A Novel Framework for Sib Pair Linkage Analysis

G. David Poznik,[1] Katarzyna Adamska,[3] Xin Xu,[2] Andrzej S. Krolewski,[1] and John J. Rogus[1]

[1]Section on Genetics and Epidemiology, Joslin Diabetes Center, and [2]Department of Environmental Health, Harvard School of Public Health, Boston; and [3]Institute of Informatics, Jagiellonian University, Krakow, Poland

Sib pair linkage analysis of a dichotomous trait is a popular method for narrowing the search for genes that influence complex diseases. Although the pedigree structures are uncomplicated and the underlying genetic principles straightforward, a surprising degree of complexity is involved in implementing a sib pair study and interpreting the results. Ascertainment may be based on affected, discordant, or unaffected sib pairs, as well as on pairs defined by threshold values for quantitative traits, such as extreme discordant sib pairs. To optimize power, various domain restrictions and null hypotheses have been proposed for each of these designs, yielding a wide array of choices for the analyst. To begin, we systematically classify the major sources of discretion in sib pair linkage analysis. Then, we extend the work of Kruglyak and Lander (1995), to bring the various forms into a unified framework and to facilitate a more general approach to the analysis. Finally, we describe a new, freely available computer program, SPLAT (Sib Pair Linkage Analysis Testing), that can perform any sib pair statistical test currently in use, as well as any user-defined test yet to be proposed. SPLAT uses the expectation maximization algorithm to calculate maximum-likelihood estimates of sharing (subject to user-specified conditions) and then plots LOD scores versus chromosomal position. It includes a novel grid-scanning capability that enables simultaneous visualization of multiple test statistics. This can lead to further insight into the genetic basis of the disease process under consideration. In addition, phenotype definitions can be modified without the recalculation of inheritance vectors, thereby providing considerable flexibility for exploratory analysis. The application of SPLAT will be illustrated with data from studies on the genetics of diabetic nephropathy.

Sib pair analysis is a powerful linkage technique based on an elegantly straightforward principle: pairs of phenotypically similar sibs will tend toward excess sharing of relevant chromosomal regions, whereas those that are dissimilar will tend toward lower sharing. Therefore, a set of sib pairs is typically ascertained such that either all pairs are concordant or all are discordant for a given trait. Then, at some region of interest or at loci across the genome, the degree of intrapair genetic similarity in the study population is assessed. This is done by estimating the *sharing pattern*—(z_0, z_1, z_2), the probabilities that a typical pair shares zero, one, or two alleles identical by descent (IBD)—and by evaluating statistical significance via likelihood theory.

Although the pedigree structures are uncomplicated and the problem is simply stated, sib pair linkage analysis turns out to be surprisingly complex. Beyond basic design issues, such as the type of sib pairs involved in the study, much of the ongoing debate focuses on two issues: how best to restrict the parameter domain and which null/alternative hypotheses to use. As for domain restriction, the baseline set of acceptable values comes from the fact that the sharing pattern is a 2-df probability vector, with z_2 fixed, given a suitable (z_0, z_1) pairing. On the z_0z_1-plane, this point must fall within the triangle bound by $\{z_0, z_1\}\geqslant 0$ and $z_0 + z_1 \leqslant 1$. However, to increase power, it is often desirable to further limit the domain of sharing patterns by imposing additional restrictions. Within the specified domain, regions corresponding to the null and alternative hypothesis must be chosen from among a growing collection of legitimate contenders. Other choices, including the method for handling multiple sibs per family, must also be made. Permutation of these various options leads to a wide array of possible analytic approaches.

Our intent in investigating such facets of sib pair analysis stems from our desire to gain control, both theoretically and practically, of the plethora of options available when searching for linkage. To accomplish this, we set out to generalize the framework of Kruglyak and Lander (1995) to encompass the broadest possible set of pairwise sib pair linkage tests. Our goal was not only to capture all variations currently described but also to allow for emerging developments in the field. This unified sib pair linkage framework served as the motivation for our software program, SPLAT (Sib Pair Linkage Analysis Testing), which was designed to handle any sib pair statistical test currently in use, as well as any user-defined test yet to be proposed. SPLAT also includes features to make sib pair analysis more efficient and more revealing.

Received July 11, 2005; accepted for publication November 10, 2005; electronically published December 8, 2005.

Address for correspondence and reprints: Dr. John Rogus, Section on Genetics and Epidemiology, 1 Joslin Place, Room 368, Boston, MA 02215. E-mail: John.Rogus@joslin.harvard.edu

Am. J. Hum. Genet. 2006;78:222–230. 0002-9297/2006/7802-0005$15.00

For example, sib pair linkage analysis is generally performed in two steps: (1) inference of familial inheritance patterns and (2) assessment as to whether this information suggests genetic linkage (Kruglyak and Lander 1995). Since the first step is usually far more time consuming, SPLAT allows this information to be retained so that linkage assessment can be performed repeatedly with some variation. The desire for such repetition may arise from a definition of "affection" that is somewhat fluid. For instance, in our studies of diabetic nephropathy, it is not always clear a priori whether to consider as affected subjects who do not present with advanced kidney disease but do express early biomarkers of disease.

Another issue addressed by SPLAT is that differing user interfaces may be more suitable for different analytical tasks. For when the analyst wishes to quickly engage in exploratory analyses, a point-and-click graphical user interface (GUI) is provided. With this GUI, the user may see the linkage results as they are generated, within a single program. When automation is desired (e.g., in a simulation controlled by a script), a command-line interface is available. A convenient method for implementing whole genome scans is also provided.

In addition, SPLAT provides a number of graphical features, such as the option to filter out irrelevant noise for visual clarity, the ability to view multiple statistics simultaneously, and the luxury of viewing contour curves of the entire LOD surface. These features are designed to aid the analyst in gaining insight from the data set.

Methods and Results

Likelihood Framework

Genetic segregation to an offspring at any locus can be likened to flipping two distinct coins representing the maternal and paternal inheritances. Sharing between siblings can be thought of as the comparative results of flipping the same coins again for a second offspring. Of the four possible outcomes, there is one in which neither coin matches its first flip, two in which exactly one coin matches its first flip, and one in which both coins match. Thus, in the absence of linkage, sibs have an expected sharing pattern of $(\frac{1}{4},\frac{1}{2},\frac{1}{4})$, and the point $(\frac{1}{4},\frac{1}{2})$ on the z_0z_1-plane is typically used as the null hypothesis for sib pair sharing.

The goal of linkage analysis is to estimate allele-sharing proportions across the study population and to assess whether there are any loci at which these proportions deviate significantly from the expectation under the null hypothesis of no linkage. Sharing-proportion estimations are performed within a likelihood framework.

The likelihood equation, based on the multinomial probability mass function, provides a way to assess various (z_0, z_1) points for compatibility with the observed genotype data. In the case of complete data, where it is known explicitly for each sib pair whether zero, one, or two alleles are shared IBD, the likelihood is simply

$$L = z_0^{n_0} z_1^{n_1} z_2^{n_2} \ , \qquad (1)$$

where n_j is the number of sib pairs sharing j alleles IBD. Maximizing L with respect to the parameters, z_j, gives an estimate of the sharing proportions. In the complete data case, the maximum-likelihood estimates are simply $\hat{z}_j = \frac{n_j}{n}$, where $n = \sum n_j$.

For incomplete data, where sib pair sharing is not known precisely, the likelihood of observing the data for each sib pair resembles a conditional probability. The overall likelihood is the product

$$L = \prod_i z_0\rho_{i0} + z_1\rho_{i1} + z_2\rho_{i2} \ ,$$

where ρ_{ij} is the probability of observing the genotype data, given that the ith sib pair shares j alleles IBD. Dividing each factor by the likelihood for the null sharing pattern, $\alpha = (\frac{1}{4},\frac{1}{2},\frac{1}{4})$, gives the likelihood ratio

$$\mathrm{LR} = \prod_i \frac{z_0\rho_{i0} + z_1\rho_{i1} + z_2\rho_{i2}}{\alpha_0\rho_{i0} + \alpha_1\rho_{i1} + \alpha_2\rho_{i2}} \ .$$

The decimal log of the likelihood ratio is a LOD score, the statistic used to assess the significance of deviation from the null hypothesis.

The first stage of linkage analysis is the calculation of ρ_{ij} values. To do so, the SPLAT user must use a program such as GENEHUNTER (Kruglyak et al. 1996) or Merlin (Abecasis et al. 2002) to generate an "IBD file" summarizing inheritance information. If the nuclear families are to be carved out of large genotyped families, the program Loki (Heath 1997) may be preferred, to fully utilize the inheritance information afforded by extended pedigrees. Loki IBDs can then easily be converted to the GENEHUNTER format. Whichever program is used, the file must be generated just once for a given genotype data set, as it can be read into SPLAT, in conjunction with an alterable phenotype file, any number of times. Detailed instructions on how to produce the IBD file can be found in the SPLAT manual.

Each line of the externally generated IBD file describes inheritance information for one sib pair at a designated chromosomal position. Specifically, GENEHUNTER reports the probabilities, π_j, that the sib pair shares j alleles, given the genotype data. We can relate π_j to ρ_j with Bayes's theorem. If J is the event that a sib pair shares j alleles, and if D is the event of observing the genotype data, then

$$\rho_j \equiv P(D|J) = \frac{\pi_j \times P(D)}{\alpha_j} \ .$$

Thus, the LOD score in terms of the IBD probabilities is

$$\begin{aligned} \mathrm{LOD} &= \log_{10}\mathrm{LR} \\ &= \sum_i \log_{10}(4\pi_{i0}z_0 + 2\pi_{i1}z_1 + 4\pi_{i2}z_2) \ . \qquad (2) \end{aligned}$$

Note that for complete data, where, for each *i*, π_{ij} is equal to 1 for exactly one value of *j* and equal to 0 for the other two, the likelihood ratio collapses to the form expected from equation (1):

$$\mathrm{LOD} = \log_{10}\frac{z_0^{n_0}z_1^{n_1}z_2^{n_2}}{\left(\frac{1}{4}\right)^{n_0}\left(\frac{1}{2}\right)^{n_1}\left(\frac{1}{4}\right)^{n_2}} \ .$$

Splat first examines the phenotype data to designate the sib pairs to be incorporated into the analysis. These can be affected sib pairs (ASPs) or discordant sib pairs (DSPs). Then, upon reading in the IBD probabilities, the program employs the expectation maximization (EM) algorithm to maximize the LOD score (subject to the restrictions and options discussed below) and to thereby determine the maximum-likelihood estimate of the sharing pattern for each chromosomal position (fig. 1).

Sharing-Pattern Domain Restrictions

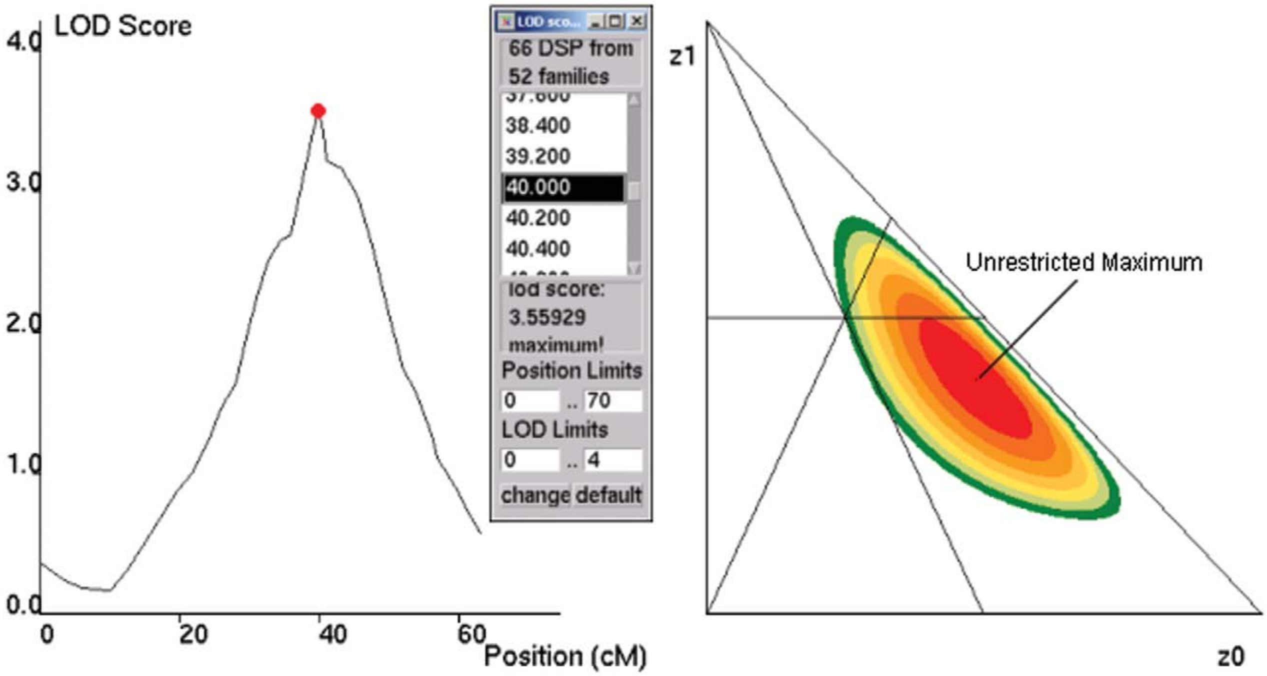
In figure 1, the maximum-likelihood sharing pattern, at the center of the contour curves plotted for the position 40 cM, emerged from an unrestricted domain; the sharing pattern was allowed a full 2 df across the entire planar region defined for probabilities (i.e., the projection of $z_0+z_1+z_2=1$, with all $z_j \geqslant 0$, onto the $z_o z_1$-plane). However, a study's power can be increased by restricting the domain over which the maximum-likelihood sharing pattern is sought to a subset where it is most likely to be found. In doing so, the threshold value for a LOD score to be considered significant is reduced, since the unrestricted 2-df reference distribution is replaced by one comprising a mixture of 0-, 1-, and 2-df distributions. Splat includes an interactive feature for the analyst to define any logically consistent linear or triangular domain restriction, and options for several common constraints are built in.

The simplest modification to the domain is to restrict to those sharing patterns for which mean sharing diverges from the null in the direction of interest. That is, one may restrict to $m>1/2$ for ASPs (or, more broadly, for sibs concordant for any trait) or to $m<1/2$ for DSPs (fig. 2*a*), where

$$m \equiv \frac{1}{2}\sum_{j=0}^{2} jz_j \ .$$

Cutting the domain in half essentially reduces the problem to a one-sided test. The utility of such restriction is apparent in our DSP example, where, in addition to being discordant for nephropathy, all sib pairs are concordant for diabetes. This being the case, we find increased sharing (and, hence, high LOD scores) in chromosomal regions linked to diabetes (fig. 3). Although this does provide an interesting means of checking sample and data integrity, the peaks can be misleading in the search for regions linked exclusively to nephropathy.

A stricter "possible triangle" restriction, based on principles of biological consistency, was proposed for ASPs by Holmans



66 DSP from
52 families

37.600
38.400
39.200
40.000
40.200
40.400

lod score:
3.55929
maximum!

Position Limits
0 .. 70

LOD Limits
0 .. 4

change default

Figure 1 DSP linkage curve for diabetic nephropathy on a region of chromosome 3q (Moczulski et al. 1998) and contour curves (LOD intervals of one-half) defining the LOD surface at the peak (40 cM).

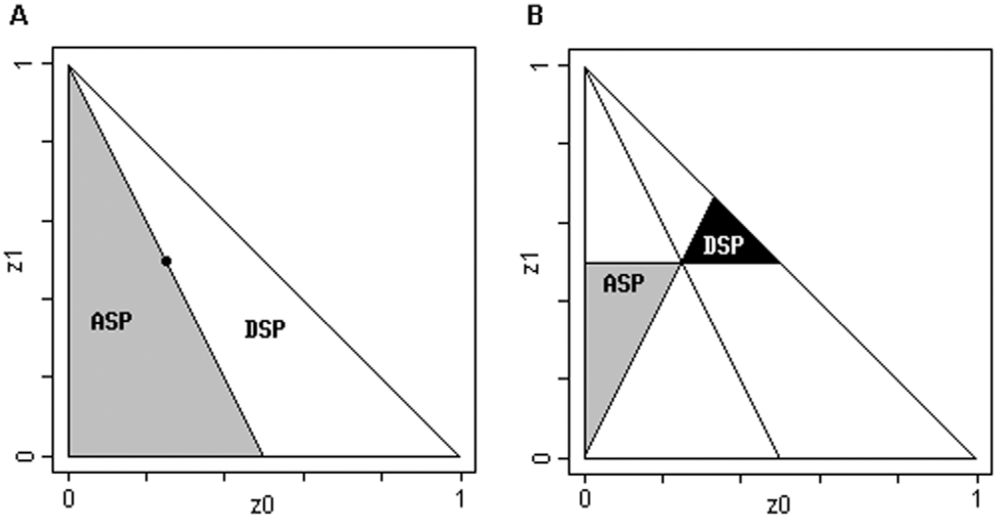


Figure 2 Domain restriction. *A,* Line defined by MSH, $z_1 = 1 - 2z_0$, bifurcates the domain into the broadest regions consistent with excess (ASP) and decreased (DSP) sharing. *B*, ASP and DSP triangles.

(1993), as was an analog for DSPs (Lunetta and Rogus 1998) (fig. 2*b*). For ASPs,

$$2z_0 \leqslant z_1 \leqslant \frac{1}{2}$$

and, for DSPs,

$$2z_0 \geqslant z_1 \geqslant \frac{1}{2} \ .$$

The "no dominance variance" test is stricter yet. It restricts to 1 df by fixing z_1 to a value of 1/2. Since the possible triangle and no dominance variance restrictions are subsets of the mean sharing–based restriction, misleading LOD scores, such as those in figure 3, would also be prevented by applying either of these.

When the analyst opts to utilize a restricted domain, such as Holmans's triangle, the program first calculates the unrestricted maximum-likelihood sharing pattern. If this point is found to be within the specified domain restriction, then it is accepted. However, if one of the restriction conditions is violated, the maximization is repeated—this time constrained to the line defining the violated condition—and the sharing pattern resulting from this second maximization is elected instead. If both restriction conditions are violated by the unrestrained maximization, then the maximum-likelihood sharing pattern is the intersection of the restriction conditions—the point $(\frac{1}{4},\frac{1}{2},\frac{1}{4})$—and the LOD score is 0. The sharing pattern arrived at by this process could alternatively be read directly from a plot of the LOD surface, as in figure 1.

Sib pair classes need not be defined by true dichotomous traits. Several ascertainment schemes are based instead on threshold values of a quantitative trait, and an appropriate domain restriction has been proposed for each. For example, in the extreme discordant sib pair (EDSP) design (Risch and Zhang 1995, 1996; Kruse et al. 1997), pairs are recruited such that one sib is in the top decile of the population distribution for a trait and the other is in the bottom decile. The appropriate domain restriction for EDSP is

$$\frac{2}{3} - \frac{2}{3}z_0 \leqslant z_1 \leqslant 2z_0 \ .$$

Other examples include pairs concordant for either high or low values (Xu et al. 1999). SPLAT is sufficiently flexible to accommodate any of these designs or any user-defined restriction.

Null Hypotheses

Rather than the point $(\frac{1}{4},\frac{1}{2})$, several statistics define the null hypothesis less restrictively as a line passing through this point. An example is the line defined by mean sharing of one-half (MSH), $z_1 = 1 - 2z_0$ (fig. 2*a*). The mean sharing test (MST) (Blackwelder and Elston 1985) assesses deviation from this line, which is equivalent to testing whether sib pairs share two alleles at a different frequency than that at which they share zero alleles. In practice, the program need not directly maximize the likelihood on the null hypothesis line. Instead, it utilizes the fact that the MST is equivalent to

$$\log_{10}\frac{L_{2\mathrm{df}}}{L_{\mathrm{MSH}}} = \log_{10}\left(\frac{\dfrac{L_{2\mathrm{df}}}{L_{(1/4,\,1/2)}}}{\dfrac{L_{\mathrm{MSH}}}{L_{(1/4,\,1/2)}}}\right)$$

$$= \mathrm{LOD}_{2\mathrm{df}} - \mathrm{LOD}_{\mathrm{MSH}} \ .$$

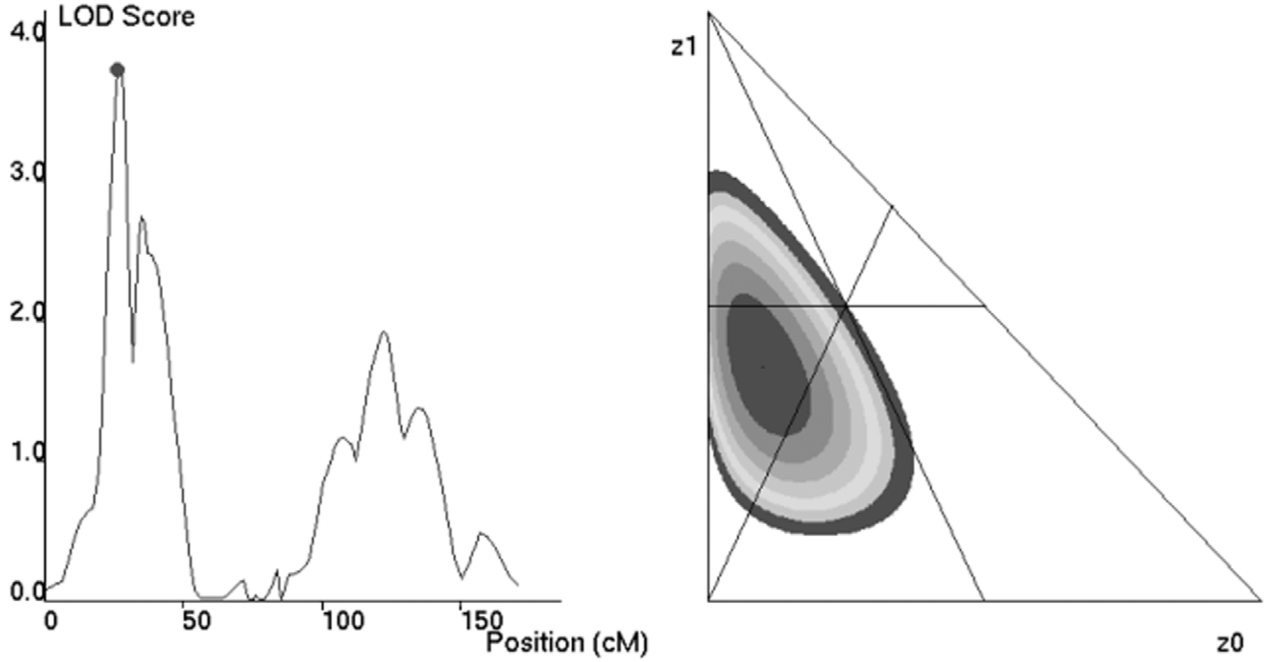


Figure 3 In regions linked to diabetes, such as the human leukocyte antigen region on chromosome 6, sib pairs discordant for nephropathy but concordant for diabetes showing *increased* sharing and, in the absence of domain restriction, achieving a potentially misleading nonzero LOD score.

Thus, the LOD score for the MST is simply the difference between the unrestricted LOD score and the MSH LOD score.

Another possible null hypothesis, the proportions test (Day and Simons 1976), stems from the expectation that $z_2 = 1/4$ in the absence of linkage. Hence, one could test for deviation from the line $z_1 = 3/4 - z_0$. Since others may be proposed in the future, Splat allows the user to define the null hypothesis as any line passing through $(\frac{1}{4},\frac{1}{2})$.

Filtering

Even within the context of a restricted domain, potentially misleading LOD score peaks may result when the sharing pattern drifts away from $(\frac{1}{4},\frac{1}{2})$ yet remains close to the MSH line. Such a situation is not representative of true linkage. To account for this and to clean up plots of the results, Splat includes an option to eliminate spurious findings by zeroing out LOD scores for all positions at which mean sharing is hovering around 1/2 (i.e., if it is less than some threshold value for ASPs or greater than a threshold value for DSPs).

Weighting

Since sibships collected for linkage studies can vary widely in size, it is normally recommended to employ a weighting scheme to prevent large families from contributing disproportionately to the results. Summands in equation (2) should be scaled down, such that the number of pairs effectively contributed by each family is equal to one fewer than the number of sibs involved in the analysis (Hodge 1984). Under such a scheme, the $\frac{n(n-1)}{2}$ ASPs in families with n affected sibs are each weighted by the factor $\frac{2}{n}$, and the nm DSPs in families with n affected and m unaffected sibs are each weighted by $\frac{n+m-1}{nm}$. Although weighting is recommended and is turned on by default, Splat gives the user the ability to turn this option off.

Grid Scan

Linkage analysis programs typically determine sharing patterns by utilizing an iterative procedure, such as the EM algorithm (Dempster et al. 1977; Little and Rubin 1987), to hone in on the maximum-likelihood solution, subject to a limited selection of constraints and null hypotheses. Although Splat does this more flexibly than other programs, the approach has an inherent limitation: it does not convey the context from which a constrained maximal solution has emerged. This can be important, since several factors have the potential to drive sharing patterns away from the null in a manner inconsistent with actual linkage. One such factor is genotype error (Olson et al. 2004), which is not always possible to eliminate, since sib pair samples tend to be drawn from nuclear families (sometimes with parents unavailable). Additional complicating factors are the possibility of phenotype misclassification and the limited sample sizes available for relatively infrequent diseases. The results of an EM procedure often provide little insight as to whether a signal connotes linkage or is merely a reflection of these obfuscating factors. It is, therefore, not prudent to ignore the broader context provided by the entire LOD surface.

To supply this context and to alleviate the "black box" feel of analysis relying solely on an iterative algorithm methodology, we have implemented a novel grid-scanning process. At any given locus, we can derive contour curves of the likelihood surface by directly computing the sum in equation (2) at regular intervals across the z_0z_1-plane. Doing so provides an informative visual representation in which multiple statistics can be depicted on the same graph. It is apparent immediately upon inspection whether a high LOD score accompanies a meaningful sharing pattern or is more likely an aberration. In SPLAT, the left panel of the split-screen main window presents a chromosomewide plot of LOD score versus position (subject to any domain restriction and null hypothesis), and the right-hand panel displays the contour plot for whichever locus the user selects from a separate position control window (fig. 1). Together they afford a potent account of potential linkage results.

User Interface

Beyond the theoretical considerations outlined above, a number of practical issues arise in performing linkage analysis. Analytic tasks must be repeated for each chromosome in a genome scan. Not only can this be tedious, but it can make documentation and file management a nontrivial task, especially when various phenotype definitions and/or analysis options are to be considered. The difficulty is exacerbated when results must be plotted with a separate program (or on a different platform) from the one in which they were generated. These bottlenecks can impede exploratory analysis. To streamline the process, we included a batch processing feature in SPLAT that can automatically repeat analyses for multiple chromosomes. In addition, the program has plotting capability and a robust GUI. Users can browse to select the desired IBD and phenotype files and to name the EM scan output file. This makes it quite easy to run multiple analyses with varying phenotype definitions. Options for study design, domain restriction, null hypothesis, filtering, and weighting are all easily set by the user in a single window. The program can also be run from the command line, which is useful for simulations in which the user would like to use a controlling script to run the program thousands of times with randomly generated data.

Multiple Testing

Any attempt to find regions of the genome linked to a phenotypic trait of interest must strike a balance between two competing goals: maximizing the chance of finding truly relevant regions (power) and minimizing the chance of implicating irrelevant regions (type I error). SPLAT was designed with the mindset that, given the cost and effort associated with typical genetic studies, researchers are usually highly motivated to uncover promising leads, even at the cost of some additional false positives. Specifically, SPLAT allows the analyst to quickly and easily reconsider analyses under different clinical scenarios and with alternative analytic options. Because such multiple testing provides additional chances to obtain spurious findings, the resulting *P* values should ideally be corrected to maintain a given significance level, one that is neither biased upward nor downward. Unfortunately, the easiest correction procedure, the Bonferroni adjustment, can be substantially conservative since it does not account for the potentially high correlation across the tests. As a practical matter, therefore, we recommend that a primary analysis plan be rigorously developed a priori and implemented exactly as planned. This means that all aspects of the analysis, from phenotype definition to choice of statistic, should be determined before the primary analysis commences. The *P* values for this analysis can then be assessed using, for example, the guidelines set forth by Lander and Kruglyak (1995). After the primary analysis, the investigator is then free to perform (possibly extensive) exploratory analysis, provided that the results are reported with full disclosure of the extent of exploration. Should the analyst wish to put forth a conservative bound on statistical significance, a Bonferroni adjustment may be made by scaling up *P* values by a factor equal to the number of tests performed.

Discussion

Many of the analytic challenges of sib pair analysis arise from the fact that siblings may share up to two chromosomes IBD at any given chromosomal location. In the absence of consanguinity, this phenomenon does not occur for other types of relative pairs, and, as a result, statistical testing for nonsib pairs is naturally based on a one-sample binomial test (Risch 1990*a*, 1990*b*, 1990*c*). For example, cousin pair analysis would test whether the frequency of sharing one chromosome IBD differs from the null value of 1/4. The multinomial analog for sib pair analysis, however, is less clear cut. Two decades ago, Blackwelder and Elston (1985) demonstrated the superiority of the ASP means test ($m = 1/2$) over the proportions test ($z_2 = 1/4$) (Day and Simons 1976) for a broad range of genetic models, but Wu and Amos (2003) outlined a region in which the reverse in true. Schaid and Nick (1990) derived a statistic based on the maximum of these two statistics, and Knapp (1991) suggested an alternative linear combination of the underlying multinomial probabilities. More recently, Whittemore and Tu (1998) described a general class of statistics encompassing both the means test and the proportions test, and they introduced the idea of choosing a "minmax" statistic from this class to achieve robust performance across all possible genetic models (i.e., to minimize the maximum possible loss of efficiency). Such examples highlight the utility of an analytic framework that allows the analyst to choose from among a broad set of alternative test statistics.

Risch (1990*b*) pointed out that most common complex diseases in man show little or no dominance effect (under which the sibling recurrence risk would be similar to that for offspring). If this condition is met, the power of a study may be improved by imposing the domain restriction of $z_1 = 1/2$. However, if this assumption appears too stringent, Holmans (1993) showed that power could still be improved by restricting maximization to the set of possible haplotype-sharing probabilities, the so-called possible triangle. In fact, other domain restriction strategies are possible (Holmans 1993; Greenwood and Bull 1999), and we have described a methodology to implement any logical constraint. It should be noted, however, that, under some ascertainment schemes, restriction to sharing in a particular direction may not necessarily be appropriate. In our example of sib pairs concordant for diabetes but discordant for nephropathy, restricting to regions of decreased sharing may exclude chromosomal regions harboring a gene that elicits joint susceptibility. In situations where the analyst feels that the ascertainment scheme may influence the underlying sharing in a manner contrary to such assumptions or that the disease model may render them inappropriate, she retains complete flexibility to perform unrestricted analysis.

One nice feature of describing ASP test selection and domain restriction in generality is that the concepts are easily adaptable to other variations of dichotomous trait sib pair analysis. For example, for diseases with high sibling recurrence risk, Rogus and Krolewski (1996) demonstrated the advantages of analyzing DSPs. As with ASPs, there is a full array of legitimate test statistics (e.g., analogs of the means test and proportions test), and domain restriction may rely on a stringent criterion of $z_1 = 1/2$ or on the Holmans-type triangle defining legal genetic models (Lunetta and Rogus 1998). A whole additional class of dichotomous trait sib pair analysis was born when Risch and Zhang (1996) conceived the idea of EDSPs. The Holmans-type triangle for EDSPs was described by Kruse et al. (1997). Later, Xu et al. (1999) supplemented EDSPs with highly concordant sib pairs in a genetic study of blood pressure.

Although many of the options described above have been implemented piecemeal in various computer programs (e.g., MAPMAKER/SIBS, GENEHUNTER, ASPEX, Allegro, and Merlin, among others), no existing program gives the user complete analytical freedom. Typically, one can set various preferences, such as restricting the program to a 1-df test; however, the range of investigative options is bound to the discrete set of statistical tests provided by any given package. Our approach, on the other hand, addresses linkage-analysis needs quite generally and brings all the possibilities outlined above under one umbrella, allowing them to be treated in the same manner. SPLAT is designed to handle any traditional sib pair statistical test currently in use, as well as any user-defined test yet to be proposed, thereby offering the analyst a high degree of flexibility and self-sufficiency for exploratory analysis. This flexibility is enhanced by a grid scan of the likelihood surface. The ability to simultaneously visualize multiple test statistics gives a complete picture of the actual genetic sharing as well as the significance of potential linkage according to any test of interest. Since SPLAT relies on other programs to calculate IBD statistics and simply imports these from a text file, it is indifferent to the source or quantity of genotype data. Thus, it is equally well suited for analyzing 10-cM microsatellite scans as for the increasingly popular higher density and more accurate SNP-based linkage panels.

Beyond the fundamental considerations of design, test, and domain restriction, additional complexity has arisen from more recent additions to the field, further underscoring the need for a flexible framework. Our approach (and the software) could be adapted to accommodate novel twists to sib pair linkage analysis involving covariates, imprinting, error detection, and the use of multiple sib pair definitions within a data set.

Greenwood and Bull (1999) generalized the ASP model to allow for covariates, arguing that since environmental factors can affect disease risk it is plausible that they could change the ratio of disease penetrances and, as a result, the evidence for linkage. Holmans's arguments, they point out, considered genetic effects only. The triangle will continue to hold when gene-environment interactions affect only the size of the genetic effect, but, when either the exposure changes the direction of the disease gene effect or there are sib pairs with differing exposures, the sharing patterns may reside outside the triangle. When this is the case, the power comparisons may not be valid, so, to develop a new set of constraints, they investigated three approaches: average constraints, subgroup-triangle constraints, and simultaneous-boundary constraints. They further raise the issue that different boundary constraints could be used if it is anticipated that a gene might act recessively in one group but additively or dominantly in another group, but there is not likely to be good support for the hypothesis of differing modes of inheritance in covariate subgroups. Interestingly, in the models they investigated, application of the no dominance variance restriction gave better power than the triangles, echoing Lunetta and Rogus (1998), who showed that there is only a very small region of the full-parameter space where triangle constraints are more powerful.

Knapp and Strauch (2004) extended the possible triangle test to account for genomic imprinting, or parent-of-origin effect, where affected individuals inherit mutant alleles preferentially from one particular parent. This effect, which can be brought about by DNA methylation or by differential packing density of DNA by histones, is known to be present at many chromosomal regions. To account for it in parametric linkage analysis, they extend the trait model to contain two different heterozygote penetrances. For model-free analysis, they propose extending the sharing pattern to a fourth component, to distinguish the maternal and paternal sharing of one allele: (z_0, z_1^F, z_1^M, z_2). The normal constraints on probabilities would then restrict the sharing pattern to a subset of $\mathbb{R}^3$, for which an analog of the Holmans triangle could be implemented, with the likelihood-ratio test statistic following a mixture of χ^2 distributions with 0, 1, 2, and 3 df. Grid scanning would become cumbersome in three dimensions, but the EM portion of our software, including allowance for generalized restrictions and null hypotheses, could be adapted to such models, provided that the appropriate four-component IBD extraction were available.

Olson et al. (2004) highlight the critical importance of detecting genotype errors in a linkage study. This issue is particularly relevant to sib pairs in which, if parents are not present, Mendelian errors cannot be found with certainty. Such errors will be even more difficult to detect as the field moves toward high-density SNP genome scans. To combat this, a quick graphical approach to linkage analysis, such as the one we propose here, can help to identify data sets for which a major source of error exists. If a sib pair distribution is found to be shifted substantially away from zero in either direction, genotype error should be suspected. For example, a shift toward increased allele sharing can indicate severe misspecification of allele frequencies.

Our framework could also be extended to handle multiple sib pair definitions within a single data set. Since ASPs may be expected to show increased sharing in the same chromosomal regions where DSPs for the same trait show decreased sharing, it would seem sensible to combine analysis of both classes (Guo and Elston 2000). The optimal strategy to proceed with such analysis must account for the relative distributions of the classes within the families, among other complications.

The framework we have laid out can handle any sib pair test based on a qualitative trait, whether it is measured qualitatively or is assessed by somehow dichotomizing a quantitative trait. Any consistent set of constraints, once derived for a particular ascertainment method, can simply be plugged into the software we have developed. Analysts working within this class of problems are thus granted a great deal of flexibility. The approach is, however, limited to pairwise statistics; it is not applicable to statistics which employ measures of groupwise sharing, such as affected sib triples (Whittemore and Tu 1998). Furthermore, we consider only siblings and do not trace sharing through generations, such as in GENEHUNTER's NPL statistic. Another limitation is that we consider only autosomal chromosomes. Data for pseudoautosomal regions could be analyzed with this approach; however, care should be taken in interpretation. The fact that same-sex pairs will generally show increased sharing in these regions can elicit misleading results if such sharing inequalities are not accounted for (Dupuis and Van Eerdewegh 2000). Finally, although quantitative traits can be considered with this approach, they must be dichotomized and, therefore, are not used directly.

Acknowledgments

We would like to acknowledge significant contribution to the development of the SPLAT software by Garretson Beebe and Christopher Botka of the Bioinformatics Resource at the Joslin Diabetes Center. This research was supported by National Institutes of Health grants DK53534 and DK58549 and by Juvenile Diabetes Research Foundation grants 9-2005-27 and 9-2000-1008.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

Allegro, http://www.decode.com/software/allegro/
ASPEX, http://aspex.sourceforge.net/
Loki, http://loki.homeunix.net/
Merlin, http://www.sph.umich.edu/csg/abecasis/Merlin/index.html/
MIT Genome Center FTP Archive, http://www.broad.mit.edu/ftp/distribution/software/ (for MAPMAKER/SIBS and GENEHUNTER)
SPLAT, http://www.joslinresearch.org/LabSites/Krolewski/splat/(for computer program and manual)

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85–97

Day NE, Simons MJ (1976) Disease susceptibility genes: their identification by multiple case family studies. Tissue Antigens 8:109–119

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Dupuis J, Van Eerdewegh P (2000) Multipoint linkage analysis of the pseudoautosomal regions, using affected sibling pairs. Am J Hum Genet 67:462–475

Greenwood CM, Bull SB (1999) Analysis of affected sib pairs, with covariates with and without constraints. Am J Hum Genet 64:871–885

Guo X, Elston RC (2000) Two-stage global search designs for linkage analysis II: including discordant relative pairs in the study. Genet Epidemiol 18:111–127

Heath SC (1997) Markov chain segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760

Hodge SE (1984) The information contained in multiple sibling pairs. Genet Epidemiol 1:109–122

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Knapp M (1991) A powerful test of sib-pair linkage for disease susceptibility. Genet Epidemiol 8:141–143

Knapp M, Strauch K (2004) Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease models with imprinting. Genet Epidemiol 26:273–285 (erratum 28:288)

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Kruse R, Seuchter SA, Baur MP, Knapp M (1997) The "possible triangle" test for extreme discordant sib pairs. Genet Epidemiol 14:833–838

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

Lunetta KL, Rogus JJ (1998) Strategy for mapping minor histocompatibility genes involved in graft-versus-host disease: a novel application of discordant sib pair methodology. Genet Epidemiol 15:595–607

Moczulski DK, Rogus JJ, Antonellis A, Warram JH, Krolewski AS (1998) Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: results of novel discordant sib-pair analysis. Diabetes 47:1164–1169

Olson JM, Song Y, Lu Q, Wedig GC, Goddard KA (2004) Using overall allele-sharing to detect the presence of large-scale data errors and parameter misspecification in sib-pair linkage studies. Hum Hered 58:49–54

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

——— (1990*b*) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

——— (1990*c*) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253 (erratum 51:673–675)

Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584–1589

——— (1996) Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. Am J Hum Genet 58:836–843

Rogus JJ, Krolewski AS (1996) Using discordant sib pairs to map loci for qualitative traits with high sibling recurrence risk. Am J Hum Genet 59:1376–1381

Schaid DJ, Nick TG (1990) Sib-pair linkage tests for disease susceptibility loci: common tests vs. the asymptotically most powerful test. Genet Epidemiol 7:359–370

Whittemore AS, Tu IP (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet 62:1228–1242

Wu CC, Amos CI (2003) Statistical properties of affected sib-pair linkage tests. Hum Hered 55:153–162

Xu X, Rogus JJ, Terwedow HA, Yang J, Wang Z, Chen C, Niu T, Wang B, Xu H, Weiss S, Schork NJ, Fang Z (1999) An extreme-sib-pair genome scan for genes regulating blood pressure. Am J Hum Genet 64:1694–1701