Am. J. Hum. Genet. 56:475–483, 1995

Methods for Genetic Linkage Analysis Using Trisomies

Eleanor Feingold,[1] Neil E. Lamb,[2] and Stephanie L. Sherman[2]

[1]Division of Biostatistics, Emory University School of Public Health, and [2]Department of Genetics and Molecular Medicine, Emory University, Atlanta

Summary

Certain genetic disorders are rare in the general population, but more common in individuals with specific trisomies. Examples of this include leukemia and duodenal atresia in trisomy 21. This paper presents a linkage analysis method for using trisomic individuals to map genes for such traits. It is based on a very general gene-specific dosage model that posits that the trait is caused by specific effects of different alleles at one or a few loci and that duplicate copies of "susceptibility" alleles inherited from the nondisjoining parent give increased likelihood of having the trait. Our mapping method is similar to identity-by-descent–based mapping methods using affected relative pairs and also to methods for mapping recessive traits using inbred individuals by looking for markers with greater than expected homozygosity by descent. In the trisomy case, one would take trisomic individuals and look for markers with greater than expected homozygosity in the chromosomes inherited from the nondisjoining parent. We present statistical methods for performing such a linkage analysis, including a test for linkage to a marker, a method for estimating the distance from the marker to the trait gene, a confidence interval for that distance, and methods for computing power and sample sizes. We also resolve some practical issues involved in implementing the methods, including how to use partially informative markers and how to test candidate genes.

Introduction

Cytogenetic abnormalities have been useful tools to identify candidate regions of chromosomes that contain genes involved in single or multigene disorders. Mapping by gene dosage, making use of partially trisomic individuals, has been particularly useful in this respect. Korenberg and her colleagues have used phenotypic, cytogenetic, and molecular analyses of individuals with partial trisomy of chromosome 21 to create a phenotypic map of the chromosome (Korenberg et al. 1992; Korenberg 1993). This paper expands that approach to the use of individuals with whole trisomy to map phenotypic components of the trisomy. Obviously, defects resulting from the presence of a whole chromosome are not the result of imbalance of one or a few genes and could not be mapped easily. As pointed out by Epstein et al. (1991), phenotypic components of trisomy that are global, such as mental retardation, are due to many genetic factors. But phenotypic components of trisomy that are present in only a proportion of cases do have the potential to be mapped. Even in the case of mental retardation, it is possible that some genetic factors have a stronger effect than others and may determine a specific feature of the retardation or of development that would be possible to map.

We present linkage analysis methods to map genes involved in defects present in a proportion of cases of whole trisomy of a specific chromosome. Our approach assumes that the defect of interest is due to a gene-specific dosage effect. That is, the presence or absence of the phenotype in a trisomic individual, or variable expression of the phenotype among individuals, is caused by specific effects of different alleles at one or a few genetic loci. The different alleles may range from deleterious alleles, i.e., rare mutations, to simple isoforms, i.e., those resulting in normal variation of function. Certain trisomic genotypes lead to greater liability or susceptibility for the defect. The increased liability may be due to different levels of gene regulation, different levels of enzyme activity, or altered structural associations due to excess gene product. Carothers (1983) described a similar model for quantitative traits in trisomy 21 on the basis of dosage effects of polymorphic genes. Further, Engel (1980) suggested that, as a result of nondisjunction, two of the three chromosomes may be wholly or partially identical, which could lead to effects different from those that may be produced by three nonidentical chromosomes.

Following Engel and Carothers, we propose that the susceptible trisomic genotypes are likely to arise in cases where the two chromosomes inherited from the nondisjoining parent are partially identical, resulting in the inheritance of double copies of "susceptibility" alleles at a specific locus. These identical chromosome portions are examples of *disomic homozygosity* or, equivalently, *reduction to homozygosity,* defined at a given locus as homozygosity by

Received June 10, 1994; accepted for publication November 7, 1994.

Address for correspondence and reprints: Dr. Eleanor Feingold, Division of Biostatistics, Emory University School of Public Health, 1518 Clifton Road, Atlanta, GA 30322.

© 1995 by The American Society of Human Genetics. All rights reserved.

0002-9297/95/5602-0017$02.00

descent in the two alleles inherited from the nondisjoining parent. The term "by descent" is used here to indicate that the homozygosity is by duplication of a single parental allele and is not the result of parental homozygosity. Disomic homozygosity for a specific allele can arise in many ways. For example, it can arise from (1) mitotic nondisjunction, resulting in disomic homozygosity for the entire chromosome; (2) meiosis II nondisjunction, resulting in disomic homozygosity in the pericentromeric region; or (3) meiosis I nondisjunction that has undergone recombination, resulting in disomic homozygosity distal to the point of exchange.

The exact mechanism leading to disomic homozygosity does not affect the phenotypic outcome. Thus, regardless of the mechanism, trisomic individuals with a particular defect are expected to show greater than normal levels of disomic homozygosity in the chromosomal region containing the gene involved in the etiology of the defect. This is true under a wide variety of specific models for the trait etiology, as long as greater numbers of "susceptibility" alleles lead to a greater likelihood of having the trait. This is analogous to the expectation that affected relative pairs will show greater than expected levels of identity by descent in regions containing genes for the trait they share, even in the presence of heterogeneity, epistasis, phenocopies, or environmental interactions (see, for example, Suarez et al. 1978 and Risch 1990). The trisomy cases that arise from mitotic nondisjunction contain no mapping information, because they are homozygous for the entire chromosome, but the meiosis I and meiosis II cases can be used to map genes for the defect.

Our mapping method is quite straightforward. It is analogous to methods dating back to Penrose (1935) that look for excess identity by descent in affected relative pairs and also to methods for mapping recessive traits by using inbred individuals and looking for markers with greater than expected homozygosity by descent (Smith 1953; Lander and Botstein 1987). In the trisomy case, we take trisomic individuals and look for markers with greater than expected disomic homozygosity at a marker. In this paper, we outline such a test for linkage of a trait gene to a marker. The test is valid under any model of the trait etiology in which excess disomic homozygosity is expected. We then specialize to a simple model that can be used to estimate the distance from the marker to the trait locus and to calculate the power of the test. The natural targets for our mapping methods are well-defined defects that appear in a proportion of a trisomic population. Our methods are equally applicable to all autosomal trisomies, if appropriate phenotypic components can be identified and sufficient data found. In practice, however, because of the large number of live-born individuals involved, trisomy 21 is of the greatest interest. It will be used as an example throughout the paper. A glossary of symbols appears in the appendix.

The Data

The data needed to test linkage of a trait gene to a marker are the genotypes at the marker of trisomic individuals with the trait and of their parents. We assume that the parental origin of the nondisjunctional error is known on the basis of other informative markers on the chromosome. When the nondisjoining parent is heterozygous, we can determine whether each trisomic individual shows disomic homozygosity at that marker, as shown in table 1. Mating types in which the nondisjoining parent is homozygous at the marker yield no information and so are not included the analysis. One mating type, CD × CD, yields partial information. The offspring of the partially informative matings can be excluded from the analysis, or a slightly more complex analysis can be done that includes them. We discuss both the simple and the complex analyses.

Testing for Linkage

The Basic Test

The basic test is analogous to identity-by-descent–based linkage tests using affected relative pairs. In such tests, the proportion of pairs showing identity by descent at the marker is observed and compared to the expected proportion under the null hypothesis of no linkage. The null hypothesis expectation is $\frac{1}{2}$ for grandparent/grandchild pairs and avuncular pairs, $\frac{1}{4}$ for cousins, etc. In the trisomy case, we observe the proportion of trisomic individuals with the trait showing disomic homozygosity at the marker and compare it to the expected proportion under the null hypothesis of no linkage. The null hypothesis proportion is more complex than for affected pairs. For the moment we simply call it x and defer its calculation to the next section. We define m to be the expected proportion under the (general) alternative hypothesis. We test the null hypothesis that $m = x$, and hope to find $m > x$, which implies linkage of the trait gene to the marker being tested. The appropriate test is a one-sample z-test (one-sided), which rejects the null hypothesis if the sample proportion, $\hat{m}$, is greater than

$$x + z_\alpha \sqrt{\frac{x(1-x)}{n}},$$

where z_α is the normal percentile and n is the sample size. Alternatively, it is possible to compute a lod score, since the likelihood ratio is

$$\frac{P\{\text{observe } \hat{m} \mid m = \hat{m}\}}{P\{\text{observe } \hat{m} \mid m = x\}} = \frac{\binom{n}{\hat{m}n} (\hat{m})^{\hat{m}n}(1-\hat{m})^{n-\hat{m}n}}{\binom{n}{\hat{m}n}(x)^{\hat{m}n}(1-x)^{n-\hat{m}n}}.$$

Table 1

Genotypes Indicating Disomic Homozygosity

GENOTYPE			
Nondisjoining Parent	Disjoining Parent	Offspring	DISOMIC HOMOZYGOSITY?
CD	EF	CDE or CDF	No
		CCE, CCF, DDE, or DDF	Yes
CD	DE	CDD or CDE	No
		CCD, CCE, DDE, or DDD	Yes
CD	CD	CCD or CDD	Maybe
		CCC or DDD	Yes
CD	EE	CDE	No
		CCE or DDE	Yes
CD	DD	CDD	No
		CCD or DDD	Yes

The lod score test and the z-test are asymptotically equivalent if equivalent error levels are used. For small sample sizes and/or extreme values of x, an exact binomial test or a skewness-adjusted z-test could be used instead.

Calculation of x

Unlike the expected proportion of affected pairs who are identical by descent, the expected proportion of trisomic individuals displaying disomic homozygosity, x, is not constant over the length of the chromosome. Instead, it varies from marker to marker, depending on the distance from the centromere to the marker, and must be estimated from data. We assume that a large amount of data is available from which to estimate x; if the data set is not large, a two-sample z-test should be used for the basic test instead of a one-sample test.

The simplest approach to calculating x is to take a large sample of trisomic individuals and their parents and find the proportion of the sample that is reduced to homozygosity at each marker, thereby directly estimating x for each marker individually. This is similar to estimating the distance between two markers (the marker of interest and the centromere) by using a two-point analysis. This is likely to lead to reasonably good results for markers near the centromere, with poorer results for markers further from the centromere.

A more robust approach uses centromeric mapping methods (Shahar and Morton 1986; Chakravarti and Slaugenhaupt 1987) combined with multipoint mapping strategies to obtain genetic maps based on the trisomic population (e.g., Sherman et al. 1991). These maps can be used to derive estimates of x at each marker that incorporate data from multiple markers. The details of this approach are somewhat data dependent and in general will vary with the trisomy being studied. For example, if all trisomic individuals were the result of meiosis II errors, x would simply be $1 - z$, where z is the tetratype frequency, which is equal to the probability of heterozygosity at the marker, given homozygosity at the centromere (Shahar and Morton 1986; Chakravarti and Slaugenhaupt 1987). The tetratype frequency can be derived from the trisomy-generated genetic map by using an appropriate mapping function to translate from map distance in centimorgans to z (Morton and MacLean 1984; Chakravarti and Slaugenhaupt 1987). If only meiosis I errors existed, all trisomic individuals would be heterozygous at the centromere and thus homozygous at the marker with probability $z/2$. In general, the data will consist of a mixture of these two types of errors, giving the weighted value,

$$x = h(z/2) + (1 - h)(1 - z),$$

where h is the proportion of meiosis I errors of the total number of meiosis I and meiosis II cases. This formula is appropriate as long as the recombination process is the same in the meiosis I and meiosis II groups. There is evidence, however, that nondisjunction resulting in trisomy is associated with aberrant recombination (Warren et al. 1987; Morton et al. 1990; Sherman et al. 1994). This means that the tetratype frequency, z, may differ in the meiosis I and meiosis II cases, necessitating the use of two separate maps. The formula for x is then

$$x = h(z_1/2) + (1 - h)(1 - z_2),$$

where z_1 is the tetratype frequency in the meiosis I cases and z_2 is the tetratype frequency in the meiosis II cases.

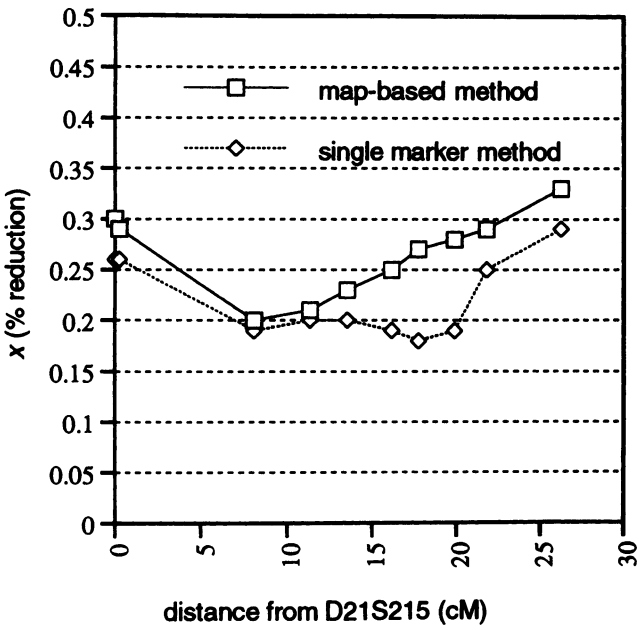

Figure 1 Values of the expected reduction, x, along chromosome 21. The markers shown are, from left to right, D21S215, D21S192, D21S214, D21S210, D21S213, IFNAR, D21S167, HMG14, D21S212, and D21S171. The horizontal axis shows distance on the average trisomy 21 map.

Figure 1 shows values of x calculated by both methods for a trisomy 21 data set of >300 cases (data described in Sherman et al. 1994). To date, there are no highly polymorphic, chromosome 21 centromeric markers available. Thus, D21S215, a pericentromeric marker, is used to represent the most centromeric marker. The most polymorphic telomeric marker available is D21S171. For the map-based method, we actually used four separate maps, for maternal and paternal meiosis I and meiosis II cases. We separated the maternal and paternal cases because there is significant reduction in recombination in the maternal, but not paternal, meiosis I cases (Sherman et al. 1994). Comparison of the estimates of x at each marker, on the basis of the two methods, shows, as one would expect, a high level of concordance for pericentromeric markers and a lesser level for more distal markers.

Including Partially Informative Individuals

The only partially informative mating type occurs when the parents share both alleles in common at the marker, i.e., CD × CD. If the offspring is DDD or CCC, it is reduced to homozygosity, but if the offspring is CCD or CDD, it could be reduced or nonreduced. If the DDD and CCC offspring are included in the sample proportion showing disomic homozygosity, $\hat{m}$, it biases $\hat{m}$ to the high side (i.e., the expected value of $\hat{m}$ is $>m$). When the methods described above are used, these individuals should be ignored. The information from CD × CD matings can be included, however, if a full likelihood approach is used to calculate $\hat{m}$. The data for this approach can be thought of as two separate samples, one of fully informative matings and one of partially informative matings, with sample sizes n_{full} and n_{part}.

For the fully informative sample, we observe v_{full}, the number of individuals in the sample with genotypes that indicate reduction. The distribution of v_{full} is binomial (n_{full}, m). For the partially informative sample, we observe v_{part}, the number of individuals with genotype DDD or CCC. The distribution of v_{part} is binomial $(n_{\text{part}}, m/2)$, since, given reduction, there is a 50% chance that the individual will be DDD or CCC rather than CCD or CDD.

The log likelihood is

$$L(m) = \ln\binom{n_{\text{full}}}{v_{\text{full}}} + \ln\binom{n_{\text{part}}}{v_{\text{part}}} + v_{\text{full}}\ln m + (n_{\text{full}} - v_{\text{full}})\ln(1 - m) + v_{\text{part}}\ln(m/2) + (n_{\text{part}} - v_{\text{part}})\ln(1 - m/2),$$

and the maximum likelihood estimate, $\hat{m}$ is the value of m that solves

$$m^2(n_{\text{full}} + n_{\text{part}}) - m(2n_{\text{full}} + n_{\text{part}} + v_{\text{full}} + 2v_{\text{part}}) + 2v_{\text{full}} + 2v_{\text{part}} = 0.$$

To test for linkage, the likelihood ratio test statistic is

$$2(v_{\text{full}} + v_{\text{part}})\ln\left(\frac{\hat{m}}{x}\right) + 2(n_{\text{full}} - v_{\text{full}})\ln\left(\frac{1 - \hat{m}}{1 - x}\right) + 2(n_{\text{part}} - v_{\text{part}})\ln\left(\frac{2 - \hat{m}}{2 - x}\right)$$

(twice the natural log of the likelihood ratio), which has an approximate χ^2 distribution with 1 df for large enough sample sizes. Alternatively, the lod score is

$$(v_{\text{full}} + v_{\text{part}})\log_{10}\left(\frac{\hat{m}}{x}\right) + (n_{\text{full}} - v_{\text{full}})\log_{10}\left(\frac{1 - \hat{m}}{1 - x}\right) + (n_{\text{part}} - v_{\text{part}})\log_{10}\left(\frac{2 - \hat{m}}{2 - x}\right).$$

Using a Specific Model to Estimate a Linkage Parameter

An important feature of our linkage test is that it does not depend on a specific model for the trait etiology. It applies to any model for which excess reduction to homo-

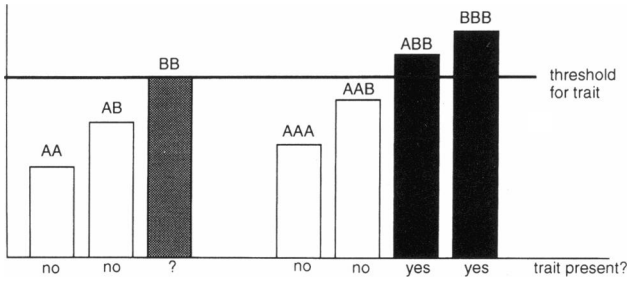

Figure 2 Hypothetical activity levels for different genotypes

zygosity is expected: essentially any gene-dosage model. However, it is desirable not only to test for linkage, but to estimate a linkage parameter: the distance from the marker to the trait locus. Estimating that distance requires using a specific model of the trait etiology. We present a simple two-allele model, but our methods can be adapted to more complex models.

The Two-Allele Model

We suppose that the trait is controlled by one locus, with two alleles. Allele A might correspond to a low dose of activity, and allele B to a high dose, with the doses being additive. Allele B is the less frequent "susceptibility" allele. A high enough total dose of activity puts an individual over a threshold and results in the abnormal phenotype. Thus, trisomic individuals who are AAA or AAB are unaffected, whereas those who are ABB or BBB are affected, as shown in figure 2. A proportion of chromosomally normal individuals with the BB genotype may exceed the threshold, depending on the variation in genotype expression. Since the B allele is less frequent, ABB and BBB individuals are primarily the result of disomic homozygosity for B.

Estimation of the Linkage Parameter

Estimating the linkage parameter requires calculating two quantities from the model of the trait etiology: the probability of having the defect, given disomic heterozygosity at the trait locus and the probability of having the defect, given disomic homozygosity at the trait locus. The general gene-dosage model implies that the latter probability should be larger than the former, but there is no other necessary relationship between these probabilities. For our simple model, the two probabilities are calculated as follows. Let the population frequencies of A and B be p and q, with $p + q = 1$. If there is reduction to homozygosity at the trait locus, the nondisjoining parent contributes either AA or BB with probabilities p and q. The other parent contributes either A or B with probabilities p and q. Then the offspring has disease genotype ABB or BBB with probabilities pq and q^2. So P{trait|homozygous at trait locus} $= pq + q^2 = q$. If there is not reduction to homozygosity at the trait locus, the nondisjoining parent contributes AA, AB, or BB with probabilities p^2, $2pq$, and q^2 (the homozygous genotypes arise here as identity by state rather than by descent). This gives the offspring genotype ABB or BBB with probabilities $2pq^2 + pq^2$ and q^3. Then P{trait|heterozygous at trait locus} $= q^3 + 3pq^2$.

The linkage parameter that we want to estimate is y, the tetratype frequency between the marker and the trait gene, defined as the probability of heterozygosity at the trait gene, given homozygosity at the marker or, equivalently, heterozygosity at the marker, given homozygosity at the gene (Chakravarti et al. 1989). It is a function of m and x and so can be estimated as a function of $\hat{m}$ and x. Since $\hat{m}$ is the maximum-likelihood estimate of m, the estimate $\hat{y}$ calculated in this manner is the maximum-likelihood estimate of y. The calculation is complicated by two additional parameters: q, the frequency of the susceptibility allele, which we presume is an unknown quantity, and K, the trisomy population prevalence of the trait, which we presume is known. Two equations relate m, x, and y, and they are both needed to solve for $\hat{y}$. For brevity, we denote homozygosity and heterozygosity by $-$ and $+$, respectively. The two equations are

$$
\begin{aligned}
m &= P\{\text{marker}-|\,\text{trait}\} \\
&= \frac{P\{\text{trait}|\,\text{marker}-\}P\{\text{marker}-\}}{P\{\text{trait}\}} \\
&\quad \text{(by Bayes's theorem)} \\
&= \frac{x}{K}\,P\{\text{trait}|\,\text{marker}-\} \\
&\quad \text{(by the definitions of } x \text{ and } K\text{)} \\
&= \frac{x}{K}\begin{pmatrix} P\{\text{gene}-|\,\text{marker}-\}P\{\text{trait}|\,\text{gene}-\} \\ +\,P\{\text{gene}+|\,\text{marker}-\}P\{\text{trait}|\,\text{gene}+\}\end{pmatrix} \\
&\quad \text{(by conditioning on the trait gene)} \\
&= \frac{x}{K}\left[(1-y)q + y(q^3 + 3pq^2)\right] \qquad (1) \\
&\quad \text{(by the definition of } y \text{ and probabilities} \\
&\quad \text{provided above by the gene-dosage} \\
&\quad \text{model), and} \\
1 - m &= P\{\text{marker}+|\,\text{trait}\} \\
&= \frac{P\{\text{trait}|\,\text{marker}+\}P\{\text{marker}+\}}{P\{\text{trait}\}} = \frac{1-x}{K} \qquad (2) \\
&\quad \times \begin{pmatrix} P\{\text{gene}-|\,\text{marker}+\}P\{\text{trait}|\,\text{gene}-\} \\ +\,P\{\text{gene}+|\,\text{marker}+\}P\{\text{trait}|\,\text{gene}+\}\end{pmatrix} \\
&= \frac{1-x}{K}\left[(y/2)q + (1-y/2)(q^3 + 3pq^2)\right].
\end{aligned}
$$

Solving for y as a function of x, K, and $\hat{m}$ first yields the intermediate step,

$$-4q^3 + 6q^2 + q = \frac{2(1 - \hat{m})K}{1 - x} + \frac{\hat{m}K}{x},$$

which must be solved numerically for $\hat{q}$, and then

$$\hat{y} = \frac{\frac{mK\tau}{x} - \hat{q}}{-2\hat{q}^3 + 3\hat{q}^2 - \hat{q}}.$$

The estimate $\hat{y}$ must be interpreted as a value on the average map that includes both meiosis I and meiosis II cases.

A Confidence Interval for the Linkage Parameter

The confidence interval for y is obtained by first calculating the confidence interval for m. For a data set that excludes partially informative cases, $\hat{m}$ is just a sample proportion, so the standard confidence interval is

$$m = \hat{m} \pm z_{\alpha/2}\sqrt{\frac{\hat{m}(1 - \hat{m})}{n}}.$$

If the partially informative cases are included and the full likelihood method used, the confidence interval for m is

$$m = \hat{m} \pm z_{\alpha/2} \times s_{\hat{m}},$$

where $s_{\hat{m}}$ is the approximate asymptotic standard error of $\hat{m}$, which is

$$\sqrt{\frac{-1}{E[L''(m)]}} \text{ evaluated at } m = \hat{m}$$
$$= \sqrt{\frac{\hat{m}(1 - \hat{m})(2 - \hat{m})}{(2 - \hat{m})n_{\text{full}} + (1 - \hat{m})n_{\text{part}}}}.$$

Substituting the left-hand endpoint of the confidence interval for m into equations (1) and (2) and solving for y gives one endpoint of the confidence interval for y. Substituting the right-hand endpoint into equations (1) and (2) gives the other. A confidence interval for q is a by-product of this procedure.

Power and Sample-Size Methods

Sample sizes for fully informative cases can be calculated by standard methods, given a particular model for the etiology of the disease. For our gene-dosage model, it is done by specifying K, x, and y and substituting them into equations (1) and (2), to give a value for m. The power to find a gene for a defect with that value of m is

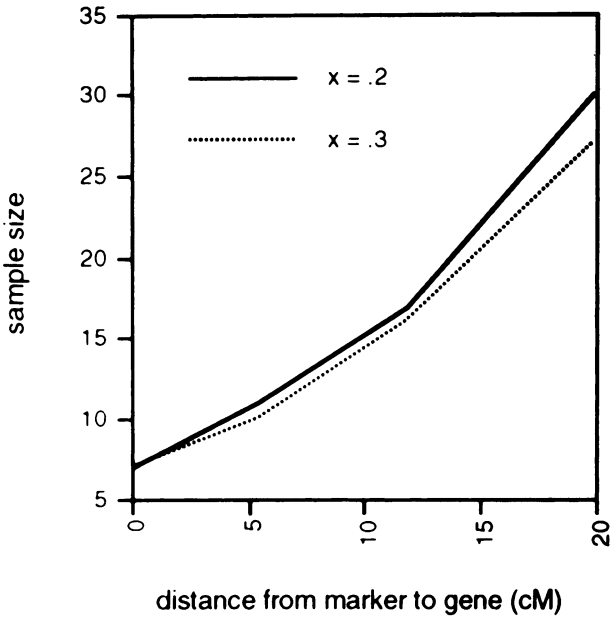

Figure 3 Sample sizes for trait frequency 1%

$$1 - \Phi\left(\frac{\sqrt{n}(x - m) + z_\alpha\sqrt{x(1 - x)}}{\sqrt{m(1 - m)}}\right), \quad (3)$$

where Φ is the standard normal distribution function. The sample size to achieve power of $1 - \beta$ is

$$n = \left(\frac{z_\beta\sqrt{m(1 - m)} - z_\alpha\sqrt{x(1 - x)}}{x - m}\right)^2. \quad (4)$$

Figures 3, 4, and 5 show sample sizes necessary to detect genes for traits with frequencies of 1%, 5%, and 15%, respectively. Mapping traits with those frequencies should be quite feasible. With markers spaced every 10 cM, so that the maximum distance from the trait gene to some marker is 5 cM, sample sizes of ~10, ~40, and ~200 are sufficient. For very common traits, necessary sample sizes are prohibitive (on the order of 5,000–10,000 cases for a trait frequency of 40%). All sample sizes are computed for power of 80%, using a significance level for the test of .01. These sample sizes, of course, apply only if our simple model is correct. More complex trait etiologies would, in general, require larger sample sizes.

The likelihood methods used for the test that includes partially informative data do not provide a way to make power and sample-size calculations. However, even if the data are expected to include both fully informative and partially informative observations, equation (4) can be used to find the necessary number of fully informative observations. Somewhat more observations will then be needed if

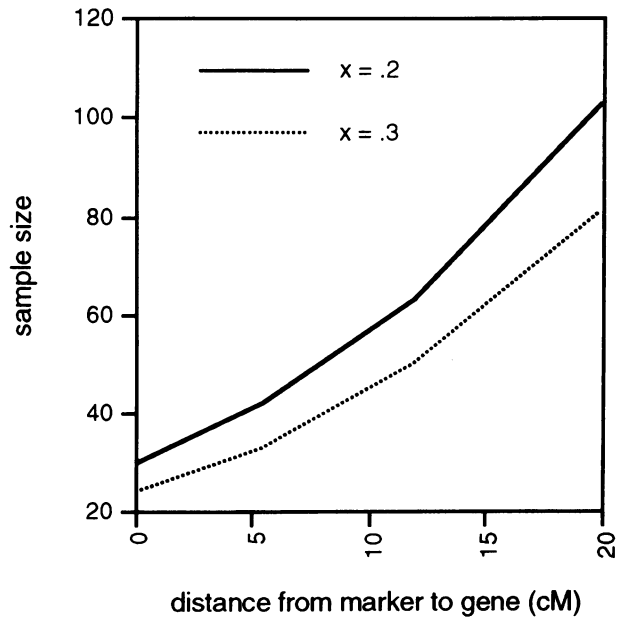

Figure 4 Sample sizes for trait frequency 5%

some of them are partially informative. The relative value of fully and partially informative observations can be found by comparing the power of the z-test for fully informative matings to the equivalent z-test for partially informative matings, that is, by setting

$$\frac{\sqrt{n_{\text{full}}}(x - m) + z_\alpha\sqrt{x(1 - x)}}{\sqrt{m(1 - m)}}$$

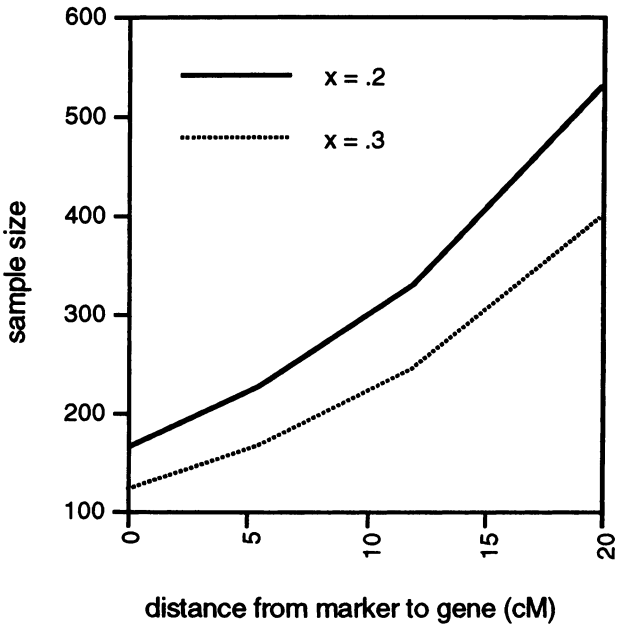

Figure 5 Sample sizes for trait frequency 15%

$$= \frac{\sqrt{n_{\text{part}}}\,\frac{(x - m)}{2} + z_\alpha\sqrt{\frac{x}{2}\left(1 - \frac{x}{2}\right)}}{\sqrt{\frac{m}{2}\left(1 - \frac{m}{2}\right)}}$$

(from eq. [3]), solving for n_{part}, and taking the derivative with respect to n_{full}. This yields the fact that one fully informative observation is equivalent to approximately

$$\frac{2 - m}{1 - m} + \frac{1}{\sqrt{n_{\text{full}}}}\left[\frac{z_\alpha\sqrt{x(2 - m)}}{(1 - m)(m - x)}\left(\sqrt{(1 - m)(2 - x)}\right.\right.$$
$$\left.\left. - \sqrt{(1 - x)(2 - m)}\right)\right]$$

partially informative observations. This value depends on n_{full}, for small values of n_{full}, but for large values of n_{full} it is approximately equal to $(2 - m)/(1 - m)$. This procedure is appropriate, as long as the partially informative observations make up only a modest percentage of the overall sample. For the values of x, y, and K for which we did power calculations, the value of $(2 - m)/(1 - m)$ ranges from ~2.5 to ~6.5.

Testing a Candidate Gene

In the case where the marker being tested is actually a candidate gene for the disorder, it is desirable to test the null hypothesis that $y = 0$. If the confidence interval for y has been calculated, one can just check whether it includes zero. Equivalently, one could calculate the value of m to which $y = 0$ corresponds (by using eqs. [1] and [2]) and then perform a z-test. This test does, however, depend on the model of the trait etiology. More nonparametric evidence for the candidate gene being involved in the defect could be obtained through a population study: examining allele distributions of the gene among trisomy 21 cases with and without the defect.

Discussion

We present a method to identify genes responsible for specific stigmata present in a proportion of trisomic individuals. This method can be applied to any autosomal trisomy for which a genetic map is available and for which specific defects are well-defined. Mapping genes responsible for the defects present in a proportion of Down syndrome individuals is probably the best application of this method, as there is a large population of live-born individuals with trisomy 21. However, other trisomies, including trisomies 13 and 18, could also be studied, and our methods could be extended to apply to sex-chromosome trisomies. It is

also possible that a series of trisomic spontaneous abortuses could be studied, if the defect in question can be diagnosed by that stage. One limitation of the method is that it is most effective for traits that appear in <15% of the trisomic population. This is due to the prohibitive sample size that is necessary to detect differences in disomic homozygosity in the most common traits. Thus, for trisomy 21, this method appears promising for traits such as leukemia or duodenal atresia, which are present in 1% and 7% of the Down syndrome population, respectively.

An important feature of our methods is that the linkage test does not depend on a specific model of the trait etiology. It is almost completely nonparametric. The only assumption it makes is that there will be increased disomic homozygosity near any gene associated with the trait. However, estimating the linkage parameter or calculating the power of the test does require relying on a specific model. It will be important in the future to extend the methods to more complex models, and to study the effects of model misspecification.

Another area for future work is developing methods that use information from more than one marker simultaneously. For example, the power of conventional linkage studies can be improved by examining flanking markers of a putative susceptibility locus, i.e., interval mapping (Lander and Botstein 1986). Moreover, it should be possible to do a whole-chromosome search, analyzing data from markers along the entire chromosome simultaneously by using methods similar to those discussed in Feingold et al. (1993).

Acknowledgments

This work is supported in part by NIH contract N01-HD 92907 and NIH grant PO1-HG 00470-01A1.

Appendix

Glossary of Symbols

- m P(disomic homozygosity) at a given marker for a trisomic individual affected with the trait of interest
- x P(disomic homozygosity) under the null hypothesis of no linkage
- n sample size (number of affected trisomic individuals)
- $\hat{m}$ maximum-likelihood estimate of m
- z tetratype frequency between marker and centromere
- z_1 tetratype frequency between marker and centromere in meiosis I cases
- z_2 tetratype frequency between marker and centromere in meiosis II cases
- h proportion of meiosis I cases of the total number of meiosis I and meiosis II cases
- n_{full} number of fully informative cases
- v_{full} number of fully informative cases showing disomic homozygosity at the marker
- n_{part} number of partially informative cases
- v_{part} number of partially informative cases showing disomic homozygosity at the marker
- q frequency of the rarer "susceptibility" gene under the two-allele model
- y linkage parameter (tetratype frequency between marker and trait gene)
- K trait frequency in trisomic population
- $\hat{y}$ maximum-likelihood estimate of y
- $\hat{q}$ maximum likelihood estimate of q

References

Carothers AD (1983) Gene dosage effects in trisomy: comment on a recent article by BL Shapiro. Am J Med Genet 16:635–637

Chakravarti A, Majumder PP, Slaugenhaupt SA, Deka R, Warren AC, Surti U, Ferrell RE, et al (1989) Gene-centromere mapping and the study of non-disjunction in autosomal trisomies and ovarian teratomas. In: Hassold TJ, Epstein CJ (eds) Molecular and cytogenetic studies of non-disjunction. AR Liss, New York, pp 45–79

Chakravarti A, Slaugenhaupt SA (1987) Methods for studying recombination on chromosomes that undergo nondisjunction. Genomics 1:35–42

Engel E (1980) A new genetic concept: uniparental disomy and its potential effect. Am J Med Genet 6:137–143

Epstein CJ, Korenberg JR, Annerén G, Antonarakis SE, Aymé S, Courchesne E, Epstein LB, et al (1991) Protocols to establish genotype-phenotype correlations in Down syndrome. Am J Hum Genet 49:207–235

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for linkage analysis using high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Korenberg JR (1993) Toward a molecular understanding of Down syndrome. In: Epstein CJ (ed) The phenotypic mapping of Down syndrome and other aneuploid conditions. Wiley-Liss, New York, pp 87–115

Korenberg JR, Bradley C, Disteche CM (1992) Down syndrome: molecular mapping of the congenital heart disease and duodenal stenosis. Am J Hum Genet 50:294–302

Lander ES, Botstein D (1986) Mapping complex traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 51:49–62

——— (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570

Morton NE, Keats BJ, Jacobs PA, Hassold T, Pettay D, Andrews V (1990) A centromere map of the X chromosome from trisomies of maternal origin. Ann Hum Genet 54:39–47

Morton NE, MacLean CJ (1984) Multilocus recombination frequencies. Genet Res 44:99–108

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

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

Shahar S, Morton NE (1986) Origin or teratomas and twins. Hum Genet 74:215–218

Sherman SL, Petersen MB, Freeman SB, Hersey J, Pettay D, Taft L, Frantzen M, et al (1994) Non-disjunction of chromosome 21 in maternal meiosis I: Evidence for a maternal-age dependent mechanism involving reduced recombination. Hum Mol Genetics 3:1529–1535

Sherman SL, Takaesu N, Freeman SB, Grantham M, Phillips C, Blackston RD, Jacobs PA, et al (1991) Trisomy 21: association between reduced recombination and nondisjunction. Am J Hum Genet 49:608–620

Smith CAB (1953) The detection of linkage in human genetics. J Royal Stat Soc 15:153–184

Suarez BK, Rice J, Reich T (1978) The generalized sib-pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94

Warren AC, Chakravarti A, Wong C, Slaugenhaupt SA, Halloran SL, Watkins PC, Metaxotou C, et al (1987) Evidence for reduced recombination on the nondisjoined chromosomes 21 in Down syndrome. Science 237:652–654