Published in final edited form as:

Genet Epidemiol. 2008 July ; 32(5): 464–475. doi:10.1002/gepi.20319.

Searching for Epistasis and Linkage Heterogeneity by Correlations of Pedigree-Specific Linkage Scores

Daniel J. Schaid^{1,*}, Shannon K. McDonnell¹, Erin E. Carlson¹, Stephen N. Thibodeau², Janet L. Stanford^{3,4}, and Elaine A. Ostrander⁵

¹*Division of Biostatistics, Mayo Clinic College of Medicine, Rochester, Minnesota* ²*Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota* ³*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington* ⁴*Department of Epidemiology, School of Public Health and Community Medicine, University of Washington, Seattle, Washington* ⁵*Cancer Genetics Branch, NHGRI, National Institutes of Health, Bethesda, Maryland*

Abstract

Recognizing that multiple genes are likely responsible for common complex traits, statistical methods are needed to rapidly screen for either interacting genes or locus heterogeneity in genetic linkage data. To achieve this, some investigators have proposed examining the correlation of pedigree linkage scores between pairs of chromosomal regions, because large positive correlations suggest interacting loci and large negative correlations suggest locus heterogeneity (Cox et al. [1999]; Maclean et al. [1993]). However, the statistical significance of these extreme correlations has been difficult to determine due to the autocorrelation of linkage scores along chromosomes. In this study, we provide novel solutions to this problem by using results from random field theory, combined with simulations to determine the null correlation for syntenic loci. Simulations illustrate that our new methods control the Type-I error rates, so that one can avoid the extremely conservative Bonferroni correction, as well as the extremely time-consuming permutational method to compute *P*-values for non-syntenic loci. Application of these methods to prostate cancer linkage studies illustrates interpretation of results and provides insights into the impact of marker information content on the resulting statistical correlations, and ultimately the asymptotic *P*-values.

Keywords

epistasis; extreme values; gene-gene interaction; locus heterogeneity

INTRODUCTION

Besides environmental factors, multiple interacting genes are likely the cause of complex traits, as suggested by linkage studies in yeast [Brem et al., 2005; Storey et al., 2005], experimental crosses [Carlborg et al., 2003; Frankel and Schork, 1996; Shimomura et al., 2001], studies of human families [Bell et al., 2006], as well as genetic association studies in humans [Hsueh et al., 2001; Naber et al., 2000; Ukkola et al., 2001; Williams et al., 2000; Yanchina et al., 2004]. In fact, it has been hypothesized that epistasis is a major factor for human trait variation when the main effects of the contributing loci cannot be detected [Culverhouse et al., 2002]. It has also been proposed as a major cause for failure to replicate linkage studies.

*Correspondence to: Daniel J. Schaid, Harwick 775, Section of Biostatistics, Mayo Clinic, 200 First Street, SW, Rochester, Minnesota 55905. E-mail: E-mail: schaid@mayo.edu.

Although likelihood methods have been developed to simultaneously search for linkage with two interacting loci [Cordell et al., 1995, 2000; Dizier and Clerget-Darpoux, 1986; Dupuis et al., 1995; Farrall, 1997; Knapp et al., 1994; Olson, 1997; Schork et al., 1993; Tang and Siegmund, 2002], they are computationally demanding, the threshold for statistical significance can be challenging to determine (although progress has been made [Bell et al., 2006; Schaid et al., 2007]) and, except for simple pedigree structures like affected sib pairs [Bell et al., 2006; Farrall, 1997], they are restricted to non-syntenic loci.

An alternative model-free approach that can capture some forms of interaction between loci, as well as locus heterogeneity, is the correlation of pedigree-specific linkage statistics between different chromosomes or between different regions of the same chromosome; positive correlation suggests interaction and negative correlation suggests locus heterogeneity [Cox et al., 1999; Maclean et al., 1993]. Although simple to implement, it can be computationally challenging to use random permutations of pedigrees to determine the statistical significance of extreme positive or negative correlations when examining all pairs of non-syntenic loci, and no methods have been developed for extreme correlations of different regions within the same chromosome. To overcome these limitations, we develop new methods to determine the statistical significance of extreme positive or negative correlations based on random field theory. To illustrate the utility and limitations of our methods, we perform limited simulations and apply them to two large studies of genetic linkage for prostate cancer susceptibility.

METHODS

CROSS-CORRELATIONS BETWEEN TWO DIFFERENT CHROMOSOMES

We shall focus on the correlation of pedigree nonparametric standardized linkage statistics (NPLs) that are computed by major software packages [e.g., Allegro, Gudbjartsson et al., 2000, 2005; Merlin, Abecasis et al., 2002; Genehunter, Kruglyak et al., 1996; Markianos et al., 2001]. Although we consider all pairs of chromosomes, for now we consider only two different chromosomes, arbitrarily labeled 1 and 2, to simplify notation. If the linkage statistics are computed at p_1 positions on chromosome-1 and p_2 positions on chromosome-2, then we consider the $p_1 \times p_2$ matrix of correlation coefficients, denoted as R . Under the null hypothesis, the expected value of R is zero, and we wish to test whether the most extreme positive or negative values of matrix R depart significantly from the null distribution. If the values in R were independent, we could use the Bonferroni correction to control the testing of many values in R . However, because the values in R are correlated, with high correlations for scores with close positions on a chromosome, we need to account for this when evaluating the distribution of its extreme values.

To describe the theoretical developments, let $X(s)$ denote a random multipoint NPL score at position s on chromosome-1 and let $Y(t)$ denote a similar score at position t on chromosome-2. In practice, $X(s)$ and $Y(t)$ are computed at increments, often at 1 cM steps, so we use i_1 to index the points along chromosome-1 and j_2 for that of chromosome-2. Hence, we use $\rho_{i_1, j_2} = \text{Cor}[X(i_1), Y(j_2)]$ to denote the correlation of NPL scores at position i_1 of chromosome-1 and position j_2 of chromosome-2; r_{i_1, j_2} denotes the analogous sample correlation.

Because the sample correlation is not robust to outliers, and we rely on Gaussian random field theory to compute P -values, we replace the observed NPL scores with their van der Waerden normal scores, $v = \Phi^{-1}(\text{rank}(\text{NPL})/[n + 1])$, by ranking the n pedigree scores at each chromosome position, and using the inverse cumulative normal distribution, $\Phi^{-1}()$. This is useful when pedigrees have different structures and information content, so that the focus is on the joint *relative ranking* of linkage scores at two chromosome regions.

To test the null hypothesis that all correlations in R are zero, we consider extremely large positive correlations and extremely large negative correlations. The theory below is outlined for extremely large positive correlations, noting that the same theory can be used for extremely large negative correlations by simply using the absolute value of the negative correlation. To test whether a single correlation, $\rho_{i_1 j_2}$, is equal to zero, we could use either an exact test based on the t -statistic,

$$t(r_{i_1, j_2}) = \frac{r_{i_1, j_2}}{\sqrt{1 - r_{i_1, j_2}^2}} \sqrt{n - 2},$$

which has a t distribution with $n-2$ degrees of freedom, or Fisher's transformation,

$$z(r_{i_1, j_2}) = \log \left(\frac{1 + r_{i_1, j_2}}{1 - r_{i_1, j_2}} \right) \sqrt{n - 3},$$

which has an asymptotic standard normal distribution; similar P -values result for a large number of pedigrees.

Although both statistics can be used to test the null hypothesis that $\rho_{i_1 j_2} = 0$, only $z(r_{i_1 j_2})$ can be extended to test non-zero values of $\rho_{i_1 j_2}$ —this will be important when we consider correlations between different segments within a single chromosome. Furthermore, the theoretical distributions of the extreme values of $z(r_{i_1 j_2})$ and $t(r_{i_1 j_2})$ have been derived, but, as discussed below, the calculations for the extreme values of $t(r_{i_1 j_2})$ have been made more precise through more intricate developments. Because we later consider correlations between segments within a single chromosome, and need to account for non-zero null values of $\rho_{i_1 j_2}$, we need to rely on $z(r_{i_1 j_2})$ statistics. But, before doing so, we want to be sure that the theoretical distribution of the extreme value of the $z(r_{i_1 j_2})$ statistics is adequate. To evaluate this, we compare the P -value for the extreme value of $t(r_{i_1 j_2})$ with the P -value for the extreme value of $z(r_{i_1 j_2})$ and perform simulations.

To compute a P -value for a pair of chromosomes, we need to approximate $P(T_{\max} > t_{\max} | H_0)$, where $T_{\max}$ is a random extreme correlation statistic and $t_{\max}$ is the observed statistic. To approximate $P(T_{\max} > t_{\max} | H_0)$, we use results from random field theory [Adler, 1976; Adler and Hasofer, 1976; Cao and Worsley, 1999] that account for spatial correlations in multidimensional space when determining the probability of observing an extreme statistic. For some practical issues and an application to three-dimensional brain imaging, see Worsley et al. [1992]. We wish to approximate the distributions of the maximum t -statistic,

$$t_{\max} = \max_{i_1, j_2} t(r_{i_1, j_2})$$

and the maximum z -statistic

$$z_{\max} = \max_{i_1, j_2} z(r_{i_1, j_2}).$$

Cao and Worsley [see also Worsley et al., 1998] emphasize that the matrix of the $t(r_{i_1 j_2})$ values does not have a standard two-dimension t -field, but rather has a slightly “rougher” field due to the random variation in each of the dimensions. That is, each element of the matrix has a t -distribution, and each row of this matrix has a one-dimensional t -field because by conditioning

on the random row variable the corresponding elements in the row meet the criteria for a t -field. However, by allowing variation across the rows, the resulting two-dimensional field has greater variation than would exist if the row variables were constant (similar arguments can be made for columns having a one-dimension t -field). The theoretical implications are that the resulting distribution of $t_{\max}$ differs slightly from that of a standard t -field, so that the resulting P -values are slightly larger (and critical thresholds are slightly larger). Although this might occur for the distribution of $z_{\max}$, we nonetheless use standard Gaussian field theory to compute its approximate P -value, but then compare the P -values of $t_{\max}$ and $z_{\max}$, as a way to assess whether a large sample size will smooth over the roughness of the field to allow us to use standard Gaussian field theory when we consider correlations of linkage statistics between regions within a single chromosome.

CROSS-CORRELATIONS AND DISTRIBUTION OF $T_{\max}$

Using results from Cao and Worsley [1999, Lemma 3.1 on p 1027, Theorem 5.1 on p 1033, and Corollary 5.2 on p 1034], it can be shown that the P -value for $t_{\max}$ can be approximated by

$$P(T_{\max} > t_{\max}) \approx f_{00}(r_{\max}) + L_1 f_{10}(r_{\max}) + L_2 f_{01}(r_{\max}) + L_1 L_2 f_{11}(r_{\max}),$$

where $r_{\max}$ is the maximum value in the R matrix, L_i is the length of chromosome i in cM, and the f functions are given below:

$$\begin{aligned} f_{00}(r) &= \int_0^1 \frac{\Gamma(n/2)}{r \pi^{1/2} \Gamma((n-1)/2)} (1-u^2)^{(n-3)/2} \partial u, \\ f_{01}(r) &= \sqrt{v_y} (2\pi)^{-1} (1-r^2)^{(n-2)/2}, \\ f_{10}(r) &= \sqrt{v_x} (2\pi)^{-1} (1-r^2)^{(n-2)/2}, \\ f_{11}(r) &= \sqrt{v_x} \sqrt{v_y} \frac{\Gamma((n-1)/2)}{2\pi^{3/2} \Gamma((n-2)/2)} r (1-r^2)^{(n-3)/2}. \end{aligned}$$

For these f functions, the terms v_x and v_y are the variances of the first derivatives of the components of $X(s)$ and $Y(t)$. These variances determine the smoothness of the search space, such that larger values imply a rougher space, which translate to larger f values and a corresponding larger P -value. More insights about the smoothness of the search space are provided below for the distribution of $z_{\max}$.

To estimate v_x , we let $X_k(i_1)$ denote the van der Waerden normal scores for pedigree k on chromosome-1 at position i_1 , we use the numerical estimate of the first derivative,

$$\partial X_k(s) = [X_k(i_1 + \delta) - X_k(i_1)] / \delta,$$

we compute the variance of this derivative over all pedigrees at position i_1 , and then average this estimate over all positions for chromosome-1. In practice, we use $\delta = 1$ cM; v_y is computed in a similar fashion for chromosome-2. As we demonstrate later, these variance estimates can change over chromosomes, depending on the information content and autocorrelation within chromosomes.

CROSS-CORRELATIONS AND DISTRIBUTION OF $Z_{\max}$

By Gaussian random field theory, the null probability that the random variable $Z_{\max}$ over a standardized two-dimensional Gaussian field is greater than the observed $z_{\max}$ is

$$P(Z_{\max} > z_{\max}) \approx L_1 L_2 |\Lambda|^{1/2} (2\pi)^{-3/2} z_{\max} e^{-z_{\max}^2/2}, \quad (1)$$

where $L_1 L_2$ is the search area, assumed to be large relative to the smoothness of the space, and Λ is the 2×2 covariance matrix of the partial derivatives of the random field in each of the two dimensions (i.e., $\partial Z(s, t)/\partial s$ and $\partial Z(s, t)/\partial t$), which measures the smoothness of the space. Because $|\Lambda|^{1/2}$ is used to scale the search area in expression (1), we refer to $|\Lambda|^{1/2}$ as the “area scale.” To approximate the first partial derivatives, we consider the change in Fisher's z -statistic as points change along the rows of the Z matrix (the matrix of z -statistics computed on the R matrix),

$$\partial Z(s, t) / \partial s = [Z(r_{i_1+\delta, j_2}) - Z(r_{i_1, j_2})] / \delta$$

and along the columns of the Z matrix,

$$\partial Z(s, t) / \partial t = [Z(r_{i_1, j_2+\delta}) - Z(r_{i_1, j_2})] / \delta.$$

Although we could use a sample variance to estimate $\text{Var}(\partial Z(s, t)/\partial s)$, we use the delta method to gain greater insights about the smoothness of the search space, which also allows computation of $\text{Var}(\partial Z(s, t)/\partial s)$ under the null hypothesis. To do so, we need

$$\text{Var}(\partial Z(s, t) / \partial s) = \frac{\text{Var}[Z(r_{i_1+\delta, j_2})] + \text{Var}[Z(r_{i_1, j_2})] - 2\text{Cov}[Z(r_{i_1+\delta, j_2}), Z(r_{i_1, j_2})]}{\delta^2}. \quad (2)$$

Because the elements of Z are standard normal variables, each having variance 1, the covariance term is derived by applying the delta method to Fisher's z -transformation,

$$\text{Cov}[Z(r_{i,j}), Z(r_{k,l})] = (n-3) \frac{\text{Cov}(r_{i,j}, r_{k,l})}{(1-r_{i,j}^2)(1-r_{k,l}^2)}.$$

Now, by the multivariate central limit theorem, it can be shown that the covariance of two sample correlation coefficients, $\text{Cov}(r_{i,j}, r_{k,l})$, depends on fourth-order moments of the distribution [see Steiger and Browne, 1984, for general formulas]. Because the computation of fourth-order moments is cumbersome and time consuming, we simplify by assuming multivariate normality of the transformed NPL scores, in which case the covariance depends only on pair-wise correlations,

$$\begin{aligned} \text{Cov}(r_{i,j}, r_{k,l}) = & \left[\frac{1}{2} \rho_{i,j} \rho_{k,l} (\rho_{i,k}^2 + \rho_{i,l}^2 + \rho_{j,k}^2 + \rho_{j,l}^2) \right. \\ & + \rho_{i,k} \rho_{j,l} + \rho_{i,l} \rho_{j,k} - (\rho_{i,j} \rho_{i,k} \rho_{i,l} \\ & + \rho_{j,i} \rho_{j,k} \rho_{j,l} + \rho_{k,i} \rho_{k,j} \rho_{k,l} \\ & \left. + \rho_{l,i} \rho_{l,j} \rho_{l,k}) \right] / (n-3), \end{aligned}$$

see Olkin and Finn [1995]. To apply this to our situation, we carefully keep track of chromosomes, because the R matrix that has the correlations between chromosome-1 (rows)

and chromosome-2 (columns) is not symmetric. By using subscripts to keep track of chromosomes, we extend this covariance to our situation as

$$\begin{aligned} \text{Cov} \left(r_{i_1, j_2}, r_{k_1, l_2} \right) \\ = \left[\frac{1}{2} \rho_{i_1, j_2} \rho_{k_1, l_2} \left(\rho_{i_1, k_1}^2 + \rho_{i_1, l_2}^2 + \rho_{k_1, j_2}^2 + \rho_{j_2, l_2}^2 \right) \right. \\ + \rho_{i_1, k_1} \rho_{j_2, l_2} + \rho_{i_1, l_2} \rho_{k_1, j_2} - \left(\rho_{i_1, j_2} \rho_{i_1, k_1} \rho_{i_1, l_2} \right. \\ + \rho_{i_1, j_2} \rho_{k_1, j_2} \rho_{j_2, l_2} + \rho_{i_1, k_1} \rho_{k_1, j_2} \rho_{k_1, l_2} \\ \left. \left. + \rho_{i_1, l_2} \rho_{l_2, j_2} \rho_{k_1, l_2} \right) \right] / (n-3), \end{aligned}$$

and replace ρ with sample correlations r to estimate this covariance. Hence, $\text{Var}(\partial Z(s, t)/\partial s)$ around the point $s = i_1, t = j_2$ can be estimated by

$$\text{Var} \left(\frac{\partial Z(s, t)}{\partial s} \right) \approx \frac{2}{\delta^2} \left\{ 1 - \text{cov} \left[Z(r_{i_1 + \delta, j_2}), Z(r_{i_1, j_2}) \right] \right\},$$

where

$$\text{cov} \left[Z(r_{i_1 + \delta, j_2}), Z(r_{i_1, j_2}) \right] = \frac{\text{Cov}(r_{i_1 + \delta, j_2}, r_{i_1, j_2})}{(1 - r_{i_1 + \delta, j_2}^2)(1 - r_{i_1, j_2}^2)}.$$

By symmetry,

$$\text{Var} \left(\frac{\partial Z(s, t)}{\partial t} \right) \approx \frac{2}{\delta^2} \left\{ 1 - \text{cov} \left[Z(r_{i_1, j_2 + \delta}), Z(r_{i_1, j_2}) \right] \right\}.$$

Finally, using the methods outlined above,

$$\begin{aligned} \text{Cov} \left(\frac{\partial Z(s, t)}{\partial s}, \frac{\partial Z(s, t)}{\partial t} \right) \\ \approx \frac{1}{\delta^2} \left\{ \begin{array}{l} 1 - \text{cov} \left[Z(r_{i_1 + \delta, j_2}), Z(r_{i_1, j_2}) \right] \\ - \text{cov} \left[Z(r_{i_1, j_2 + \delta}), Z(r_{i_1, j_2}) \right] \\ + \text{cov} \left[Z(r_{i_1 + \delta, j_2}), Z(r_{i_1, j_2 + \delta}) \right] \end{array} \right\} \end{aligned}$$

One of the conditions for the validity of the approximation $P(z_{\max} > z_{\max})$ in expression (1) is that the distribution of the elements in Z can be modeled by a smooth homogeneous Gaussian random field with zero expectation and unit variance, with homogeneity implying that the local correlation structure around any point in Z is the same for all points in Z . Under this condition, we compute the local covariance structure for all points in Z , and then average over all these points to compute Λ . See Worsley et al. [1992] for more details on empirical local correlation structures for three-dimensional brain image data.

Although we use the above method to estimate Λ , insights into the above formulas can be gained by noting that under the null hypothesis, all correlations between chromosome-1 and chromosome-2 are zero, and only the within-chromosome autocorrelations make contributions:

$$\begin{aligned}
\text{Var}\left(\frac{\partial Z(s,t)}{\partial s}\right) &\approx \frac{2}{\delta^2} \left\{1 - \rho_{i_1+\delta, j_1}\right\}, \\
\text{Var}\left(\frac{\partial Z(s,t)}{\partial t}\right) &\approx \frac{2}{\delta^2} \left\{1 - \rho_{j_2+\delta, j_2}\right\}, \\
\text{Cov}\left(\frac{\partial Z(s,t)}{\partial s}, \frac{\partial Z(s,t)}{\partial t}\right) &\approx \frac{1}{\delta^2} \left\{1 - \rho_{i_1+\delta, j_1} - \rho_{j_2+\delta, j_2} + \rho_{i_1+\delta, i_1} \rho_{j_2+\delta, j_2}\right\}.
\end{aligned}$$

These simplifications illustrate that as the autocorrelations increase, the variances and covariance of these partial derivatives decrease, emphasizing that smoothness of the search space is determined by the autocorrelation. Substituting our estimate of Λ into expression (1) gives a way to compute the P -value for $z_{\max}$. Likewise, critical values can be determined from expression (1) by finding z such that $P(Z_{\max} > z) = \alpha$ for a specified Type-I error rate of α .

CORRELATIONS WITHIN A CHROMOSOME

We now consider the correlation of pedigree linkage statistics when a chromosome is paired with itself. For example, for chromosome-1 we consider the $p_1 \times p_1$ symmetric matrix of correlation coefficients. In contrast to correlations between different chromosomes, the expected value of R is not zero under the null hypothesis of no epistasis and no linkage heterogeneity. Suppose for now that we can determine the null expected value of R and denote this matrix by Ω . Applying Fisher's z statistic to the element r_{i_1, j_1} in the R matrix, it is well known that $z(r_{i_1, j_1}) = \log[(1 + r_{i_1, j_1})/(1 - r_{i_1, j_1})]$ has an asymptotic normal distribution with mean $\mu(\rho_{i_1, j_1}) = \log[(1 + \rho_{i_1, j_1})/(1 - \rho_{i_1, j_1})]$ (where ρ_{i_1, j_1} is an element from matrix Ω) and variance $1/(n-3)$. Hence, a standard normal statistic to test departures from the null is

$$w_{i_1, j_1} = \left[z(r_{i_1, j_1}) - \mu(\rho_{i_1, j_1}) \right] \sqrt{n-3}.$$

We can use the formulas in the previous section that are based on Gaussian random field theory by simply replacing j_2 for chromosome-2 with j_1 for chromosome-1. Furthermore, because R is a symmetric matrix, with 1s down the diagonal, some adjustments are needed for the search area. When considering two different chromosomes, the size of the search areas is $L_1 L_2$. When considering a single chromosome, we only use values in R that are above the diagonal, so the maximum search area size is $L_1(L_1-1)/2$.

Our simulations, however, suggest improvements by restricting the search area even smaller. When the distance between two points on a chromosome is small, say less than 10–20 cM, P -values from the Gaussian approximation can be inaccurate due to the high autocorrelation under the null hypothesis. Because there is little power to detect departures from the high correlations in these regions, it seems prudent to reduce the search area to points that are at least s cM apart (i.e., $|i_1 - j_1| \geq s$); we use simulations to guide the choice of s . In this case, the search area size is reduced to $(L_1 - s)(L_1 - s + 1)/2$.

ESTIMATING THE NULL CORRELATION

To estimate the null correlation matrix (Ω) for each chromosome, we use a simulation-based method. It is easy to simulate pedigree-specific NPL scores under the null hypothesis of no linkage, yet it is time consuming to perform a large number of simulations per pedigree across all chromosomes. To reduce the computational burden, we explored the use of a bootstrap replicate pool, as others have proposed [Leal and Ott, 1993; Song et al., 2004; Terwilliger and Ott, 1992]. This method requires two steps: (1) simulate a relatively small number (b) of replicates per pedigree and perform the necessary computer-intensive calculations to compute NPL scores per replicate, so that each pedigree will have a small pool of replicate NPL scores

for each chromosome and (2) randomly sample a pedigree-specific NPL score from each pedigree's replicate pool, and then perform the necessary calculations on the entire random sample. Initially, we used this approach to estimate both Λ for all chromosome pairs (both within and between chromosomes) and the null correlation matrix, Ω . Yet, we found very little change in the estimates of Λ over the bootstrap samples, and that the average over bootstraps differed little from the estimate of Λ based on the observed data. For this reason, we used the estimate of Λ based on the observed data. Furthermore, to obtain the most precise estimates, the estimate of Ω for each chromosome was based on the simple average over all simulations (b replicates $\times n$ pedigrees).

Although we performed our simulations under the null of no linkage, in fact the null hypothesis of no epistasis and no linkage heterogeneity for loci on the same chromosome do not necessarily mean that there is no linkage of a single disease susceptibility locus (DSL) on the chromosome. Nonetheless, we took our approach because we postulate that the presence of a single DSL on a chromosome has little impact on the autocorrelation of the NPL scores, particularly for genetically complex traits. Although this is difficult to prove when a sample is composed of different types of pedigrees, it is possible to examine the effect of a single DSL for the special case of a pedigree with a single affected sib pair (ASP). As illustrated in the Appendix, the correlation of the estimated fraction of alleles shared identical-by-descent (IBD) for an ASP at positions s and t , $\text{Cor}(\pi(s), \pi(t))$, depends on the distance between the marker loci, their distances to the trait-locus, and the magnitude of genetic effect.

The important question for our analysis is how much the autocorrelation, $\text{Cor}(\pi(s), \pi(t))$, depends on the magnitude the genetic effect, parameterized as the recurrence risk ratio λ . To illustrate this, we plot the autocorrelation in Figure 1 as position s moves along the x -axis (for a chromosome of length 100 cM), and allowing position t to be at $s + \text{gap}$, where gap was 1, 5, or 10 cM. Furthermore, we allowed λ to have values of 1 (no genetic effect), 2, 4 (moderate effect), and 1,000 (large effect). Figure 1 illustrates that in the vicinity of the DSL, the autocorrelation $\text{Cor}(\pi(s), \pi(t))$ decreases, and the decrease becomes greater as the marker spacing between s and t increases, or as the trait-locus effect, λ , increases. However, the effect of λ on $\text{Cor}(\pi(s), \pi(t))$ is quite small, even for very large values of λ . This suggests that we can approximate $\text{Cor}(\pi(s), \pi(t))$ by assuming no genetic effect ($\lambda = 1$). In this case, $\text{Cor}(\pi(s), \pi(t)) = \exp(-0.04|s - t|)$ for ASP. For more general types of affected relative pairs, Feingold et al. [1993] showed that $\text{Cor}(\pi(s), \pi(t)) = \exp(-0.01\beta|s - t|)$, where 0.01 is the crossover rate per cM and β depends on the type of relationship; larger values imply weaker correlations for a fixed distance. Some values of β for different types of affected relative pairs (ARPs) are 2 for grandparent-grandchild; 4 for half-sibs and full-sibs; and 5 for uncle-nephew and 16/3 for first cousins.

SIMULATIONS

To evaluate the Type-I error rates of our proposed methods, we simulated 1,000 three-generation pedigrees with the structure illustrated in Figure 2. For each pedigree we simulated NPL scores for two chromosomes. One chromosome was 253 cM (e.g., representing the longest human chromosome-1) with 35 markers, spaced at approximately 10 cM intervals, and the other chromosome was 42 cM (e.g., representing chromosome-22 which is the shortest human chromosome) with 6 markers. The genetic marker maps for these chromosomes are identical to real data [Janer et al., 2003], although we assumed 10 equally frequent alleles at each marker. All simulations assumed no DSL, and were performed with the Merlin software [Abecasis et al., 2002]. We then randomly sampled without replacement, 200 pedigrees from this pool of 1,000 pedigrees. All statistical methods were applied to the sample of 200 pedigrees, and this process was repeated 10,000 times to compute the empirical Type-I error rates.

RESULTS

SIMULATIONS

For the between-chromosome correlations, z -max and t -max gave similar Type-I error rates, and both were close to the true nominal rates for the max statistics, yet slightly conservative for the min statistics (Table I). In fact, the P -values for the z - and t -statistics were quite close over all simulations (see Fig. 3).

For the within-chromosome correlations, the z -statistics were extremely liberal when the search area included short distances between points on a chromosome (see Table II). In contrast, Type-I error rates were slightly conservative for the longer chromosome-1 (253 cM) when the search area excluded points within 10 cM from each other. However, for the short chromosome-2 (42 cM), the Type-I error rates did not achieve their nominal levels until the search area excluded points within 20 cM from each other (although at the nominal 5% level, the z -statistics were still slightly liberal). This is likely due to the short total length of chromosome-2, so that correlations between points in the search area are not small. This can be seen in the contour plots of Figure 4 for the within-chromosome correlations for all 1,000 simulated pedigrees. Note that the within-chromosome correlations quickly approach zero after 50 cM on chromosome-1, yet chromosome-2 is only 42 cM in length, so for the shortest chromosomes, there will be residual correlations that can influence the nominal Type-I error rates. Nonetheless, our simulations suggest that by restricting the search area to distances at least 20 cM apart, the empirical Type-I error rates will be acceptable for small P -values, say less than 0.01.

APPLICATION TO PROSTATE CANCER LINKAGE SCANS

Our newly developed methods were applied to two different studies of linkage for familial prostate cancer. The Mayo Clinic study included 159 pedigrees with a total of 429 ARPs, with both microsatellite ($n = 397$) and single nucleotide polymorphism ($n = 4,376$) genetic markers, and so provided high information content, despite not having genotypes for the founders of the pedigrees [Cunningham et al., 2003; Schaid et al., 2004]. The *PROGRESS* study, based at the Fred Hutchinson Cancer Research Center, included 247 pedigrees with a total of 878 ARPs, with microsatellite markers ($n = 438$) [Janer et al., 2003]. Although both sets of pedigrees are similar with respect to the characteristics of the affected men, there are some important differences that influence the smoothness of the search space. The Mayo pedigrees have a much higher information content than the *PROGRESS* pedigrees because of the much denser microsatellite and single nucleotide polymorphism marker scan, despite the *PROGRESS* pedigrees having more genotyped subjects per pedigree. Details of pedigree sampling and genotyping can be found in the above mentioned published reports. Each study was approved by the Institutional Review Boards where samples were collected and genotyped.

MAYO CLINIC STUDY

For the application of our methods to the Mayo pedigrees, we also simulated, under no linkage, 20 replicates per pedigree for each chromosome to compare the estimates of the area scale ($|A|^{1/2}$) between the observed and simulated data, and to estimate the within-chromosome correlation structure. These simulations were performed using Merlin software, which uses the marker allele frequencies and the missing data pattern of the observed pedigrees. The area scale within chromosomes and between chromosomes was remarkably close between the observed and simulated data (not shown). These observations, combined with our observation that the coefficient of variation of the estimates of area scale was less than 6% over bootstrap replicates, suggest that estimating the area scale on the observed data should suffice.

The area scale based on the observed Mayo data, Figure 5, illustrates a trend for shorter chromosomes to have smaller area scales, due to their higher autocorrelation, and that the X chromosome tends to have much larger area scale. This likely has several causes: (1) the X chromosome is one of the longest; (2) the higher recombination rate in females versus males weakens the autocorrelation; and (3) the information content of the X chromosome tends to be less than that for autosomes. For example, seven pedigrees were completely non-informative for the X chromosome because they contained only affected paternal cousins. Similar patterns (not shown) were observed for the *PROGRESS* data. For these reasons, we computed the area scale separately for each pair of chromosomes (both within and between chromosomes).

The most extreme observed positive and negative correlations, and their null expectations, are presented in Table III, for pairs of different chromosomes and each chromosome paired with itself. Although some pairs suggest extreme correlations, none of the correlations were statistically significant after using the Bonferroni correction for 276 pairs of chromosomes.

PROGRESS STUDY

The most extreme correlations for the *PROGRESS* pedigrees are presented in Table IV. The magnitude of extreme correlations are similar to those observed for the Mayo pedigrees, and only the much less than expected correlation between two regions on chromosome-9 was statistically significant after correcting for testing multiple pairs of chromosomes. We also compared the results between the two data sets, and found that correlations between chromosomes 18 and 19 were extreme for both data sets, with maximum correlations of 0.26 for the *PROGRESS* data and 0.33 for the Mayo data (both with uncorrected P -value = 0.005). Although this might be an important observation worth attempting to replicate in other data sets, it was not statistically significant after correcting for the multiple tests.

MARKER DENSITY, AUTOCORRELATION, AND SMOOTHNESS OF SEARCH AREA

Insights about the influence of marker density can be gained by comparing results between the two studies. We found the estimated area scale parameters to be 18–167% smaller for the *PROGRESS* data than the Mayo data (86% of the between-chromosome area scales were at least 50% smaller for the *PROGRESS* pedigrees). This results from the *PROGRESS* pedigrees having less information content due to a sparser genetic map. When the marker map is sparse, multipoint IBD calculations tend to smooth the linkage statistics [Teng and Siegmund, 1998], which in turn causes a larger autocorrelation of the linkage statistics along the chromosomes. This increased autocorrelation translates to smaller values of the area scale parameters. Although a smaller area scale means a smaller P -value for a given value of the $z_{\max}$ statistic, the increase in autocorrelation makes it more difficult to obtain large values of $z_{\max}$.

Others have reported that when searching over a three-dimensional volume V , the P -values approximated by a Gaussian random field tend to be accurate for P -values < 0.10 when the scaled search area, $V|A|^{1/2}$, was at least 30 [Worsley et al., 1992]. In our applications, we found the scaled search area for correlations between different chromosomes to range 21–525 for the *PROGRESS* pedigrees and 91–1,808 for the Mayo pedigrees. For correlations within the same chromosome, the scaled search area ranged 4–257 for *PROGRESS* pedigrees and 17–798 for the Mayo pedigrees. The smaller values for the *PROGRESS* pedigrees resulted from a sparser marker density, and the smaller values for correlations within the same chromosome were for the shortest chromosomes (e.g., chromosomes 20–22), particularly because the search area was restricted to distances of at least 20 cM. Overall, we expect the asymptotic P -value approximations to work well for correlations between different chromosomes and for within-chromosome correlations for all but the shortest chromosomes.

DISCUSSION

Following the suggestions of others to use correlations of linkage statistics to screen for interactions among linked loci [Cox et al., 1999; Maclean et al., 1993], we derived the asymptotic null distribution of extreme positive and negative correlations of NPL statistics between pairs of chromosomes. Some advantages of this approach are (1) the statistics are rapid to compute for genome wide linkage scans; (2) they can be applied to pedigrees of arbitrary structure; (3) they do not depend on genetic models, making them robust to departures from model assumptions; (4) they simultaneously evaluate both extreme positive correlations (implying interactions) and extreme negative correlations (implying locus heterogeneity); and (5) they can be used to examine extreme correlations between regions on the same chromosome. Furthermore, because our methods are based on correlations, the results from Gaussian random fields are expected to hold when there is linkage to some chromosomes yet without correlations (i.e., under the null hypothesis of no epistasis and no linkage heterogeneity), mainly because the correlations implicitly account for marginal linkage (by subtracting the mean NPL in the covariances and dividing by the marginal standard deviations in the correlation). This means that our methods should prove useful for screening for extreme correlations whether or not marginal linkage is identified for some chromosomes. To meet the assumption of Gaussian distributions, we used van der Waerden normal scores, and handled ties by the usual approach of using the mean rank for tied observations. If there are many tied ranks, then this would reduce the variation in the linkage scores, yet should have little impact on the estimated correlations. We came to this conclusion by simulating bivariate normal data and truncating observations (resulting in integers ranging from -3 to $+3$) to create highly tied data; this had little influence on the estimated correlations.

Our applications to the Mayo and *PROGRESS* pedigrees illustrate that our methods adapt to the marker information content. The lesser the information content, the greater the smoothing by multipoint calculations, which in turn increases the autocorrelations along chromosomes, translating to smaller values of the scale for the two-dimensional search. Smaller scale implies less conservative corrections for multiple testing than the Bonferroni method, hence accounting for the correlated statistics. Because information content can vary across different chromosomes, we advocate computing the area scale separately for different pairs of chromosomes.

Despite the advantages of our approach there can be hidden disadvantages, in terms of reduced power, compared to a likelihood-based method that uses a genetic model that is close to the true underlying genetic architecture. Wang and Zhao [2007] emphasize that the power of competing tests for gene-gene interactions depends on the disease prevalence and the definitions of interaction, and that tests based on correlations of NPL scores tend to achieve greater power for rare diseases. For affected sib pairs, Hallgrimsdottir and Speed [2007] found that modeling the joint IBD at two loci with a two-locus triangle constraint gave greater power to detect linkage than using correlation of NPL scores. This finding should be cautiously interpreted, however, when considering the power of a method to detect linkage in the presence of interaction [the question addressed by Hallgrimsdottir and Speed, 2007] versus the power to detect interaction alone. The former considers a general model with both main effects and interaction, yet the latter approach focuses solely on interaction. It is not surprising that the correlation of NPL scores is a weak test for linkage in the presence of interaction, but its power to detect interaction, relative to a likelihood-based method, is not obvious. Furthermore, due to computational restrictions, Hallgrimsdottir and Speed [2007] did not perform a simultaneous search over all pairs of markers, but rather performed a conditional search, conditioning only on those markers that had a maximum lod score at least 1.5. Because others have cautioned about potential limitations of imposing IBD triangle constraints in the presence of gene-environment interaction [Dizier et al., 2000; Guo, 2000], it is not clear whether model-based

methods will always have greater power to detect linkage with interacting loci than using correlations of NPL scores, and whether the conclusions of Hallgrimsdottir and Speed [2007] will generalize to more general pedigrees with different types of affected relative pairs, particularly when considering complex gene-gene interactions. Furthermore, the two-locus model-based methods only focus on interaction, and not locus heterogeneity, such as the extreme negative correlation of linkage scores can detect. Nonetheless, Hallgrimsdottir and Speed [2007] provide additional insights. They found that two-locus tests can have substantial power to detect interacting loci if there is at least a moderate increase in allele sharing at one of the two loci. Conversely, power to detect interacting loci decreases as the marginal allele sharing at the two loci decreases, suggesting that it will be extremely difficult to detect interacting loci if neither locus alone shows an increase in allele sharing.

To explore how different genetic architectures influence the correlation of linkage statistics, we considered the simple setting of fully informative ASPs and determined expected correlations for the mean IBD sharing statistics. Because our findings are fairly intuitive, we merely outline our general assumptions and strategy, and provide intuitive explanations, without presenting all the mathematical details. For the i th ASP, the fraction of alleles shared IBD is $\pi_i = f_{i2} + f_{i1}/2$, where f_{ij} is the estimated probability of sharing j alleles IBD. Then, the

mean sharing statistic is $z = \sqrt{8n} \left(\bar{\pi} - 0.5 \right)$, which is equivalent to the NPL statistic in this situation. To evaluate how heterogeneity influences negative correlations, we assumed that a fraction p_1 of ASPs were linked to one chromosome, another fraction p_2 linked to a different chromosome, and the remaining fraction not linked. For a large number of ASPs, the correlation of the z -statistics for the two different chromosomes can be shown to be negative, with more extreme values as p_1 or p_2 increase. The most extreme value of -1 occurs when $p_1 = p_2 = 0.5$.

To evaluate how epistasis influences positive correlations, we used the two-locus IBD sharing probabilities derived by Cordell et al. [1995]. They showed that the joint IBD sharing can be expressed as epistatic additive and dominance variances of the penetrances. One should not confuse these variance terms with mode of inheritance. For a single locus, recessive inheritance implies a much larger dominance variance than additive variance, whereas dominant inheritance implies a much larger additive variance. Epistatic variances between locus-1 and locus-2 can be decomposed into additive-additive, additive-dominance, dominance-additive, and dominance-dominance. Using formulas from Cordell et al. [1995], we numerically calculated the expected correlations of the z -statistics. We found that when there were no main effects for each locus, only epistatic terms, the maximum correlation for a model with only additive-additive epistasis was 0.15. In contrast, a model with only dominance-dominance epistasis had correlation as high as 0.63. Including both additive-additive and dominance-dominance epistasis reduced the correlation to 0.26. In general, adding main effects to these models tended to reduce the correlations. We interpret these findings to mean that extreme large positive correlations are most likely to occur when each locus has little main effect and the mode of inheritance is jointly recessive for both loci, meaning that both loci are required for an increase in disease risk, and that risk increases when both loci are homozygous for the risk alleles. Although these numerical evaluations provide intuitive guidelines, it is important to recognize that the definition of epistasis (in our case, epistatic additive and dominance variances of the penetrances) influences power of our methods, as well as the rarity of the disease in the population [Wang and Zhao, 2007].

Our simulations to examine the Type-I error rates suggest that our methods provide reasonable approximations to the P -values. Although we have evaluated some simple conditions for when correlations of NPL statistics are expected to be extreme, it is not obvious how our methods will compare to likelihood methods that model two-locus penetrances or two-locus IBD sharing. Because there are a wide variety of genetic mechanisms to explore, including the

impact of environmental covariates, we plan to perform extensive simulations in future work to compare competing methods under a broad range of genetic and environmental architectures of complex traits. Nonetheless, we advocate the use of our methods because they are computationally rapid, and positive findings would provide strong guidance regarding the design of studies aimed at following-up linkage signals.

ACKNOWLEDGMENTS

This work was supported by Grant Numbers R01GM67768 (D.J.S.) from the National Institute of General Medical Sciences; R01CA72818 and R01CA15083 (S.N.T.); R01CA80122 (J.L.S.); CA78836 and CA90754 (E.A.O.) from the National Cancer Institute, with additional support for the *PROGRESS* study from the Prostate Cancer Foundation and the Fred Hutchinson Cancer Research Center. EAO is supported by the Intramural Program of the National Human Genome Research Institute. The content is solely the responsibilities of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences, the National Cancer Institute, or the National Institutes of Health.

The authors are particularly grateful to Keith Worsely, Ph.D. (Department of Mathematics and Statistics, McGill University, Montreal Quebec, Canada) for answering questions about random field theory, although any errors in understanding or presentation are the sole responsibilities of the authors. Also, we are grateful for a reviewer's advice to explore how correlations are influenced by different genetic architectures.

Contract grant sponsor: National Institute of General Medical Sciences; Contact grant number: R01GM67768; Contract grant sponsor: National Cancer Institute; Contract grant number: R01CA72818; R01CA15083; R01CA80122; CA78836; CA90754; Contact grant sponsors: Prostate Cancer Foundation; Fred Hutchinson Cancer Research Center; Intramural Program of the National Human Genome Research Institute.

APPENDIX

AUTOCORRELATION FOR ASP AND A SINGLE DSL

To determine the autocorrelation for an ASP, let $\pi(t)$ denote the estimated fraction of alleles shared IBD for an ASP at chromosome position t . This is estimated by $\pi(t) = f_2(t) + f_1(t)/2$, where $f_m(t)$ is the estimated probability of sharing m alleles IBD at position t , conditional on the observed marker data. Assuming the Haldane mapping function and allowing for a trait-locus at position τ , Liang et al. [2001] showed that

$$\text{Var}(\pi(t)) = \frac{1}{8} - \exp(-0.08|t - \tau|) \left(\frac{1}{8} - \frac{\text{Var}(S(\tau))}{4} \right),$$

where $\text{Var}(S(\tau))$ is the variance of the *number* of alleles shared IBD at the trait-locus; they also showed that the covariance of $\pi(s)$ and $\pi(t)$ is

$$\begin{aligned} \text{Cov}(\pi(s), \pi(t)) &= \frac{\exp(-0.04|s-t|)}{8} - \exp(-0.04|s-\tau|) \\ &\quad \times \exp(-0.04|t-\tau|) \left(\frac{1}{8} - \frac{\text{Var}(S(\tau))}{4} \right). \end{aligned}$$

When there is no dominance variation so that the sibling recurrence risk ratio, λ_S , is equal to the offspring recurrence risk ratio, λ_O , Liang et al. [2001] showed that

$$\text{Var}(S(\tau)) = \frac{1}{2} - \frac{(\lambda - 1)^2}{4\lambda^2}.$$

This illustrates that the variance of the number of alleles shared IBD at the trait-locus decreases as the genetic effect, λ , increases. Under the assumption of no dominance,

$$\begin{aligned}\text{Var}(\pi(t)) &= \frac{1}{8} - \exp(-0.08|t - \tau|) \left(\frac{(\lambda-1)^2}{16\lambda^2} \right), \\ \text{Cov}(\pi(s), \pi(t)) &= \frac{\exp(-0.04|s-t|)}{8} - \exp(-0.04|s-\tau|) \\ &\quad \times \exp(-0.04|t - \tau|) \left(\frac{(\lambda-1)^2}{16\lambda^2} \right).\end{aligned}$$

This emphasizes that the covariance of the process depends on the distance between the marker loci, their distances to the trait-locus, and the magnitude of genetic effect.

REFERENCES

- Abecasis G, Cherny S, Cookson W, Cardon L. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30:97–101. [PubMed: 11731797]
- Adler, R. The Geometry of Random Fields. Wiley; New York: 1976.
- Adler R, Hasofer A. Level crossings for random fields. *Ann Prob* 1976;4:1–12.
- Bell JT, Wallace C, Dobson R, Wiltshire S, Mein C, Pembroke J, Brown M, Clayton D, Samani N, Dominiczak A, Webster J, Lathrop GM, Connell J, Munroe P, Caulfield M, Farrall M. Two-dimensional genome-scan identifies novel epistatic loci for essential hypertension. *Hum Mol Genet* 2006;15:1365–1374. [PubMed: 16543358]
- Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 2005;436:701–703. [PubMed: 16079846]
- Cao J, Worsley KJ. The geometry of correlation fields with an application to functional connectivity of the brain. *Ann Appl Probability* 1999;9:1021–1057.
- Carlborg O, Kerje S, Schutz K, Jacobsson L, Jensen P, Andersson L. A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res* 2003;13:413–421. [PubMed: 12618372]
- Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type I diabetes. *Am J Hum Genet* 1995;57:920–934. [PubMed: 7573054]
- Cordell HJ, Wedig GC, Jacobs KB, Elston RC. Multilocus linkage tests based on affected relative pairs. *Am J Hum Genet* 2000;66:1273–1286. [PubMed: 10729111]
- Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat Genet* 1999;21:213–215. [PubMed: 9988276]
- Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 2002;70:461–471. [PubMed: 11791213] Epub 2002 Jan 08
- Cunningham JM, McDonnell SK, Marks A, Hebbring S, Anderson SA, Peterson BJ, Slager S, French A, Blute ML, Schaid DJ, Thibodeau SN. Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate* 2003;57:335–346. [PubMed: 14601030]
- Dizier MH, Clerget-Darpoux F. Two-disease locus model: sib pair method using information on both HLA and Gm. *Genet Epidemiol* 1986;3:343–356. [PubMed: 3781239]
- Dizier M, Quesneville H, Prum B, Selinger-Leneman H, Clerget-Darpoux F. The triangle test statistic (TTS): a test of genetic homogeneity using departure from the triangle constraints in IBD distribution among affected sib-pairs. *Ann Hum Genet* 2000;64:433–442. [PubMed: 11281281]
- Dupuis J, Brown P, Siegmund D. Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* 1995;140:843–856. [PubMed: 7498758]
- Farrall M. Affected sibpair linkage tests for multiple linked susceptibility genes. *Genet Epidemiol* 1997;14:103–115. [PubMed: 9129957]

- Feingold E, Brown PO, Siegmund D. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am J Hum Genet* 1993;53:234–251. [PubMed: 8317489]
- Frankel WN, Schork NJ. Who's afraid of epistasis? *Nat Genet* 1996;14:371–373. [PubMed: 8944011]
- Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* 2000;25:12–13. [PubMed: 10802644]
- Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsson A. Allegro version 2. *Nature Genet* 2005;37:1015–1016. [PubMed: 16195711]
- Guo S-W. Gene-environment interactions and the affected-sib-pair designs. *Hum Hered* 2000;50:271–285. [PubMed: 10878472]
- Hallgrimsdottir IB, Speed TP. The power of two-locus affected sib-pair linkage analysis to detect interacting disease loci. *Genet Epidemiol* 2007;32:84–88. [PubMed: 17654608]
- Hsueh WC, Cole SA, Shuldiner AR, Beamer BA, Blangero J, Hixson JE, MacCluer JW, Mitchell BD. Interactions between variants in the beta3-adrenergic receptor and peroxisome proliferator-activated receptor-gamma2 genes and obesity. *Diabetes Care* 2001;24:672–677. [PubMed: 11315829]
- Janer M, Friedrichsen DM, Stanford JL, Badzioch MD, Kolb S, Deutsch K, Peters MA, Goode EL, Welti R, DeFrance HB, Iwasaki L, Li S, Hood L, Ostrander EA, Jarvik GP. Genomic scan of 254 hereditary prostate cancer families. *Prostate* 2003;57:309–319. [PubMed: 14601027]
- Knapp M, Seuchter SA, Baur MP. Two-locus disease models with two marker loci: the power of affected-sib-pair tests. *Am J Hum Genet* 1994;55:1030–1041. [PubMed: 7977340]
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;58:1347–1363. [PubMed: 8651312]
- Leal S, Ott J. A bootstrap approach to estimating power for linkage heterogeneity. *Genet Epidemiol* 1993;10:465–470. [PubMed: 8314045]
- Liang KY, Chiu YF, Beaty TH. A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. *Hum Hered* 2001;51:64–78. [PubMed: 11096273]
- Maclean CJ, Sham PC, Kendler KS. Joint linkage of multiple loci for a complex disorder. *Am J Hum Genet* 1993;53:353–366. [PubMed: 8328454]
- Markianos K, Daly MJ, Kruglyak L. Efficient multipoint linkage analysis through reduction of inheritance space. *Am J Hum Genet* 2001;68:963–977. [PubMed: 11254453]
- Naber CK, Husing J, Wolfhard U, Erbel R, Siffert W. Interaction of the ACE D allele and the GNB3 825T allele in myocardial infarction. *Hypertension* 2000;36:986–989. [PubMed: 11116112]
- Olkin I, Finn J. Correlations redux. *Psychol Bull* 1995;118:155–164.
- Olson JM. Likelihood-based models for genetic linkage analysis using affected sib pairs. *Hum Hered* 1997;47:110–120. [PubMed: 9097092]
- Schaid D, Guenther J, Christensen G, Hebring S, Rosenow C, Hilker C, McDonnell S, Cunningham J, Slager S, Blute M, et al. Comparison of microsatellites versus single nucleotide polymorphisms by a genome linkage screen for prostate cancer susceptibility loci. *Am J Hum Genet* 2004;75:948–965. [PubMed: 15514889]
- Schaid DJ, McDonnell SK, Carlson EE, Thibodeau SN, Ostrander EA, Stanford JL. Affected relative pairs and simultaneous search for two-locus linkage in the presence of epistasis. *Genet Epidemiol* 2007;31:431–449. [PubMed: 17410530]
- Schork NJ, Boehnke M, Terwilliger JD, Ott J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am J Hum Genet* 1993;53:1127–1136. [PubMed: 8213836]
- Shimomura K, Low-Zeddies SS, King DP, Steeves TD, Whiteley A, Kushla J, Zemenides PD, Lin A, Vitaterna MH, Churchill GA, Takahashi JS. Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res* 2001;11:959–980. [PubMed: 11381025]
- Song KK, Weeks DE, Sobel E, Feingold E. Efficient simulation of P values for linkage analysis. *Genet Epidemiol* 2004;26:88–96. [PubMed: 14748008]
- Steiger J, Browne M. The comparison of interdependent correlations between optimal linear composites. *Psychometrika* 1984;49:11–24.
- Storey JD, Akey JM, Kruglyak L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* 2005;3:e267. [PubMed: 16035920]

- Tang HK, Siegmund D. Mapping multiple genes for quantitative or complex traits. *Genet Epidemiol* 2002;22:313–327. [PubMed: 11984864]
- Teng J, Siegmund D. Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics* 1998;54:1247–1265. [PubMed: 9883537]
- Terwilliger JD, Ott J. A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. *Cytogenet Cell Genet* 1992;59:142–144. [PubMed: 1737483]
- Ukkola O, Perusse L, Weisnagel SJ, Bergeron J, Despres JP, Rao DC, Bouchard C. Interactions among the glucocorticoid receptor, lipoprotein lipase, and adrenergic receptor genes and plasma insulin and lipid levels in the Quebec Family Study. *Metabolism* 2001;50:246–252. [PubMed: 11229437]
- Wang S, Zhao H. Sample size needed to detect gene-gene interactions using linkage analysis. *Ann Hum Genet* 2007;71:828–842. [PubMed: 17521308]
- Williams SM, Addy JH, Phillips JA 3rd, Dai M, Kpodonu J, Afful J, Jackson H, Joseph K, Eason F, Murray MM, Epperson P, Aduonum A, Wong LJ, Jose PA, Felder RA. Combinations of variations in multiple genes are associated with hypertension. *Hypertension* 2000;36:2–6. [PubMed: 10904004]
- Worsley K, Evans A, Marrett S, Neelin P. A three dimensional statistical analysis for CBF activation studies in the human brain. *J Cereb Blood Flow Metab* 1992;12:900–918. [PubMed: 1400644]
- Worsley KJ, Cao J, Paus T, Petrides M, Evans AC. Applications of random field theory to functional connectivity. *Hum Brain Mapp* 1998;6:364–367. [PubMed: 9788073]
- Yanchina ED, Ivchik TV, Shvarts EI, Kokosov AN, Khodzhayantz NE. Gene-gene interactions between glutathione-s transferase M1 and matrix metalloproteinase 9 in the formation of hereditary predisposition to chronic obstructive pulmonary disease. *Bull Exp Biol Med* 2004;137:64–66. [PubMed: 15085249]

Fig. 1.

Autocorrelation of fraction of alleles shared identical-by-descent for an affected sib pair [$\text{Cor}(\pi(s), \pi(t))$] as a function of marker spacing (gap = $|t-s|$) and their sliding distance to a trait locus at 50cM with genetic effect size λ (assuming a single locus with no dominance effect).

Fig. 2.
Pedigree structure for simulations.

Fig. 3.
P-values for *z*- and *t*-statistics over all simulations (transformed to $-\log_{10}$).

Fig. 4. Contours of within-chromosome correlations for all 1,000 simulated pedigrees for chromosome-1 (253 cM; panel A) and chromosome-2 (42 cM; panel B).

Fig. 5. Area scale ($|A|^{1/2}$) within chromosomes (upper left panel) and between pairs of different chromosomes (remaining panels) for Mayo observed data.

TABLE I

Type-I error rates for extreme statistics for between-chromosome correlations

Nominal error rate	<i>z</i> -max	<i>t</i> -max	<i>z</i> -min	<i>t</i> -min
0.05	0.0431	0.0382	0.0309	0.0267
0.01	0.0092	0.0082	0.0062	0.0053
0.001	0.0012	0.0009	0.0004	0.0004

TABLE II
Type-I error rates for extreme correlations within chromosomes

Within chromosome		Minimum distance in search area (s, cM)	z-Statistic	Nominal error rate		
				0.05	0.01	0.001
1	1	Max	0.7408	0.5521	0.3352	
		Min	0.9820	0.9036	0.6483	
	5	Max	0.1129	0.0444	0.0115	
		Min	0.0781	0.0218	0.0034	
	10	Max	0.0291	0.0077	0.0013	
		Min	0.0192	0.0037	0.0003	
	20	Max	0.0117	0.0017	0.0001	
		Min	0.0103	0.0015	0.0002	
2	1	Max	0.4337	0.2533	0.1236	
		Min	0.8831	0.6943	0.3833	
	5	Max	0.1475	0.0515	0.0122	
		Min	0.1521	0.0456	0.0080	
	10	Max	0.0945	0.0241	0.0033	
		Min	0.0783	0.0155	0.0020	
	20	Max	0.0849	0.0154	0.0011	
		Min	0.0685	0.0105	0.0001	

TABLE III
Most extreme correlations for Mayo data: between different chromosomes and within the same chromosome

Chromosome pair	Chrom-1	Chrom-2	Extreme correlation	P-value	
				Per pair	Bonferroni
Different	16	21	0.35 (null = 0.0)	0.0015	0.41
	4	21	-0.33 (null = 0.0)	0.0113	1.00
Same	X	X	0.76 (null = 0.64)	0.0161	1.00
	12	12	-0.16 (null = -0.028)	0.0017	0.47

TABLE IV
Most extreme correlations for *PROGRESS* data: between different chromosomes and within the same chromosome

Chromosome pair	Chrom-1	Chrom-2	Extreme correlation	P-value	
				Per pair	Bonferroni
Different	2	4	0.30 (null = 0.0)	0.0014	0.39
	6	17	-0.27 (null = 0.0)	0.0045	1.00
	21	21	0.77 (null = 0.67)	0.0004	0.11
Same	9	9	-0.07 (null = 0.004)	7.1e-07	0.0002