ASPEX: an integrated toolkit for nonparametric linkage analysis of sib pair data

Running title: Nonparametric sib pair analysis

David A. Hinds and Neil Risch
Department of Genetics
Stanford University

September 18, 2003

Address for correspondence:

Dr. Neil Risch
Department of Genetics
Stanford University Medical Center
300 Pasteur Drive room M322
Stanford, CA 94305-5120
Voice: (650) 725-5967
FAX: (650) 725-1534
E-mail: risch@lahmed.stanford.edu

Introduction

Nonparametric sib pair methods are based on the expectation that affected relative pairs should tend to share more alleles identical by descent (IBD) at markers in the vicinity of a predisposing gene. They do not require specification of a particular mode of inheritance for a putative disease gene. They also have the practical advantage that for rare and/or late-onset diseases, it is often much easier to collect sib pairs than larger pedigrees with multiple affected individuals. For both reasons, sib pair methods are often chosen for studies of complex diseases.

Affected sib pair (ASP) methods have been reviewed elsewhere (Holmans 1998). The first ASP methods were based on simple genotype counting statistics at single loci (Penrose 1935), but most are based on directly or indirectly counting the number of alleles shared IBD. In the absence of linkage, the probabilities of a sib pair being IBD for 0, 1, or 2 alleles are 0.25, 0.5, and 0.25, respectively. Likelihood ratio methods (Risch 1990a, 1990b) score the likelihood of the observed marker data under alternative sets of IBD probabilities, and can also be used to calculate maximum likelihood estimates for the IBD probabilities. A strength of likelihood methods is that they permit inclusion of partially informative data where IBD cannot be determined unambiguously, such as when one or both parents are untyped. However, even if both parents are typed, incomplete polymorphism can still prevent unambiguous determination of IBD.

Multipoint extensions of likelihood methods integrate IBD information across multiple linked markers, to extract the maximum amount of linkage information. Interval methods (Olson 1995; Holmans and Clayton 1995) determine likelihoods across flanking pairs of linked markers. Hidden Markov model (HMM) methods permit multipoint likelihoods to be efficiently calculated across many markers, with or without typed parents. The MAPMAKER/SIBS program (Kruglyak and Lander 1995) was the first general HMM implementation for sib pair analysis.

We have developed a software package for affected sib pair analysis of qualitative traits, ASPEX, that implements interval and HMM likelihood methods. Table 1 lists the main analysis programs included in the ASPEX package. While each program performs just one type of analysis, the tools are integrated in the sense that all accept the same parameter and input files. We have incorporated several novel improvements to traditional sib pair analysis techniques, which we have validated with simulated and real data.

Multipoint likelihoods for sibships with typed parents

In a sibship with both parents genotyped, nearly all the available multipoint linkage information can be extracted from just the nearest informative paternal and maternal meioses on either side of a given location. Likelihood calculations are then very similar to other interval methods, but instead of using a fixed interval, we pick the most informative interval separately for each meiosis. For most parental mating types, the likelihoods of the marker data given the IBD state for one parent are conditionally independent of the IBD state for the other parent. The only exception is in $AB \times AB$ matings, for sibs with AB genotypes. These matings give some sharing information, but the parental origin of a transmitted allele can never be determined unambiguously. For reasonably polymorphic marker data, $AB \times AB$

mating types should be uncommon and will represent a small fraction of the available linkage information.

In the *sib_ibd* program, for each sib pair, we determine the paternal and maternal IBD status at each marker position, wherever this is unambiguous. Missing parents may be reconstructed using additional unaffected sibs as well as just one of the sibs in the pair being scored. For each location to be scored, we determine the nearest flanking positions that have sharing information, separately for each parent. Considering just one parent, let θ_A and θ_B be the recombination fractions for the nearest markers on either side of the probe location that are informative for sharing through that parent, and let θ_{A+B} be the recombination fraction for the entire flanking interval. In the absence of interference, $\theta_{A+B} = \theta_A + \theta_B - 2\theta_A\theta_B$.

$$\begin{aligned}
\theta_{AB} &= (\theta_A + \theta_B - \theta_{A+B})/2 \\
\theta_{\bar{A}B} &= (\theta_B - \theta_A + \theta_{A+B})/2 \\
\theta_{A\bar{B}} &= (\theta_A - \theta_B + \theta_{A+B})/2 \\
\theta_{\bar{A}\bar{B}} &= 1 - (\theta_A + \theta_B + \theta_{A+B})/2
\end{aligned}$$

Let i_j be the number of alleles (0 or 1) shared at location $j \in \{A, D, B\}$. For a triplet (i_A, i_D, i_B), we can calculate a corresponding probability from the number of implied recombinations:

$$p(i_A, i_D, i_B) = \begin{cases} \theta_{AB}^2 + \theta_{\bar{A}B}^2 + \theta_{A\bar{B}}^2 + \theta_{\bar{A}\bar{B}}^2 & \text{if } i_A = i_D = i_B \\ 2(\theta_{AB}\theta_{A\bar{B}} + \theta_{\bar{A}B}\theta_{\bar{A}\bar{B}}) & \text{if } i_A = i_D \neq i_B \\ 2(\theta_{AB}\theta_{\bar{A}B} + \theta_{A\bar{B}}\theta_{\bar{A}\bar{B}}) & \text{if } i_A \neq i_D = i_B \\ 2(\theta_{AB}\theta_{\bar{A}\bar{B}} + \theta_{A\bar{B}}\theta_{\bar{A}B}) & \text{if } i_A = i_B \neq i_D \end{cases}$$

Let i_{kj} be the number of alleles shared at location $j \in \{A, D, B\}$ via parent $k \in \{f, m\}$. We can calculate ω_l where $l \in \{0, 1, 2\}$ is the total number of alleles IBD at the disease locus, as products of terms for each parent:

$$\begin{aligned}
\omega_0 &= p(i_{mA}, 0, i_{mB})p(i_{fA}, 0, i_{fB}) \\
\omega_1 &= p(i_{mA}, 0, i_{mB})p(i_{fA}, 1, i_{fB}) + p(i_{mA}, 1, i_{mB})p(i_{fA}, 0, i_{fB}) \\
\omega_2 &= p(i_{mA}, 1, i_{mB})p(i_{fA}, 1, i_{fB})
\end{aligned}$$

From these likelihoods, we calculate likelihood ratios and maximum likelihood sharing estimates (Risch 1990b).

If only one parent is typed (and heterozygous), we can sometimes determine IBD with respect to that parent even if the other parent cannot be reconstructed. However, unless care is taken, the resulting IBD scores will be biased against sharing. We only score IBD for one parent if the sibs are both heterozygotes and each have an allele that does not match either of the typed parent's alleles (i.e., typed parent AB, sibs AX or BX).

Likelihood calculations with incomplete data

The basic HMM likelihood algorithm has been described elsewhere (Lander and Green 1987), and we only summarize the method here. Consider data for a nuclear family with n siblings and m linked markers $M_1..M_m$, where the genetic distance between M_i and M_{i+1} is θ_i. At any marker, there are $N = 2^{2n-2}$ distinct inheritance states, where we define an inheritance

state as a pattern of sharing of maternal and paternal alleles identical by descent with the first sibling in the family. At a single marker M_i, we can determine a vector q_i giving the likelihood of the observed genotypes at that marker for all N possible inheritance states. Let $T(\theta)$ be a matrix giving the transition probabilities between all pairs of inheritance states for a given genetic distance, and let the "$\circ$" operator be the element-wise product of two vectors of size N, yielding a result with N elements. We build up multipoint likelihoods in stepwise fashion from each end of the list of markers:

$$\begin{aligned} P_0^L &= P_{m+1}^R = I_N = \{1, 1, \ldots\} \\ P_{i+1}^L &= P_i^L T(\theta_i) \circ q_{i+1} \\ P_{i-1}^R &= P_i^L T(\theta_{i-1}) \circ q_{i-1} \end{aligned}$$

The product of a likelihood vector and a transition matrix using straightforward matrix multiplication requires $\mathrm{O(N^2)}$ operations. More efficient algorithms taking advantage of symmetries in the transition matrix can reduce this to $\mathrm{O(N \log_2 N)}$ operations (Kruglyak and Lander 1998; Idury and Elston 1997). However, larger reductions can be obtained by exploiting some special features of the likelihood vectors, particularly for nuclear family data.

Consider the possible parental mating types at a single marker where both parents are typed. In an $AA \times AA$ or $AA \times BB$ mating, q_i will equal the identity vector, I_N: the probability of the marker data will be equal for all inheritance states. We do not explicitly calculate likelihood vectors at uninformative markers: instead, we keep track of the distance to the nearest informative marker, so these matings are "free". In an $AA \times AB$ or $AA \times BC$ mating, $\sqrt{N}$ states of the likelihood vector will have non-zero probabilities: the rest will be excluded by data from the informative parent. In an $AB \times AC$ or $AB \times CD$ mating, only one inheritance state will have nonzero probability. $AB \times AB$ matings are a special case: the number of allowed states will depend on which alleles are transmitted, but will never exceed $\sqrt{N}$. In any of these cases, we store just the non-zero likelihoods in a sparse array. Now, when iteratively calculating P^L or P^R, we only evaluate terms for transitions from states with non-zero probability in the previous P to states with non-zero probability in q_i. This will either be $\mathrm{O}(1 \times 1)$, $\mathrm{O}(1 \times \sqrt{\mathrm{N}})$, or $\mathrm{O}(\sqrt{\mathrm{N}} \times \sqrt{\mathrm{N}})$ depending on the mating types at the two positions.

If one or both parents are untyped, it will often be impossible to deduce the parental mating type with certainty. The likelihood vector then is a weighted sum of vectors for each possible mating type, with the weights determined by allele frequencies. In cases where an $AA \times AA$ mating type is possible, all N elements of the vector would have nonzero likelihoods. Rather than evaluating and storing this full vector, we split it into a sum of a weighted identity vector (which need not be stored), and a weighted vector with only $\mathrm{O}(\sqrt{\mathrm{N}})$ non-zero elements. Thus, we still only need to store $\mathrm{O}(\sqrt{\mathrm{N}})$ distinct likelihoods. We evaluate matrix products as sums of products of these components. For example:

$$\begin{aligned} q_i &\Rightarrow q_i^{AA \times AA} + q_i^{AA \times AB} \\ P_i^L &\Rightarrow P_{i-1}^L T(\theta_{i-1}) q_i^{AA \times AA} + P_{i-1}^L T(\theta_{i-1}) q_i^{AA \times AB} \end{aligned}$$

Here, we have decomposed a q_i that would be $\mathrm{O(N)}$ into an uninformative term, and a term that is $\mathrm{O}(\sqrt{\mathrm{N}})$. We calculate P^L as two terms which are stored separately. After evaluating

a product, we merge any result vectors that span the same set of inheritance states. This prevents exponential growth in the number of terms in the P^L results as more markers are evaluated. The end result is that regardless of whether or not parents are typed, the calculation of the P^L and P^R vectors is at worst O(N).

Now consider a putative disease locus between M_i and M_{i+1}, at a distance δ from M_i. From the one-sided likelihoods, we can calculate a likelihood for all the genotype data given an assumed IBD state $j \in 0, 1, 2$ for a particular sib pair at this locus:

$$\begin{aligned} \omega_j &= \sum \left[P_i^L T(\delta) \circ \Phi_j \circ P_{i+1}^R T(\theta_i - \delta) \right] \\ \Phi_j &= \begin{cases} 1 & \text{for all states with } IBD = j \\ 0 & \text{for all states with } IBD \neq j \end{cases} \end{aligned}$$

A direct evaluation of this equation would require two transition matrix multiplication steps. Each of these would be an expensive calculation, because there is no genotype data at the disease locus to limit the set of inheritance states that need to be considered. However, we can avoid explicitly calculating these products by recognizing that for a given inheritance state at M_i and M_{i+1}, we can partition the likelihood of the marker data into pieces for $j \in 0, 1, 2$ based on the inheritance state of just the target sib pair on either side of the interval.

We have implemented this likelihood algorithm in the ASPEX *sib_phase* program. The *sib_phase* program can generate multipoint maximum likelihood LOD scores as well as exclusion maps. The program can accommodate gender-specific recombination rates and either autosomal or sex-linked data. For autosomal data, separate LOD scores for maternal and paternal sharing can be calculated under the assumption of a multiplicative model. And multipoint statistics can be calculated for affected, unaffected, and discordant sib pairs.

Relative performance of the multipoint algorithm

One consequence of our scheme for decomposing inheritance vectors into simpler components is a dramatic reduction in memory requirements. Other implementations of the HMM method all require at least O(N) space to hold temporary results. Our method requires only $O(\sqrt{N})$ space: a huge reduction for larger families. In some cases, memory requirements are an even more severe limit on HMM methods than raw computer speed.

All HMM-based linkage algorithms have an exponential dependence on the number of non-founders in a pedigree. For MAPMAKER/SIBS and GENEHUNTER, this translates into roughly four-fold increases in memory and time for each additional sib within a sibship. ASPEX shows the same exponential time dependence, but is still up to 350 times faster for larger sibships. The algorithms used in MAPMAKER and GENEHUNTER have a practical upper limit of 9 or 10 sibs before memory requirements reach into the hundreds of megabytes; ASPEX memory usage increases more slowly and pedigrees of up to 16 sibs are easily handled. For sibships of size 10, GENEHUNTER requires more than 500 times as much memory as ASPEX.

Likelihood-based Detection of Genotype Errors

Small numbers of genotyping errors can significantly impact linkage results (i.e., Buetow 1991). Lathrop et al. (1983) proposed modeling typing errors by treating observed genotypes as phenotypes, with the true genotypes having incomplete penetrance. Lincoln and Lander (1992) used this idea to develop a multipoint likelihood-based test statistic for errors in experimental crosses, and Ehm et al. (1996) developed a similar statistic for general pedigrees. We have developed a somewhat different approach: instead of identifying individual genotypes most likely to be in error, we identify marker positions for which a whole pedigree's data appears likely to contain errors. For a pivot marker M_i, we compare the likelihood of all the data at other markers given the inheritance information at M_i, with the likelihood of the rest of the data assuming that M_i is instead uninformative:

$$E_i = \frac{\sum \left[P_{i-1}^L T(\theta_{i-1}) \circ q_i \circ P_{i+1}^R T(\theta_i) \right]}{\sum \left[P_{i-1}^L T(\theta_{i-1} + \theta_i) \circ P_{i+1}^R \right] \sum q_i}$$

Here, the numerator is just the multipoint likelihood of all the marker data, and is independent of the choice of pivot marker. These test statistics can be efficiently calculated using the same P^L and P^R vectors we use for other analyses. When checking data for a pedigree, we score each marker's error likelihood ratio E_i, then delete the most likely error, repeating until no error likelihood ratio meets a preset threshold. In exchange for a faster detection procedure, we thus sacrifice some information by discarding all of a family's data at a suspicious marker.

Power to detect genotyping errors

In simulations, we modeled the effect of genotyping errors by randomly replacing individual genotypes with random allele pairs selected with the population allele frequencies. We generated a series of 1000 datasets for 500 nuclear families with two typed parents and two sibs, for 20 equally spaced markers at various spacings, with an error frequency of 1% per genotype. We then used *sib_phase* to identify likely typing errors, using an odds ratio threshold of $E < 0.01$.

With four individuals per family, the per-family error frequency was slightly less than 4%. In our simulations, about 72% of all typing errors resulted in Mendelian inconsistencies, so the overall rate of errors consistent with Mendelian inheritance was about 1.1%. We calculated power and false positive rates by comparing the predicted errors with a record of the true errors from the simulations (table 2). The power to detect an error is directly related to the extent to which it perturbs the observed inheritance pattern. There is no power to detect errors that are consistent with surrounding data, but such "null" errors should have minimal impact on linkage-based analyses. In table 2, we also show the fraction of errors which caused a locus to either be uninformative or to have the same parental sharing as the nearest informative markers on either side, and our power to detect errors when these cases are excluded.

Our power to detect errors decreases rapidly as the intermarker distance increases. The false positive rates are very low for very dense maps, because all but the closest true multiple recombinants will have corroborating evidence from several adjacent markers. False positive

rates also fall off for larger map distances: as true recombinations become more likely, fewer sharing patterns meet the error odds ratio cutoff. For map densities less than 4 cM, most undetected errors are null errors.

Under this random error model, typing errors will introduce a bias towards increased sharing IBD. If a homozygous parent is incorrectly scored as a heterozygote, then sibs will appear to share an allele IBD for that parent when in fact they may only share IBS. Scoring a parent as a homozygote would make that parent uninformative or introduce Mendelian incompatibilities, and will not bias sharing results. Typing errors in sibs also do not bias IBD in either direction. Averaged over larger simulations, we found that with a 1% error rate, this effect yielded an apparent average sib allele sharing rate of 50.20%. While our random model may not accurately represent the distribution of errors in real data, we can predict that any error mechanism that can cause parents to be misidentified as heterozygotes will lead to excess sharing. One such mechanism is the miscalling of PCR stutter bands.

Small differences in overall sharing can have large effects on LOD scores. This is particularly important when genes of relatively small effect are studied, as a small difference in apparent sharing can push a marginal score from one side of a significance threshold to the other. Our analysis makes the conservative assumption that genotyping errors are uniformly distributed across markers: in practice, error rates vary wildly between markers (Broman et al., 1998). A marker with an unusually high error rate should have a correspondingly larger bias towards positive sharing. One might expect that in a full genome scan, regions with apparent excess sharing would even tend to be enriched with markers with relatively high error rates.

Map Construction with Sib Pair Data

Genetic maps based on sib pair data may be more accurate than standard reference maps, particularly for fine mapping. A recent comprehensive human genetic map (Broman et al. 1998) is based on eight CEPH pedigrees with about 200 meioses, while a large sib pair study can represent 1000 or more meioses. Map comparison can also be useful for identifying systematic genotyping problems. However, most mapping programs are not well suited for handling linkage data from large sib pair studies. LINKAGE (Lathrop et al. 1984) is limited to three-point analyses, and CRI-MAP (Lander and Green 1987) does not make full use of incomplete data.

The ASPEX *sib_map* program uses the same basic multipoint algorithm as *sib_phase*, to calculate global multipoint maximum likelihood maps from nuclear family data. Since we do not anticipate the use of sib pair data for de novo map construction, the *sib_map* tool is designed for validating a given map order, or positioning individual new markers on an established framework map, rather than automatically ordering large numbers of markers. We use a two-stage hill climbing algorithm to find the map distances that maximize the likelihood of the observed marker data. To ascertain the accuracy of the final map, for each intermarker map distance, we report a support interval for which the likelihood of the marker data is within a specified number of LOD units of the maximum likelihood.

In each cycle of the hill climbing algorithm, for each intermarker map distance, we calculate multipoint LOD scores for several candidate distances, holding all the other intermarker distances fixed at their best-guess values from the previous cycle. This yields a new set

of best-guess map distances, which are used in the next cycle. Typically, we converge to a global maximum LOD score in 10 or 15 cycles, independent of the number of markers mapped.

The *sib_map* program also incorporates the automated error detection algorithm for removal of unlikely recombinants. A typical run consists of three or four cycles, in which map construction is followed by detection and deletion of likely errors based on the current map. Removal of unlikely recombinants results in a shorter maximum likelihood map in the next cycle, which may reveal additional genotyping errors. The algorithm terminates when no new errors are discovered.

Support intervals based on likelihoods cannot be substituted for confidence intervals. However, they can be calculated efficiently, and give a convenient but imprecise measure of how well determined a distance estimate may be. Using simulated data, we estimated the frequency with which a *sib_map* support interval would include the true map distance. Support levels of 0.2, 0.6, and 0.8 LOD units yielded support intervals that spanned the true distances about 70%, 90%, and 95% of the time, respectively.

Transmission/Disequilibrium Tests

The original transmission/disequilibrium test (Spielman et al. 1993) measured linkage disequilibrium between a biallelic marker and a hypothetical disease locus. Subsequently, related tests were proposed for multiallelic data (Bickeböller and Clerget-Darpoux 1995; Spielman and Ewens 1996). Several groups (Kaplan et al. 1997; Morris et al. 1997; Cleves et al 1997) suggested using Monte Carlo permutation sampling to estimate empirical p-values for their multiallelic test statistics.

In the ASPEX *sib_tdt* program, we have implemented the original TDT as well as two multiallelic test statistics. For transmissions from heterozygous parents to an affected child, let n_i be the number of transmissions of allele i, and $n_{\bar{i}}$ be the number of instances where allele i was not transmitted. We implement the TDT_{max} statistic proposed by Morris et al.:

$$TDT_{max} = \max_i \left(\frac{(n_i - n_{\bar{i}})^2}{n_i + n_{\bar{i}}} \right)$$

and what we call TDT_{sum}, similar to TDT_{mhet} of Spielman and Ewens:

$$TDT_{sum} = \sum_i \left(\frac{(n_i - n_{\bar{i}})^2}{n_i + n_{\bar{i}}} \right)$$

The TDT_{max} statistic is more sensitive if association is largely with one allele, while TDT_{sum} is more sensitive if association is distributed across multiple alleles. For both composite statistics, we calculate empirical p-values with a Monte Carlo approximation. As suggested by Lazzeroni and Lange (1998), we condition our permutations on the IBD status of affected siblings in multiplex families, so the estimated p-values are insensitive to linkage evidence.

Relationship Validation

Several methods have been developed to verify pedigree structures in genotype data for nuclear families. The SibError algorithm (Ehm and Wagner, 1998) is based on comparing identity-by-state distributions across many markers with expected values based on allele frequencies. The RELPAIR algorithm (Boehnke and Cox, 1997) calculates multipoint likelihoods of observed identity-by-state data given a marker order, map distances, and a set of hypothetical relationships. In typical use with genome scan data for many polymorphic markers, the methods have similar power.

In *sib_kin*, we have implemented a pseudo-likelihood-based algorithm which neglects linkage between markers in the input data, but which is otherwise similar to the Boehnke and Cox method. For a given relationship, let α_i be the likelihood of an arbitrary marker being IBD for $i \in 0, 1, 2$ alleles. For a pair of subjects typed at n markers, at each marker j, we calculate the likelihood ω_{ij} of the marker data given that they are IBD for i alleles. Our score statistic is then:

$$T = \sum_{j=1}^{n} \log_{10} \left(\frac{\alpha_0\omega_{0j} + \alpha_1\omega_{1j} + \alpha_2\omega_{2j}}{\omega_{0j}} \right)$$

which gives a LOD score (assuming unlinked markers) relative to the null hypothesis that the pair are unrelated. Scores are typically reported for parent-child, full-sib, and half-sib relationships. We classify relative pairs by identifying the top-scoring relationship, but if the next best alternative is within 3 LOD units of the top score, a pair is scored as indeterminate.

A strength of our method is that we can calculate, for each relative pair, maximum likelihood estimates for the α parameters. This cannot be easily done using the Boehnke and Cox method, because a full likelihood expression accounting for linkage between markers cannot be easily maximized over all possible relationships. We use an EM algorithm to iteratively approximate the α parameters from the identity-by-state data and allele frequencies, using the recurrence relation:

$$\alpha_i' = \frac{1}{n} \sum_{j=1}^{n} \frac{\alpha_i\omega_{ij}}{\alpha_0\omega_{0j} + \alpha_1\omega_{1j} + \alpha_2\omega_{2j}}$$

While this does not yield a simple score statistic, the ML α parameters may be useful in diagnosing why data for a relative pair does not fit a putative relationship.

A problem with likelihood based approaches is that a single genotyping error in the input data may result in a likelihood of zero for the correct relationship. This happens for relationships where $\alpha_0 = 0$, if an error results in no alleles shared identical by state at a marker. To accommodate such errors, when scoring parent-child and monozygous-twin relationships, we adjust the α parameters such that α_0 is equal to the presumed error rate. A similar procedure was recently suggested by Broman and Weber (1998).

Discussion

In large part, the efficiency of our multipoint algorithm stems from the observation that for nuclear family data, a single informative founder is sufficient to shrink the number of allowed inheritance states from N to $\sqrt{N}$. The same approach would not be nearly as effective if

applied to family structures with more than two founders or to very incomplete data. With more founders, a much larger proportion of possible inheritance states can be consistent with the observed data and have non-zero likelihoods. Our method is also optimized for the normal situation where an individual is either typed at most markers, or completely untyped. If a dataset sharply deviates from this, ASPEX will not perform as well. There may be other useful ways of decomposing inheritance vectors in more complex pedigrees.

Two recent studies have compared the power of ASPEX with other linkage programs under various disease models and family structures (Davis and Weeks 1997; Badner et al. 1998). There are some systematic small differences among likelihood-based sib pair statistics, but generally, the SIBPAL program (SAGE 1994), SPLINK (Holmans 1993; Holmans and Clayton 1995), MAPMAKER/SIBS, and ASPEX have similar power to detect linkage and false positive rates. The SIBPAL and SPLINK programs are limited to single point analyses. If larger pedigrees are available, a program like GENEHUNTER or GENEHUNTER-PLUS (Kong and Cox 1997) will generally has greater power, particularly with rare susceptibility alleles.

Analysis of sib pair data is generally much less computationally intensive than large pedigree analysis. It could be argued that ASPEX merely provides a somewhat better solution to an already-solved problem. However, the significant speed advantages of ASPEX may make some new kinds of analysis practical. For example, permutation testing and resampling could be used to calculate empirical p-values and confidence intervals for maximum likelihood model parameters. While they are uncommon, ASPEX can also handle very large sibships that cannot be evaluated in their entirety using any other analysis program.

Proper accounting for genotyping errors in linkage analysis is likely to become increasingly important as attention shifts towards genes of smaller effect. Genotyping error rates, and distributions of error types, generally do not seem to be very well characterized. While very low genotyping error rates have been reported in some cases (i.e., Ghosh et al. 1997), more typical error rates seem to be in the 1% range, which we have shown is high enough to significantly distort marginally significant linkage results. Error detection seems particularly useful in the common situation where a low-resolution genome screen is followed up with dense genotyping of positive regions. Dense genotyping increases the odds that the most positive marker in a region may simply be the marker with the most errors in the positive direction. But it also gives much more power to detect unlikely sharing patterns.

While some of the individual ASPEX components perform analyses that can also be performed using other commonly available linkage tools, there is a real benefit to having a simple analysis framework where each tool accepts the same parameter files and data files. Once a few parameters have been set, the full range of ASPEX analyses can be performed with simple commands from the UNIX shell prompt. All the ASPEX tools have output formats that are easily parsed, and follow the UNIX convention of separating normal program output from diagnostic output, so results can often be fed directly into other analysis programs.

Program availability

Source code and documentation is available at `http://aspex.sourceforge.net/`.

References

Badner, JA, Gershon ES, Goldin LR (1998) Optimal ascertainment strategies to detect linkage to common disease alleles. Am J Hum Genet 63: 880-888

Bickeböller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidemiol 12: 865-870

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63: 861-869

Broman KW, Weber JL (1998) Estimation of pairwise relationships in the presence of genotyping errors. Am J Hum Genet 63: 1563-1564

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49: 985-994

Cleves MA, Olson JM, Jacobs KB (1997) Exact transmission-disequilibrium tests with multiallelic markers. Genet Epidemiol 14: 337-347

Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, Mitra M, et al (1998) A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. Nat Genet 19: 292-296

Davis S, Weeks DE (1997) Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. Am J Hum Genet 61: 1431-1444

Ehm MG, Kimmel M, Cottingham RW Jr (1996) Error detection for genetic data, using likelihood methods. Am J Hum Genet 58: 225-234

Ghosh S, Karanjawala, ZE, Hauser, ER, Ally D, Knapp JI, Rayman JB, Musick A et al (1997) Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. Genome Res 7: 165-178

Hauser ER, Boehnke M, Guo SW, Risch N (1996) Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. Genet Epidemiol 13: 117-137

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

Holmans P, Clayton D (1995) Efficiency of typing unaffected relatives in an affected-sib-pair linkage study with single-locus and multiple tightly linked markers. Am J Hum Genet 57: 1221-1232

Holmans P (1998) Affected sib-pair methods for detecting linkage to dichotomous traits: review of the methodology. Hum Bio 70: 1025-1040

Idury RM, Elston RC (1997) A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. Hum Hered 47: 197-202

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60: 691-702

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61: 1179-1188

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

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 (1998) Faster multipoint linkage analysis using Fourier transforms. J Comp Bio 5: 1-7

Lander ES, Green P (1987) Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 84: 2363-2367

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181

Lathrop GM, Hooper AB, Huntsman JW, Ward RH (1983) Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping. Am J Hum Genet 35: 241-262

Lathrop, GM, Lalouel, JM, Julier, C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81: 3443-3446

Lazzaroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequlibrium test. Hum Hered 48: 67-81

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14: 604-610

Morris AP, Curnow RN, Whittaker AC (1997) Randomization tests of disease-marker association. Ann Hum Genet 61: 49-60

Penrose LS (1935) The detection of autosomal linkage in a data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6: 133-138

Risch N (1990a) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46: 229-241

Risch N (1990b) 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

SAGE (1994) Statistical analysis for genetic epidemiology, version 2.2. Department of Biometry and Biostatistics, Rammelkamp Center for Education and Research, Metro Health Campus, Case Western Reserve University, Cleveland

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506-516

Spielman RS, Ewens WJ (1996) The TDT and other family based tests for linkage disequlibrium and association. Am J Hum Genet 59: 983-989

Table 1: Analysis programs included in the ASPEX package

Program	Analysis function
sib_ibd	multipoint exclusion mapping using only unambiguous IBD
sib_phase	exclusion mapping using allele frequencies to estimate IBD from IBS
sib_map	multipoint maximum likelihood genetic mapping
sib_tdt	transmission disequilibrium testing with empirical p-value estimation
sib_kin	pairwise relationship validation

Table 2: Power to detect typing errors versus marker density

Spacing	α	$1-\beta$	nulls	$1-\beta'$
1 cM	0.0018	0.50	0.37	0.81
2 cM	0.0034	0.45	0.36	0.71
4 cM	0.0051	0.29	0.32	0.42
6 cM	0.0031	0.13	0.29	0.18
8 cM	0.0008	0.05	0.27	0.07

Note. α = false positive rate; $1-\beta$ = power to detect errors; nulls = fraction of errors that are consistent with flanking linkage information; $1-\beta'$ = power to detect non-null errors.