, (2) exact PDA, (3) exact K&C linear, (4) exact K&C exponential, (5) SimWalk PDA, (6) SimWalk K&C, and (7) a combined K&C linear method that combines exact analysis on small pedigrees and estimation results on large pedigrees into a single test. In this data set we also include results from an exact PDA analysis to illustrate the fact that MCMC-based PDA is more susceptible to missing data than exact PDA. For data set 3, which contains only large pedigrees, we used only the two SimWalk procedures because they are the only methods considered able to analyze such large pedigrees. We compared the results under various missing data regimes, for both the NPL_{All} and NPL_{Pairs} , and with and without pedigree weighting. Both Mendel and SimWalk allow the user to specify that the pedigrees should not be weighted, and for the sake of accurate comparisons between Merlin, Mendel, and SimWalk, the pedigree weights were turned off.

For the quantitative simulated data sets, we compared three different non-parametric methods: (1) Q-NPL defined here, (2) Merlin QTNPL, and (3) Merlin-Regress. The Merlin QTNPL method generates a score for each inheritance vector by summing squared scores for each founder allele, where the score for each founder allele is calculated by simple mean deviates $(y_i - \mu)$ for all individuals who carry the founder allele [26]. Merlin QTNPL converts the scores into Z-scores as described by Whittemore and Halpern [2] and uses the K&C linear model to assess significance [14]. Merlin-Regress is a regression-based method motivated by the Haseman-Elston method that regresses multipoint IBD sharing on trait squared sums and squared differences among all pairs of relatives [27]. For the quantitative data sets we compared the following seven results: (1) simulated null with 10,000 replicates, (2) Mendel Q-NPL PDA, (3) Mendel Q-NPL K&C linear, (4) SimWalk Q-NPL PDA, (5) SimWalk Q-NPL K&C linear, (6) Merlin QTNPL, and (7) Merlin-Regress.

Results

Qualitative NPL Results

In our simulated data sets, the results for weighted versus unweighted statistics are virtually the same. For the sake of brevity, we will only present the unweighted results. We will likewise show only results for NPL_{All} statis-

Fig. 1. Data set 3 (large pedigrees): NPL results. **a** 0%, **b** 20%, **c** 30%, **d** 40% missing genotypes.

tics. The same trends and conclusions apply to NPL_{Pairs} statistics. We will focus our presentation of results on data sets 3, 4, and the real data set because they present novel analyses not discussed in previous discussions of PDA and the K&C method.

Figure 1 shows the results of analyzing the large pedigrees in data set 3. These very large 45-individual pedigrees have a tremendous amount of information as seen by the very significant p values when there are no missing genotypes. Even with the tremendous amount of information in the pedigrees, SimWalk PDA drastically loses the linkage signal with increasing amounts of missing data. With 40% missing data, SimWalk PDA just barely achieves a significant finding ($-\log_{10}(\text{p value}) = 2$) for the marker 5 cM away from the disease locus. Whereas with the SimWalk K&C linear model results, after an initial loss of significance with the introduction of 20% missing genotypes, the signal remains robust and consistent with increasing amounts of missing data. The missing data patterns were constructed by eliminating all genotypes for nine individuals at the top of the pedigree (20%), additionally eliminating a large fraction of genotypes for Marker5 and a small fraction of genotypes for Marker1

(30%), and finally eliminating all genotypes for more people at the top of the pedigree (40%).

Figure 2 displays analysis results for data set 4 which contains a mix of different pedigree types, and most closely resembles real data sets. Due to the computational limits of exact analyses, the p values for the simulated null, exact PDA, exact K&C linear, and exact K&C exponential scores were constructed ignoring the three largest pedigrees in the data set, but to investigate the benefits of MCMC-based methods, they were analyzed in the SimWalk analyses. For computational reasons, the minimum p value we entertain is 10^{-4} . For the simulated null and exact PDA, when the p value equaled zero due to the limitation of the sampling, we substituted 0.00009 which equals 4.0457 on the graphs. Figure 2a shows that this bound is reached for the simulated null and exact PDA with no missing values in data set 4, and that the simulated null also reaches this bound with 20% missing data. This figure nicely illustrates the superiority of the K&C methods over PDA for determining significance. It also clearly demonstrates that MCMC-based PDA is more susceptible to the limitations of PDA than exact methods, but that the implementation of the K&C linear model in an MCMC

Fig. 2. Data set 4 (mixed pedigrees): NPL results. **a** 0%, **b** 20%, **c** 30%, **d** 40% missing genotypes.

framework recovers the loss of power due to PDA. This figure also demonstrates the excellent framework the K&C linear model provides for unifying exact and MCMC-based NPL analysis. The combined K&C curve was generated by letting Mendel calculate Z-scores for 47 pedigrees and SimWalk for 3 pedigrees and performing a single K&C analysis. The combined K&C curve has the same characteristics as the exact K&C linear curve.

The results of these simulation studies are striking, but how relevant are they to real data? To investigate this question, we re-analyzed a published real data set previously subjected to NPL analysis with SimWalk under the PDA model [18]. The study sample consisted of 73 pedigrees containing from 4 to 59 people each. On average, 56% of the genotypes were missing, with a per-pedigree range of 25 to 94%. The high missing data rate is partly a consequence of combining three separate study samples into one. Of the nine markers investigated, one sample was genotyped at six markers, another at seven markers, and the third at eight markers. The separate samples have

missing data percentages of 38, 55, and 36%. To cut laboratory costs, only nuclear families were genotyped. Of course missing genotypes, regardless of their source, reduce the information content of pedigree data and adversely impact analysis. The missing data patterns in these pedigrees were taken as a model for introducing missing data in our simulated data sets.

Figure 3 displays the results of our re-analysis. The graphs in the figure beg comment. Bowing to computational reality, we were forced to discard 4 complex pedigrees in our exact re-analysis. These were retained in the approximate re-analysis and the combined analysis. Figure 3 shows that excluding these large pedigrees obscures a significant linkage signal at roughly 13 cM along the map. The figure reveals again that the SimWalk K&C method outperforms the SimWalk PDA method and that exact PDA performs worse than exact K&C. The analysis also demonstrates that the K&C exponential model is not inherently superior in the real analysis of complex traits. Finally, these graphs reveal that the combined exact and

Fig. 3. Re-analysis of Pajukanta et al.'s data [18].

approximate K&C analysis closely mirrors the approximate SimWalk K&C, but at a significant computational savings. The compute time for SimWalk to analyze the entire data set was ≈ 100 min, while the combined analysis took ≈ 32 min, a savings of approximately 68 min.

Quantitative NPL Results

The first step in evaluating a new NPL statistic is to examine its behavior under the null hypothesis of no linkage. We simulated 1,000 replicates of both data sets 1 and 2 under the null hypothesis with a quantitative trait with 50% heritability. The distribution of p values should be uniform, and figure 4 shows that the distribution of p values for Marker5 from data set 1 is approximately uniform. Figure 4b focuses on the left tail of the distribution with p values between 0 and 0.1, where false positives will occur. As the figure illustrates there are 48 replicates with p values ≤ 0.05 and 7 replicates with p values ≤ 0.01 out of the 1,000 replicates tested. For the analysis undertaken in this project, we considered a significant finding as a p value of ≤ 0.01 , therefore 0.7% of the replicates would have yielded a false positive. All markers under all conditions investigated showed a similar behavior. Additionally, we investigated the distribution of the score statistic itself across all pedigrees. For data set 1, the mean of the score statistic distribution for Marker5 is 0.0018 with a standard deviation of 0.9368, and for data set 2, the mean of the score distribution for Marker5 is -0.0007 with a standard deviation of 0.9406. The p value and score distributions illustrate that our method performs correctly under the null hypothesis.

Fig. 4. Q-NPL p value distribution using data set 1 (small pedigrees), Marker5, and 50% heritability. **a** Distribution between 0 and 1, **b** distribution between 0 and 0.1.

To assess the power of Q-NPL, we used data set 1 comprised of 51 nuclear families. We simulated the quantitative trait with heritability equal to 10, 25, and 50%. The Q-NPL statistic was not able to detect a significant linkage signal at 10% heritability (data not shown), but it did detect linkage with heritabilities of 25% (fig. 5) and 50% (fig. 6). There are several striking features revealed in these figures. The first being that the new Q-NPL defined here performs better than the Merlin QTNPL method under all circumstances. Second is that the exact implementation of the new Q-NPL method is robust against missing data both using PDA and the K&C linear model. Additionally, the Mendel Q-NPL K&C is as powerful as Merlin-Regress in the analysis of a trait with 25% heritability. Interestingly, both of Mendel's implementations are more powerful than Merlin-Regress with a trait of 50% heritability. The figure also reveals that SimWalk's implementation of the new Q-NPL method struggles with a trait of 25% heritability, but the SimWalk Q-NPL K&C performs well and even better than Merlin-Regress with a trait of 50% heritability. Furthermore, the K&C results closely mirror the results from the simulated null method at considerably less expense, with the simulated

Fig. 5. Data set 1 (nuclear families): Q-NPL results using 25% heritability. **a** 0%, **b** 20%, **c** 30%, **d** 40% missing genotypes.

Fig. 6. Data set 1 (nuclear families): Q-NPL results using 50% heritability. **a** 0%, **b** 20%, **c** 30%, **d** 40% missing genotypes.

null requiring ~ 4 h and the Mendel K&C analysis ~ 5 s for the 25% heritability analysis.

For the mixed pedigrees of data set 4, the three large pedigrees included are too complex to be handled by exact PDA, exact K&C, or the simulated null procedures, so the results for those methods exclude the three pedigrees. Data set 4 was again analyzed with heritabilities of 10% (data not shown), 25% (fig. 7), and 50% (fig. 8). With 10% heritability, we detected a linkage signal only when no missing genotypes were present. As the figures illustrate, the Q-NPL statistic is able to detect a linkage signal for both higher heritabilities under all imposed missing data patterns. With a heritability of 25%, Merlin-Regress is clearly the most powerful statistic. Of the remaining statistics tested, the combined Q-NPL K&C that combines scores from Mendel and SimWalk into a single K&C test is the best statistic and is robust to the missing data patterns. With a heritability of 50%, the linkage signal is very strong and all methods find a significant signal under all situations. In fact the signal is so strong that the simulated null and exact PDA significance testing return p values of 0 for markers under all missing data patterns. The Merlin QTNPL performs very well in figure 8. One striking feature in both simulations is that the SimWalk PDA method for assessing significance is the most susceptible to missing data, as with the qualitative investigations.

As with all new statistics, it is their performance on real data sets that is important. We tested Q-NPL and the other methods on a real quantitative trait collected as part of a study of Mexican-Americans by the Texas Biomedical Research Institute (TBRI). These data harbor a linkage signal picked up by traditional QTL analysis [28]. The study sample contains 1,835 individuals spread over 46 extended pedigrees. Sample individuals were genotyped at 26 microsatellites spanning 179 cM, with an average of 7 cM between markers. A majority of the 35.5% missing genotypes come from 517 completely untyped individuals. Our Q-NPL analysis relies on standardized residuals corrected for age, sex, age², age $\times$ sex, and age² $\times$ sex. Standard variance components analysis estimates a heritability of about 60%. We accordingly analyzed the data set with QTL heritability specified at 60%. Of the 46 pedigrees, only 13, 15, and 13 were small enough to be analyzed by Mendel Q-NPL, Merlin-Regress, and Merlin QTNPL, respectively. Figure 9 displays the results of the analysis. The most extreme p value was found by the combined Q-NPL K&C analysis with 13 pedigrees being analyzed by Mendel and 33 pedigrees being analyzed by SimWalk. The p value peak occurs at the same location

as the peak found by parametric QTL analysis. The best non-parametric p value is less impressive than the best parametric p value, but this is to be expected. If the parametric model is correct, parametric linkage will always be at least as powerful as non-parametric linkage, and for large pedigrees usually more powerful. This analysis also illustrates the importance of MCMC-based methods in the analysis of large pedigrees, because large pedigrees hold most of the information for linkage.

We have illustrated that our method works well for microsatellite-based studies, but many researchers have moved to single nucleotide polymorphism (SNP)-based linkage studies. SNPs have less information than microsatellites due to reduce heterozygosity, therefore much denser marker maps are required. The smaller inter-marker distances pose challenges for MCMC-based methods. To address our method's performance on SNP data, we re-analyzed the TBRI data set utilizing SNP genotypes from the Illumina Human Hap550 array in place of microsatellite genotypes. The Illumina Human Linkage V panel has 101 SNPs that span chromosome 6 where the microsatellite signal resides. To generate a linkage panel of SNPs from the TBRI data, we identified the SNPs on the Hap550 array that are closest to the SNPs in the Linkage V array. This process generated a SNP map of 101 SNPs with an average spacing of 1.89 cM that exactly reproduces the cM positions and spacing of the Linkage panel. Only 807 of the 1,835 individuals had SNP genotype data thereby generating very sparse pedigrees, which is known to drastically increase compute times and reduce the accuracy of the IBD estimation. Therefore we utilized the pedigree trimming function in Mendel to remove uninformative individuals from the pedigrees and reconstruct core pedigrees. Pedigree trimming resulted in transforming the original 46 pedigrees into 64 pedigrees, of which 20 were singletons and therefore uninformative in the analysis. Of the 807 genotyped individuals, the missing genotype percentage of the 101 SNPs analyzed was 0.8%, showing the extremely high success rate of modern SNP arrays. We performed the same analyses on the SNP data and reconfigured pedigrees as for the microsatellite data, including performing traditional QTL analysis, and the results are shown in figure 10. The first noticeable difference is that the reduced sample size reduces the traditional QTL peak from ≈ 13 to ≈ 8 , and also reduces the signal for the NPL methods. This analysis shows that our new Q-NPL method extends well to the analysis of SNPs. This analysis also further supports the need to analyze large pedigrees with MCMC-based methods, with the SimWalk

Fig. 7. Data set 4 (mixed pedigrees): Q-NPL results using 25% heritability. **a** 0%, **b** 20%, **c** 30%, **d** 40% missing genotypes.

Fig. 8. Data set 4 (mixed pedigrees): Q-NPL results using 50% heritability. **a** 0%, **b** 20%, **c** 30%, **d** 40% missing genotypes.

Fig. 9. QTL and Q-NPL analysis of a real data set.

Fig. 10. QTL and Q-NPL analysis of a real data set with SNP data.

and combined Mendel & SimWalk analyses finding significant evidence of linkage. Of the 44 informative pedigrees only 20 were small enough to be analyzed by Mendel, Merlin, and Merlin-Regress, and figure 10 illustrates that excluding the 24 large pedigrees from analysis removes most of the linkage signal. But analyzing large

pedigrees typed on a dense SNP map with MCMC-based methods comes with a huge computational cost because the Markov chain must run for a considerably longer time to converge with the small inter-marker distances. The SimWalk analysis of all 44 informative pedigrees required ≈ 30 days and 19 h of compute time to analyze.

Allowing Mendel to analyze the 20 pedigrees amenable to exact analysis and SimWalk to analyze the 24 large pedigrees, the compute time was reduced to ≈ 27 days 18 h, a savings of 3 days of compute time.

Discussion

We have compared five methods for testing the significance of NPL scores for qualitative traits. Three of these methods are designed for exact analyses of small pedigrees, and two for approximate analyses of large pedigrees. We were particularly interested in what effect (a) the degree of missing data; (b) the pedigree weights; (c) approximate versus exact calculations, and (d) the choice of NPL_{All} or NPL_{Pairs} statistic would have on statistical power. For the samples we studied, we found almost no differences driven by pedigree weights. Approximate statistics were very similar to exact statistics, and the choice of NPL_{All} or NPL_{Pairs} did not alter any statistical conclusions. In contrast, we found some of the testing procedures were quite sensitive to the degree of missing data, particularly SimWalk's implementation of the perfect data approximation. With 30% or more missing data, the SimWalk PDA method could miss linkage signals. With the K&C adjustment, SimWalk often recovers these signals. Because 30% missing data levels are common in real, family-based, microsatellite data sets, we recommend that researchers consider re-analyzing data from previous failed NPL studies with substantial missing data that were analyzed with SimWalk. The replicate pool method [8, 21, 24] is an alternative to the K&C adjustment. Our analysis shows that in applications of the replicate pool method, the pools do not have to be terribly deep to give good results. Wigginton and Abecasis [25]

came to similar conclusions about this computation saving tactic.

We have proposed and evaluated a new test statistic, Q-NPL, for linkage analysis with a quantitative trait. Q-NPL is motivated by a simple random effects model with fixed heritability. The test statistic looks for excess contributions by a major gene and requires no parameter estimation. It is computationally efficient and applicable to general pedigrees. In our experience, it works well in exact analyses on small pedigrees and approximate analyses on large pedigrees. It also fits within the framework of the one-parameter K&C linear adjustment. Properly implemented, Q-NPL correctly controls type I error, displays a good ability to detect linkage, and is robust to missing genotypes, particularly in its K&C implementation.

Finally, we found that the K&C model allows straightforward combining of exact and estimated NPL scores. Since $\hat{\delta}$ is based solely on standardized NPL scores through equ. 4, these can be combined across exact and approximate analyses, and an overall $\hat{\delta}$ calculated and tested. This procedure works for almost any NPL statistic, not just NPL_{All} and Q-NPL. In practice, it speeds up computations, alleviates the conservative nature of PDA, and reclaims as much information as possible in the total sample.

Family-based linkage analyses are complementary to case-control association studies. They both find genes that influence traits, but they employ different strategies. Common variants inherited from ancient mutations are amenable to population-based association, but rare variants of recent vintage are more easily found through family-based linkage. The unified NPL analysis procedure we propose here will make many linkage studies easier and more productive.

References

- Altshuler D, Daly M: Guilt beyond a reasonable doubt. *Nature Genet* 2007;39:813–815.
- Whittemore A, Halpern J: A class of tests for linkage using affected pedigree members. *Biometrics* 1994;50:118–127.
- Whittemore A, Halpern J: Probability of gene identity by descent: computation and applications. *Biometrics* 1994;50:109–117.
- Abecasis G, et al: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30:97–101.
- Kruglyak L, et al: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;58:1347–1363.
- Lange K, et al: Mendel version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet* 2001;69(suppl): A1886.
- Sobel E, Lange K: Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996;58:1323–1337.
- Lange E, Lange K: Powerful allele sharing statistics for nonparametric linkage analysis. *Hum Hered* 2004;57:49–58.
- McPeck M: Optimal allele sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol* 1999;16:225–249.
- Heath S, et al: Segregation and linkage analysis using MCMC. *Genet Epidemiol* 1997;14: 1011–1016.
- Badner J, Gershon E, Goldin L: Optimal ascertainment strategies to detect linkage to common disease alleles. *Am J Hum Genet* 1998;63:880–888.

- 12 Cordell H: Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. *Am J Hum Genet* 2004;74:1294–1302.
- 13 Jung J, Weeks D, Feingold E: Gene-dropping vs. empirical variance estimation for allele-sharing linkage statistics. *Genet Epidemiol* 2006;30:652–665.
- 14 Kong A, Cox N: Allele-sharing models: Lod scores and accurate linkage tests. *Am J Hum Genet* 1997;61:1179–1188.
- 15 Nicolae D, Kong A: Measuring the relative information in allele sharing linkage studies. *Biometrics* 2004;60:368–375.
- 16 Schork N, Greenwood T: Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. *Am J Hum Genet* 2004;74:306–316.
- 17 Sobel E, et al: Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered* 2001;52:121–131.
- 18 Pajukanta P, et al: Combined analysis of genome scans of Dutch and Finnish families reveals a susceptibility locus for high-density lipoprotein cholesterol on chromosome 16q. *Am J Hum Genet* 2003;72:903–917.
- 19 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998;62:1198–1211.
- 20 Hopper J, Mathews J: Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* 1982;46:373–383.
- 21 Lange K: *Mathematical and Statistical Methods for Genetic Analysis*, ed 2. New York, Springer, 2002.
- 22 Millar K: *Some Eclectic Matrix Theory*. Malabar, Robert E Krieger Publishing, 1987.
- 23 Gudbjartsson D, et al: Fast multipoint linkage analysis and the program allegro. Technical Report. Reykjavik, DeCode Genetics, 2000.
- 24 Song K, et al: Efficient simulation of p values for linkage analysis. *Genet Epidemiol* 2004;24:1–9.
- 25 Wigginton J, Abecasis G: Evaluation of the replicate pool method: quick estimation of genome-wide linkage peak p values. *Genet Epidemiol* 2006;30:320–332.
- 26 Abecasis G: Non-parametric statistics for quantitative traits. www.sph.umich.edu/csg/abecasis/Merlin/reference/qtl.html.
- 27 Sham P, et al: Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet* 2002;71:238–253.
- 28 Goring H, et al: Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 2007;39:1208–1216.