Am. J. Hum. Genet. 69:1255–1265, 2001

Multipoint Genetic Mapping with Trisomy Data

Jinming Li,[1] Stephanie L. Sherman,[2] Neil Lamb,[2] and Hongyu Zhao[1]

[1]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven; and [2]Department of Genetics, Emory University School of Medicine, Atlanta

Trisomy is the most common genetic abnormality in humans and is the leading cause of mental retardation. Although molecular studies that use a large number of highly polymorphic markers have been undertaken to understand the recombination patterns for chromosome abnormalities, there is a lack of multilocus approaches to incorporating crossover interference in the analysis of human trisomy data. In the present article, we develop two statistical methods that simultaneously use all genetic information in trisomy data. The first approach relies on a general relationship between multilocus trisomy probabilities and multilocus ordered-tetrad probabilities. Under the assumption that no more than one chiasma exists in each marker interval, we describe how to use the expectation-maximization algorithm to examine the probability distribution of the recombination events underlying meioses that lead to trisomy. One limitation of the first approach is that the amount of computation increases exponentially with the number of markers. The second approach models the crossover process as a χ^2 model. We describe how to use hidden Markov models to evaluate multilocus trisomy probabilities. Our methods are applicable when both parents are available or when only the nondisjoining parent is available. For both methods, genetic distances among a set of markers can be estimated and the pattern of overall chiasma distribution can be inspected for differences in recombination between meioses exhibiting trisomy and normal meioses. We illustrate the proposed approaches through their application to a set of trisomy 21 data.

Introduction

Trisomy is the most commonly identified chromosome abnormality in humans, occurring in 0.3% of live births, 4% of stillbirths, and as many as 25% of spontaneous abortions (Hassold and Jacobs 1984). To produce a trisomic offspring, the parent in whom nondisjunction occurs transmits a disomic gamete, whereas the other parent transmits the usual monosomic gamete. Recent studies of trisomy 21 have shown that both altered levels of recombination and altered exchange patterns are associated with maternal nondisjunction (Lamb et al. 1996, 1997). Although these studies have revealed that the recombination patterns among meioses that lead to nondisjunction may be different from those among normal meioses, existing statistical treatments of trisomy data are not entirely satisfactory, as will be reviewed in the next paragraph. The objective of the present article is to develop general statistical approaches that overcome the limitations of the existing methods for the analysis of trisomy data.

Received July 26, 2001; accepted for publication September 26, 2001; electronically published November 5, 2001.

Address for correspondence and reprints: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, 60 College Street, Yale University School of Medicine, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

© 2001 by The American Society of Human Genetics. All rights reserved.
0002-9297/2001/6906-0011$02.00

Genetic mapping methods for nondisjoined chromosomes have been discussed by Ott et al. (1976), Shahar and Morton (1986), Chakravarti and Slaugenhaupt (1987), Chakravarti et al. (1989), Feingold et al. (2000), Yu and Feingold (2001), and other researchers. In most studies, genetic map construction is divided into two steps. In the first step, the more-proximal marker is treated as a pseudocentromere, and pairwise LOD scores are calculated for each pair of markers, through the observed patterns of nonreduction (heterozygous genotype) and reduction (homozygous genotype) of markers along the nondisjoined chromosome pair. In the second step, these pairwise LOD scores are compiled to derive an estimated genetic map. The limitations of such methods are as follows: (1) Instead of using multilocus information jointly, they only use markers sequentially; thus, many informative cases are discarded in the pairwise analysis, because not all the markers are typed or are informative. (2) The procedures in compiling pairwise LOD scores are ad hoc, and the direction of bias is difficult to evaluate. (3) Crossover interference can be accounted for only at the stage where pairwise distances are combined, although crossover interference has been observed in humans (e.g., Hulten 1974; Broman and Weber 2000). (4) Joint recombination patterns across a set of intervals cannot be recovered from such analysis. Chakravarti et al. (1989) proposed two approaches for multilocus analysis. One was to assume at most three chiasmata across the region under study,

with no more than one chiasma in a given marker interval. The other was to treat the proximal marker as a pseudocentromere, relative to the distal marker. The first approach is not applicable to chromosomes likely to have more than three chiasmata or for studies involving large marker intervals, whereas the second approach implicitly assumes the absence of chiasma interference. Feingold et al. (2000) derived multipoint likelihoods for trisomy data, under the assumption of no crossover interference, and their method allows for partial information and the absence of one parent for individuals with trisomy. However, the genetic-distance estimates from their approach may be biased, because crossover interference does appear to occur during normal human meiosis (Broman and Weber 2000). In a recent article, in which they assumed that the total number of crossovers on a chromosome in a meiosis is observable, Yu and Feingold (2001) discussed issues related to the estimation of tetrad crossover–frequency distributions from genetic recombination data, including nondisjunction data. Given the limitations of the existing methods, in the context of analyzing uniparental disomy (UPD) data, Zhao et al. (2000) first established a general relationship between multilocus UPD probabilities and multilocus ordered-tetrad probabilities and then showed how to use the expectation-maximization (EM) algorithm (Dempster et al. 1977) to estimate joint recombination probabilities under the assumption that there is at most one chiasma within each marker interval. Because the amount of computation required by this approach increases exponentially with the number of genetic markers, Zhao et al. (2001) described how to use a hidden Markov model (HMM) to evaluate multilocus UPD probabilities when the chiasma process is assumed to follow the χ^2 model (Zhao et al. 1995). In the present article, we extend these two approaches from our previous work on UPD data to trisomy data. Both approaches can simultaneously consider all genetic markers for all individuals and consistently incorporate crossover interference in the analysis.

In the Methods section, we describe the two approaches for trisomy data in detail. In the Results section, we summarize the performance of the HMM approach, under a variety of simulation scenarios. We then apply our method to the analysis of a trisomy 21 data set. Finally, in the Discussion section, we conclude with comments on our methods and related issues.

Methods

Notation for Multilocus Ordered-Tetrad Data

In the present article, markers are denoted by script letters. For example, we use $\mathcal{A}$ to denote a genetic marker. Alleles are denoted by italic letters. For example, A and a denote two alleles of marker $\mathcal{A}$. We use [*E, F; H, W*] to denote the observed marker configuration for an ordered tetrad, where E and F are attached to one centromere and H and W are attached to the other centromere. For example, [*AB, Ab; aB, ab*] represents an ordered tetrad with two strands carrying AB and Ab attached to one centromere and with two strands carrying aB and ab attached to the other centromere. The centromere is denoted by CEN. For patterns between a pair of markers, we use P to denote a parental ditype in which all four strands retain the parental type, T to denote a tetratype in which two of the four strands show recombination, and N to denote a nonparental ditype in which all four strands are recombinants.

For a genetic marker $\mathcal{A}$ that segregates with two alleles A and a, there are six distinguishable patterns for ordered tetrads: (1) [*A, A; a, a*], (2) [*A, a; A, a*], (3) [*A, a; a, A*], (4) [*a, A; A, a*], (5) [*a, A; a, A*], and (6) [*a, a; A, A*]. Patterns 1 and 6 are called the "first division segregation" (FDS) pattern, and patterns 2–5 are called the "second division segregation" (SDS) pattern (Griffiths et al. 1996). For ordered tetrads, we distinguish $2 \times 3^{n-1}$ states for n markers in the order of $CEN, \mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$. Each of these $2 \times 3^{n-1}$ states is represented by $J_n = (j_1, j_2, \ldots, j_n)$, where $j_1 = 0$ or 1 corresponds to FDS or SDS, respectively, at $\mathcal{A}_1$, and $j_r = 0$, 1, or 2 corresponds to *P, T,* or *N*, respectively, between $\mathcal{A}_{r-1}$ and $\mathcal{A}_r$, for $r = 2, \ldots, n$. We denote the probability of ordered-tetrad state J_n by p_{J_n}.

Notation for Multilocus Data for the Two Chromosomes from the Nondisjoining Parent

Consider n markers, with each marker—$\mathcal{A}_r$, $r = (1, \ldots, n)$—being heterozygous with alleles $\mathcal{A}_r$ and a_r in the nondisjoining parent (NDJP). At any given locus, the two chromosomes inherited from the NDJP are described as "reduced to homozygosity" (denoted by R) if they are identical by descent, or as "nonreduced" (denoted by N) if they are not identical by descent. When the phases in the parent are unknown, we distinguish 2^n distinct states for joint genotypes on the two nondisjoined chromosomes. Each of these states is denoted by $I_n = (i_1, i_2, \ldots, i_n)$, where $i_k = 0$ or 1 corresponds to the kth marker being R or N, respectively. The probability for each pattern I_n is denoted by u_{I_n}. Note that throughout this article, we use "*J*" to denote an ordered-tetrad state and "*I*" to denote a state for the two chromosomes from the NDJP. Their corresponding probabilities are denoted by p_J and u_I, respectively. Meiotic nondisjunction events are classified as meiosis I (MI) nondisjunction, if the two copies of the same chromosome are homologous, and as meiosis II (MII) nondisjunction, if the two copies are sister chromatids (Orr-Weaver 1996).

Elsewhere, we have derived general relationships between u_I and the p_J for both MI and MII nondisjunction events, and we have described how to use these relationships to estimate, on the basis of UPD data, joint ordered-tetrad–state probabilities at the four-strand stage during meiosis (Zhao et al. 2000).

Notation for Multilocus Trisomy Data

Because of the complexity of trisomy data, we may not always identify whether a marker is *R* or *N* in an individual with trisomy, even if the NDJP is heterozygous at this marker. Feingold et al. (2000) gave a comprehensive discussion on the format of trisomy data. In the next paragraph, we briefly summarize the notation discussed by Feingold et al. (2000), and detailed explanations can be found in their article.

For an individual with trisomy, either one or two parents are available for study. When two parents are available, we distinguish six mating types between the NDJP and the correctly disjoining parent (CDJP): (1) *ab* × *cd*, (2) *ab* × *bc*, (3) *ab* × *cc*, (4) *ab* × *bb*, (5) *ab* × *ab*, and (6) *aa* × anything. For the first four mating types, we can always unambiguously determine whether the two chromosomes inherited from the NDJP are *R* or *N*. The fifth mating type is an intercross. It can produce an *R*, when the trisomy individual's genotype is *aaa* or *bbb*, or an *X*, when the trisomy individual's genotype is *aab* or *abb*. The marker status of *X* indicates that the marker is partially informative. Although we do not have unequivocal information about whether the true state is *N* or *R*, some information is added because the probability of observing an *X* depends on what the true state is. For the sixth mating type, the NDJP is homozygous, and the marker is completely uninformative, which we denote by *U*. Untyped markers can also be considered uninformative. When only one parent is available, we distinguish four mating types between the NDJP and the CDJP: (1) *ab* × missing, (2) *aa* × missing, (3) missing × *aa*, and (4) missing × *ab*. As in the case when two parents are available, the genotype at a marker in an individual with trisomy can be coded in four ways: *N*, *R*, *U*, and *X*. When the above notation is used, each individual with trisomy can be represented as a character string, with the use of *R*, *N*, *X*, and *U*, for example, "*…NRNUNX….*"

Maximum-Likelihood Estimates of Multilocus Ordered-Tetrad Probabilities from Trisomy Data if There Is at Most One Chiasma within Each Marker Interval

Assume that in our trisomy data there are a total of *S* cases, each typed at some of the *n* genetic markers. Elsewhere, we have established general relationships between multilocus UPD probabilities and multilocus ordered-tetrad probabilities for both MI and MII errors (Zhao et al. 2000). Assuming at most one chiasma between adjacent markers, we described how to use the EM algorithm to estimate multilocus ordered-tetrad probabilities from UPD data (Zhao et al. 2000). Using this method, we can estimate the genetic distances between consecutive markers and can obtain the overall chiasma distribution. Because of the complexity of trisomy data, we have to modify the EM algorithm described elsewhere (Zhao et al. 2000) to analyze trisomy data. The details on the E step and the M step are described in Appendix A. In general, for either MI or MII nondisjunction, we start the EM algorithm with initial estimates of multilocus probabilities p_J^0. The E step computes the expected number of each possible ordered-tetrad pattern *J*, conditional on the observed trisomy data and the initial values p_J^0. The M step maximizes the likelihood of this "expected" data set and thus generates updated estimates of p_J. These new estimates are fed back into the E step, and the algorithm iterates until convergence. From the maximum-likelihood estimates of ordered-tetrad probabilities p_J, we can derive the distribution of the number of chiasmata on the chromosome and the joint distribution of chiasmata among the marker intervals. It is also straightforward to estimate the genetic distance between each pair of consecutive markers and to estimate the total genetic distance between the centromere and the most distant marker. The detailed procedure for these estimates has been reported elsewhere (Zhao et al. 2000).

The HMM Approach to Trisomy Data

Although no specific models are assumed in the analysis of trisomy data in the above approach, the amount of computation increases exponentially with the number of markers. For human UPD data, an HMM, in which the crossover process is modeled by the χ^2 model, has been developed (Zhao et al 2001). The χ^2 model for crossovers has a long history (Bailey 1961). Foss et al. (1993) represented the model in the form of $Cx(Co)^m$ as follows: assume that the crossover intermediates (*C* events) are randomly distributed along the four-strand bundle and that every intermediate resolves either as a crossover (denoted by "*Cx*") or as a noncrossover (denoted by "*Co*"). When an intermediate resolves as a *Cx*, the next *m* intermediates must resolve as a *Co*, and after *m Co*s, the next intermediate must resolve as a *Cx*. The process is made stationary by allowing the leftmost crossover intermediate an equal chance to be one of $Cx(Co)^m$. Note that the Poisson (no interference) model corresponds to $m = 0$. Among the many crossover process models that have been proposed in the literature, the χ^2 model has been found to provide a good fit to data from many organisms (Zhao et al. 1995). Because there are no partially informative markers in UPD data,

we need to extend the HMM approach to accommodate human trisomy data, as described in the next paragraph.

As discussed by Rabiner (1989), an HMM has the following five components: (1) the set of hidden states $S = \{S_1, \ldots S_i \ldots, S_L\}$ and $1 \leq i \leq L$; (2) the set of distinct observation symbols $V = \{V_1, \ldots V_k, \ldots, V_M\}$ and $1 \leq k \leq M$; (3) the state transition probability distribution $A = \{a_{ij}\}$, where $a_{ij} = P(q_{r+1} = S_j | q_r = S_i)$, $1 \leq i$, and $j \leq L$ and where q_r denotes the hidden state at time r; (4) the observation symbol probability distribution in state S_j $B = \{b_j(V_k)\}$, where $b_j(V_k) = P(O_r = V_k | q_r = S_j)$, $1 \leq j \leq L$, and $1 \leq k \leq M$ and where O_r denotes the observation symbol at time r; and (5) the initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i)$. The exact forms of these five components of UPD data were discussed by Zhao et al. (2001). In Appendix B, we discuss the forms of these five components of trisomy data. Having defined the five elements of the HMM, we can use the forward algorithm to calculate the probability of any trisomy genotype and can obtain the maximum-likelihood estimates of the parameters of interest—that is, the genetic distances among the markers and the interference parameter m. If we let $m = 0$ in the χ^2 model, then our HMM approach is the same as the no-interference model discussed by Feingold et al. (2000).

Results

Simulation Results

In this section, we summarize the simulation results under the HMM approach for trisomy data with eight equally spaced markers. The genetic distance between each pair of consecutive markers is 10 cM. We varied the sample size ($S = 200$ or 400), the interference parameter ($m = 0$, 1, or 2), and the proportion of missing data ($\mu = 0.2$ or 0.4) in our simulations. In addition, to examine the effect of partially informative markers, we varied the proportion of intercross-mating type c ($c = 0.1$ or 0.2), when both parents are available. We assumed that the parameter c is the same across all the markers. For each parameter combination, we generated 100 simulated data sets and estimated the genetic distances among the markers. The results are summarized in table 1. We list only the estimates for the fourth interval, for both MI and MII nondisjunction. It can be seen that when both parents are available, the maximum-likelihood estimates of genetic distances are almost unbiased.

If only the NDJP is available, instead of varying the proportion of intercross-mating types c in our simulations, we varied the parameter q (the probability that the CDJP contributes an allele that is different from either of the alleles of the NDJP), to consider the effect of partially informative markers. We assumed that the parameter q is the same across all the markers. The results are summarized in table 2. We list only the estimates of the fourth interval of both MI and MII nondisjunctions. Even when the CDJP is missing, the maximum-likelihood estimates of the genetic distances still performed quite well.

Table 1

Simulation Results for MI and MII Nondisjunctions When Both Parents Are Available

Order[a]	Probabilities[b]	Sample Size	$d_{M1}(sd_{M1})$[c]	$d_{M2}(sd_{M2})$[d]
0	$\mu = .2$, $c = .1$	400	10 (.1)	10 (.1)
		200	10 (.3)	10 (.2)
0	$\mu = .4$, $c = .1$	400	10 (.3)	10 (.2)
		200	11 (.5)	10 (.4)
0	$\mu = .2$, $c = .2$	400	11 (.2)	10 (.2)
		200	11 (.3)	10 (.3)
0	$\mu = .4$, $c = .2$	400	10 (.4)	10 (.2)
		200	10 (.6)	11 (.4)
1	$\mu = .2$, $c = .1$	400	10 (.2)	10 (.2)
		200	11 (.3)	10 (.2)
1	$\mu = .4$, $c = .1$	400	10 (.2)	10 (.2)
		200	10 (.3)	10 (.3)
1	$\mu = .2$, $c = .2$	400	9 (.2)	10 (.1)
		200	11 (.4)	10 (.2)
1	$\mu = .4$, $c = .2$	400	11 (.2)	10 (.2)
		200	10 (.3)	10 (.3)
2	$\mu = .2$, $c = .1$	400	10 (.2)	10 (.2)
		200	11 (.2)	11 (.2)
2	$\mu = .4$, $c = .1$	400	10 (.2)	10 (.1)
		200	10 (.3)	10 (.2)
2	$\mu = .2$, $c = .2$	400	10 (.1)	10 (.2)
		200	10 (.2)	9 (.2)
2	$\mu = .4$, $c = .2$	400	10 (.2)	10 (.2)
		200	10 (.3)	10 (.3)

[a] Order as indicated by the χ^2 model.

[b] μ is the probability of being a completely uninformative marker, and c is the probability of being an intercross-mating type.

[c] d_{M1} indicates the average of the estimates of genetic distance of the fourth marker interval for MI nondisjunction, and sd_{M1} indicates the associated standard deviation of the genetic distance estimates.

[d] d_{M2} is the average of the estimates of genetic distance of the fourth interval for MII nondisjunction, and sd_{M2} is the associated standard deviation of the genetic distance estimates.

Application of Simulation Results to Trisomy 21 Data

Trisomy 21 is the most common viable chromosomal abnormality in humans and is responsible for >95% of instances of Down syndrome. The incidence is ~1 per 600 live births (Sherman et al. 1994). Recent molecular studies indicate that most trisomy 21 is maternally derived and is typically a result of nondisjunction at MI (Lamb et al. 1996). In this section, we apply our two approaches to a maternal MI trisomy data set that consists of 434 individuals with trisomy. Ten intervals spanning the full length of the chromosome were included in our analysis. To maximize the amount of linkage information for a given interval, we grouped several mark-

Table 2

Simulation Results for MI and MII Nondisjunction When Only the NDJP Is Available

Order	Probabilities	Sample Size	$d_{M1}(sd_{M1})$	$d_{M2}(sd_{M2})$
0	$\mu = .2, q = .8$	400	10 (.2)	10 (.2)
		200	11 (.3)	11 (.3)
0	$\mu = .4, q = .8$	400	11 (.3)	10 (.2)
		200	10 (.5)	10 (.3)
0	$\mu = .2, q = .6$	400	10 (.3)	10 (.2)
		200	10 (.4)	10 (.3)
0	$\mu = .4, q = .6$	400	10 (.3)	10 (.2)
		200	11 (.5)	10 (.4)
1	$\mu = .2, q = .8$	400	11 (.2)	10 (.2)
		200	10 (.2)	11 (.3)
1	$\mu = .4, q = .8$	400	10 (.2)	10 (.2)
		200	10 (.3)	10 (.3)
1	$\mu = .2, q = .6$	400	11 (.2)	10 (.2)
		200	10 (.3)	9 (.3)
1	$\mu = .4, q = .6$	400	11 (.3)	9 (.3)
		200	10 (.3)	10 (.4)
2	$\mu = .2, q = .8$	400	10 (.1)	10 (.2)
		200	10 (.2)	10 (.2)
2	$\mu = .4, q = .8$	400	10 (.2)	10 (.2)
		200	11 (.3)	10 (.3)
2	$\mu = .2, q = .6$	400	10 (.2)	10 (.2)
		200	9 (.3)	9 (.3)
2	$\mu = .4, q = .6$	400	9 (.3)	9 (.2)
		200	10 (.3)	10 (.3)

NOTE.—Data are as described in footnotes to table 1.

ers into a region, defined as a set of markers known to be tightly linked in normal individuals and among which no recombination was observed in the trisomic data set. The markers and marker groups used in our analysis are *D*21s13/*D*21s16/*D*21s192, *D*21*s*214/*D*21*s*232, *D*21*s*210, *D*21*s*213, *D*21*s*223/*D*21*s*224/*IFNAR*, D21*s*17/D21*s*167, ETS2/D21*s*156, HMG14, D21*s*212/D21*s*113, and D21s1575/D21*s*1446. We describe our results based on the two approaches separately.

Maximum-likelihood estimates of multilocus ordered-tetrad probabilities assuming at most one chiasma within each marker interval.—The results are summarized in table 3. The total genetic length was estimated as 36.3 cM for the MI trisomy data set, which is shorter than that of the normal female map. Several estimates of the total length of the normal female map are available: the Marshfield map estimates the total genetic length at 64.6 cM, the analysis based on normal female meiotic events that uses genotype data from CEPH puts the estimate at 72.1 cM, and the total genetic length of the most recent normal female map, maintained at the Web site of the Genetic Location Database, is 59.8 cM.

From the maximum-likelihood estimates of ordered probabilities p_{Jn}, we can also study the distribution of chiasma number and the exchange patterns for a given number of chiasmata. For these ten intervals, the estimated proportions of tetrads with 0, 1, 2, and 3 chiasmata are 41.7%, 46.2%, 9.9%, and 2.0%, respectively, for this MI trisomy data set. The majority of the tetrads (87.9%) had zero or one chiasma. For a given number of chiasmata (1 or 2), we summarize the distribution of the chiasmata for MI trisomy in table 4. The exchange patterns conditional on three chiasmata are not listed, because they are much less reliable, being derived from only a small proportion of the total cases. When there was a single chiasma on the whole chromosome, it most likely occurred in the last five intervals. For the two-chiasmata case, the modes of the first chiasma was in the first two intervals, and the majority of the second chiasma occurred in the fourth, sixth, and ninth intervals.

The HMM approach.—The estimated genetic distances among the markers and the associated standard errors under different χ^2 models are shown in table 5. For MI trisomy, the total estimated genetic length across these 10 markers ranges from 38.4 cM to 44.3 cM for different *m* values. The maximized log-likelihoods from different χ^2 models were very similar, with the no-interference model having the largest log-likelihood (-456.5). The log-likelihood of *CxCo* model is -459.4, which is very close to that of the no-interference model.

Under the assumption of no crossover interference, Feingold et al. (2000) analyzed the same MI trisomy data set. However, our analyses differed from those of Feingold et al. (2000) in the number of marker groups analyzed: we analyzed 11 marker groups for the present report, whereas Feingold et al. (2000) reported analysis of 14 marker groups. This was because it is not feasible for our program to analyze >10 markers, using the EM approach, because of the large computer memory required for the EM approach; therefore, we were forced to combine several marker groups. Using 14 marker groups, Feingold et al. (2000) estimated the total genetic length to be 45.1 cM. Note that we converted their estimated *y* values to genetic distances and summed over the intervals to arrive at the value of 45.1 cM. If we had

Table 3

Genetic Distance Estimates in the MI Trisomy 21 Data Set

Interval	*Cx*[a]	EM	Soton Map[b]
d_1	2.5 (.5)	1.8 (.9)	5.4
d_2	3.5 (.7)	3.1 (1.2)	23.3
d_3	1.4 (.4)	1.9 (.9)	1.9
d_4	2.5 (.5)	2.2 (.9)	8.9
d_5	1.6 (.4)	1.3 (.9)	.8
d_6	9.4 (.7)	7.3 (1.6)	7.2
d_7	2.0 (.2)	1.8 (.9)	.9
d_8	1.7 (.0)	1.7 (1.0)	.4
d_9	8.9 (.2)	7.6 (1.8)	9.7
d_{10}	7.4 (.2)	7.5 (2.2)	1.4
Total	40.9	36.3	59.8

[a] No interference model.

[b] Represents genetic distances in the normal female map.

Table 4

The Exchange Patterns for Chiasmata of MI Trisomy Data When the Total Number of Chiasmata on the Tetrads is $k = 0$, 1, or 2

	Frequency, Conditional on k (EM Method), in Interval									
k (Distribution)	1	2	3	4	5	6	7	8	9	10
0 (.417)	…	…	…	…	…	…	…	…	…	…
1 (.462)	.00	.01	.08	.03	.05	.25	.04	.05	.20	.28
2 (.099):										
First	.36	.47	.00	.08	.00	.00	.00	.00	.08	.00
Second	.00	.00	.00	.23	.00	.30	.08	.00	.31	.08
Map distances (cM)	1.8	3.1	1.9	2.2	1.3	7.3	1.8	1.7	7.6	7.5
	Frequency, Conditional on k (HMM method with $m = 0$), in Interval									
	1	2	3	4	5	6	7	8	9	10
0 (.452)	…	…	…	…	…	…	…	…	…	…
1 (.377)	.06	.08	.03	.06	.04	.23	.05	.04	.22	.18
2 (.124):										
First	.12	.18	.06	.10	.06	.29	.05	.04	.09	.00
Second	.00	.02	.00	.02	.01	.15	.04	.03	.34	.39
Map distances (cM)	2.5	3.5	1.4	2.5	1.6	9.4	2.0	1.7	8.9	7.4

Note.—. For each k, this table summarizes the marginal distribution of the jth chiasma, where $j = 1, \ldots, k$.

fixed the order in our HMM approach at $m = 0$ and had used the same marker groups as those used by Feingold et al. (2000), we would have arrived at the same results as those obtained by Feingold et al. (2000) with the HMM approach.

For the HMM approach, we can also study the distribution of chiasma number and the exchange patterns for a given number of chiasmata through examining the estimates of ordered-tetrad probabilities. When we fix the HMM model at $m = 0$ for this MI trisomy data set, the estimated proportions of tetrads with 0, 1, 2, and 3 chiasmata are 45.2%, 37.7%, 12.4%, and 0.3%, respectively. The majority (82.9%) of the tetrads had zero or one chiasma, which is similar to the estimates based on the EM approach. The distribution of the chiasmata for MI trisomy, given the chiasmata number, is summarized in table 4. Because the proportion of tetrads that have more than two chiasmata is as small as 0.3%, we did not list the related results in table 4.

Table 3 lists the comparisons between the results obtained with the EM approach and the HMM approach for the MI trisomy data. For the HMM approach, we chose the *Cx* model to perform the comparison. From table 3 we can see that our proposed two approaches generate consistent results for most intervals.

Discussion

In the present article, we describe two general multilocus approaches that make use of all the marker information in the observed trisomy data and that can consistently incorporate crossover interference. The first approach is based on the relationships we have established between multilocus half-tetrad probabilities and multilocus ordered-tetrad probabilities (Zhao et al. 2000). When many markers are available in genetic studies, we may assume that there is at most one chiasma in each marker interval on the tetrads. For this particular model, we have described how to use the EM algorithm to estimate multilocus ordered-tetrad probabilities from the observed trisomy data. Our approach can analyze data that contain many untyped and uninformative markers and partially informative markers. From the maximum-likelihood estimates of ordered-tetrad probabilities, we can study the probability distribution of the number of chiasmata and the exchange patterns along the chromosome. Our EM-based method can be used when both parents are available or when only the CDJP is available. The second approach is based on the χ^2 model for the crossover process. One advantage of this approach is that the amount of computation increases linearly with the number of markers, allowing us to include many genetic markers in the same analysis. Our method includes the method developed by Feingold et al. (2000) as a special case. This is because the χ^2 model when $m = 0$ corresponds to the no-interference model discussed by Feingold et al. (2000). We have implemented our methods in computer programs, which can be downloaded from our Web site.

The limitation of the previous approaches for trisomy analysis was demonstrated by Feingold et al. (2000), who performed an extensive simulation study to compare their multipoint NDJMap method with other methods, such as TETRAD and DSLINK. Throughout their simulations, NDJMap generated completely consistent and unbiased results for both small and large data sets. On the other hand, both DSLINK and TET-

Table 5

Distances Between the Genetic Markers from the MI Trisomy 21 Data Set

	Estimated Genetic Distances in cM (Standard Error) under χ^2 Models					
Interval	0	1	2	3	4	5
d_1	2.5 (.5)	2.3 (.1)	2.3 (.1)	2.3 (.1)	2.3 (.1)	2.3 (.1)
d_2	3.5 (.7)	3.6 (.1)	3.9 (.1)	4.1 (.1)	4.4 (.1)	4.6 (.1)
d_3	1.4 (.4)	1.7 (.1)	2.1 (.1)	2.4 (.1)	2.8 (.1)	3.1 (.1)
d_4	2.5 (.5)	2.6 (.1)	2.9 (.1)	3.2 (.1)	3.5 (.1)	3.7 (.1)
d_5	1.6 (.4)	1.6 (.1)	1.6 (.1)	1.6 (.1)	1.6 (.1)	1.6 (.1)
d_6	9.4 (.7)	8.4 (.1)	8.6 (.1)	9.0 (.1)	9.3 (.1)	9.7 (.1)
d_7	2.0 (.2)	1.7 (.1)	1.7 (.1)	1.7 (.1)	1.6 (.1)	1.6 (.1)
d_8	1.7 (.0)	1.6 (.1)	1.6 (.1)	1.6 (.1)	1.6 (.1)	1.6 (.1)
d_9	8.9 (.2)	7.9 (.2)	8.1 (.2)	8.3 (.2)	8.6 (.3)	8.9 (.3)
d_{10}	7.4 (.1)	6.9 (.2)	7.0 (.2)	7.1 (.2)	7.2 (.3)	7.3 (.3)
Total	40.9	38.4	39.6	41.3	42.8	44.3
Log−L	−456.5	−459.4	−465.8	−472.8	−479.8	−486.5

RAD showed occasional inconsistencies in estimates derived from data sets of different sizes. Because our HMM approach includes the NDJMap approach as a special case, $m = 0$, we expect the HMM approach to generate results as consistent as those of the NDJMap method, whereas previous methods may fail. In addition, our HMM approach has the further advantage that it can incorporate cross-interference in the analysis, and our simulation results for both MI and MII cases show that our method produces unbiased estimates of genetic distance in the presence of crossover interference. One major limitation of the EM approach is that both computer-memory requirements and computation time increase exponentially with the number of markers, making the approach inapplicable to cases where many genetic markers need to be jointly analyzed.

Recall that the EM approach assumes that each marker interval has at most one chiasma, whereas the HMM approach assumes that the crossover process follows the χ^2 model. When these two approaches were applied to analyze a real MI trisomy data set, we found that they gave similar estimates of genetic distance. However, the two approaches yielded somewhat different patterns of the exchange distributions conditional on a given number of chiasmata on the four-strand bundle. For example, conditional on having two chiasmata in the region, the more-proximal exchange was inferred by use of the EM approach to occur mostly in the first two intervals, whereas, by use of the HMM approach, the more-proximal exchange was inferred to be more uniformly distributed across the first six intervals. Given the small proportion of trisomies estimated to have two chiasmata in the region, it may be difficult to establish a statistically significant result in the comparison between the two approaches, and these discrepancies will be addressed in our future studies. Because the EM approach imposes less-stringent assumptions on the crossover process, especially when the markers are close to each other, discrepancies between the two approaches, such as those just described, may point to some limitation of the χ^2 model and may lead to modification or developments of new models for the crossover process.

The HMM approach can be extended to the Poisson-skip model (Lange et al. 1998), a generalization of the χ^2 model. Although the χ^2 model has been found to be a good fit for data from many organisms, it may not accurately describe the crossover processes underlying meiotic nondisjunction, such as the discrepancies observed between the results under the HMM approach and those under the EM approach. Model checking and model comparison will be performed in our future work, using the observed data to investigate the usefulness of the HMM discussed in the present article.

Throughout the present article, we have assumed that no genotyping error is present in the observed data, although genotyping errors do occur in real genetic studies. In a study published elsewhere, we examined the effects of genotyping errors and interference on estimation of genetic distances, using recombination data from single spores (Goldstein et al. 1997). It was found that genotyping errors inflate genetic distance estimates, and we expect similar results for nondisjunction data, especially for closely linked markers. However, the magnitude of such effects and the ability of statistical methods to incorporate genotyping errors are beyond the scope of the present article and will be addressed in our future studies.

In general, the state of origin of trisomy is evaluated by comparing chromosome 21 pericentromeric markers of the parent who contributed the extra chromosome with those of the trisomic offspring. If parental heterozygosity was retained in the trisomic offspring, we conclude that an MI error is present, and if parental heterozygosity was reduced to homozygosity, we conclude

an MII or mitotic error is present. Because highly polymorphic chromosome 21 pericentromeric markers are not available to us, these determinations are based on the most-proximal informative markers. This means that some proportion of assignments may be in error (Sherman et al. 1994). Moreover, in our real trisomy data set, a large proportion of proximal markers are untyped or uninformative, and this, too, may lead to assignment error. Our experience with human UPD15 data analysis suggests that these assignment errors may reduce our power to detect crossover interference.

Acknowledgments

We thank two anonymous referees for their constructive comments. This work was supported in part by National Institutes of Health grant HD36834 and March of Dimes Birth Defects Foundation research grant FY98-0752.

Appendix A

Below we describe the EM algorithm for the maximum-likelihood estimates of multilocus ordered-tetrad probabilities from trisomy data, assuming at most one chiasma within each marker interval.

E Step

Denote the current estimates of ordered-tetrad probabilities by $p^c_{J_n}$. Our data reconstruction is performed in two steps: (1) we calculate the expected number C_{I_n} of each possible trisomy state I_n, conditional on the observed data and the current estimates p^c_J; and (2) we calculate the expected number D_{J_n} of each possible ordered-tetrad state J_n from C_{I_n}.

Section A.—We calculate multilocus trisomy probabilities u_{I_n} via the relationships established between UPD probabilities and ordered-tetrad probabilities in Zhao et al. (2000) as follows. For each individual from the observed data, if there are no partially informative, no untyped, and no uninformative markers, the sample case corresponds to a particular pattern in $I_n = (i_1, i_2, \ldots, i_n)$, where $i_k \in \{0, 1\}$ and $1 \leqslant k \leqslant n$. Therefore, the contribution of this sample case to pattern I_n is 1, and its contribution to all other patterns is 0. If there are h untyped or uninformative markers (denoted by U) and f partially informative markers (denoted by X) for this sample case, where $h \geqslant 0$, $f \geqslant 0$, and $1 \leqslant h + f \leqslant n$, this case corresponds to 2^{h+f} different string patterns I_n. Denote the positions of untyped and uninformative markers by $m_1, m_2, \ldots, m_h$, and denote the positions of partially informative markers by $z_1, z_2, \ldots, z_f$; the 2^{h+f} different patterns can be represented by $L_n = (l_1, l_2, \ldots l_n)$, where if $k \in \{m_1, m_2, \ldots, m_h\} \cup \{z_1, z_2, \ldots, z_f\}$, then l_k can take value 0 or 1, and if $k \notin \{m_1, m_2, \ldots, m_h\} \cup \{z_1, z_2, \ldots, z_f\}$, then l_k takes a fixed value $j_k \in \{0, 1\}$. Note that for a $k \in \{z_1, z_2, \ldots, z_f\}$, the contribution of symbol X to N is two times as great as that to R at this marker. This is because $P(X) = P(X/N)P(N) + P(X/R)P(R) = \alpha P(N) + 0.5\alpha P(R)$, where α is a constant. For each L_n, we use g_{L_n} to denote the number of those k with $k \in \{z_1, z_2, \ldots, z_f\}$ and $l_k = 0$. Therefore, the contribution of this sample case to each of the 2^{h+f} different patterns can be calculated by

$$(\tfrac{1}{2})^{g_{L_n}} u_{L_n} \Big/ \sum_{l_{m_1}, l_{m_2}, \ldots, l_{m_h}, l_{z_1}, l_{z_2}, \ldots, l_{z_f}} (\tfrac{1}{2})^{g_{L_n}} u_{L_n} ,$$

and the contribution of this sample case to all other patterns I_n besides these 2^{h+f} patterns is 0. If we go through the whole data set in this fashion, we can obtain the expected number C_{I_n} for each possible pattern UPD I_n.

Section B.—With the relationships between u_{I_n} and p_{J_n} established in our earlier study (Zhao et al. 2000), we calculate D_{J_n} from the C_{I_n} (obtained in Section A), as follows:

$$D_{J_n} = \sum_{I_n} C_{I_n} [c(I_n, J_n) p^c_{J_n} \Big/ \sum_{J_n} c(I_n, J_n) p^c_{J_n}] .$$

M Step

The updated estimates of multilocus ordered-tetrad probabilities p_{J_n} are

$$p^{\text{new}}_{J_n} = \frac{D_{J_n}}{S} ,$$

where S is the sample size.

Repeat the E step and the M step until convergence.

Appendix B

The Five Components of the HMM for Trisomy Data

Because we relate trisomy data to ordered-tetrad data (as in the analysis of UPD data), three of the five components are identical to the UPD case. These three identical components are: S, the set of hidden states; A, the state transition probability distribution; and π, the initial state distribution. For the general $Cx(Co)^m$ model, Zhao et al. (2001) defined $\mathcal{N} = 6(m + 1)$ hidden states for each marker A_r. Each hidden state is represented by $S_{i,l}$, where $i = 1, \ldots, 6$ denotes one of the six patterns for an ordered-tetrad (these six patterns were defined earlier, in the subsection Notation for Multilocus Ordered-Tetrad Data), and l denotes the number of Co events after the last Cx event before marker A_r. Let $p = m + 1$, the elements in the $6p \times 6p$ transition matrix between the hidden states at consecutive markers given elsewhere (Zhao et al. 2001). The initial hidden state can only be one of the $2(m + 1)$ states $S_{1,l}$ and $S_{6,l}$, where $l = 0, \ldots, m$. If we assume that the crossover process is stationary, these $2m + 2$ states have the same probability of being the initial state. Therefore, we discuss only the two components that are different from UPD data: V, the set of distinct observation symbols; and B_j, the observation symbol probability distribution in state S_j.

As discussed above, there are four possibilities at each marker for trisomy data; therefore, $V = \{R, N, X, U\}$. For the observation symbol probability distribution in state $S_{i,l}$, it depends on which two strands are observed in the nondisjoined chromosome of an individual with trisomy. Because meiotic nondisjunction events can be classified as either MI nondisjunction or MII nondisjunction, we consider MI and MII nondisjunction, in turn, in our discussion.

If we use $[E, F; H, W]$ to denote an ordered tetrad, we define E as the first strand and H as the third strand. Under MI nondisjunction, without loss of generality, we assume that the first and third strands are observed in the nondisjoined chromosomes of an individual with trisomy. We consider two cases separately: (1) both parents are available or (2) only the NDJP is available.

MI Nondisjunction, Both Parents Available

We first note that the observation symbol distribution depends only on the first component in the hidden state $S_{i,l}$. Given a completely informative marker, according to Zhao et al. (2001), we have the following observation symbol probabilities for the first four mating types defined in the Notation for Multilocus Trisomy Data section:

	$S_{1,l}$	$S_{2,l}$	$S_{3,l}$	$S_{4,l}$	$S_{5,l}$	$S_{6,l}$
R	0	1	0	0	1	0
N	1	0	1	1	0	1

.

If the parental mating type is an intercross at the marker, we may observe only an X or an R for the trisomy individual. When $i = 1, 3, 4, 6$, the two strands from the NDJP are nonreduced, so the probability that we observe an X at this marker is 1. When $i = 2, 5$, the two strands from the NDJP are reduced, so we have the same probability of observing an X or an R at this marker. We have the following observation symbol probabilities:

	$S_{1,l}$	$S_{2,l}$	$S_{3,l}$	$S_{4,l}$	$S_{5,l}$	$S_{6,l}$
R	0	0.5	0	0	0.5	0
N	0	0	0	0	0	0
X	1	0.5	1	1	0.5	1

.

When the marker is uninformative or untyped, it is easy to see that $P(U|S_{i,l}) = 1$.

MI Nondisjunction When Only the NDJP Is Available

If the NDJP is heterozygous at a marker, assume that the genotype of the NDJP at this marker is $A_G B_G$. Let p_{A_G} and p_{B_G} denote the allele frequency of A_G and B_G in the population. Therefore, $q_G = 1 - (p_{A_G} + p_{B_G})$ is the probability that the CDJP contributes an allele that is different from either of the alleles in the NDJP. For $i = 1, 3, 4, 6$, the two strands from the NDJP are nonreduced. In this case, we observe an N if and only if the CDJP

contributes an allele different from those in the NDJP. This occurs with probability q_G. If the CDJP contributes an allele that is the same as one of the two alleles in the NDJP, we observe a Z, and this occurs with probability $1 - q_G$. For $i = 2, 5$, the two strands from the NDJP are reduced. In this case, there are two possibilities that we may observe an R: (1) the CDJP contributes an allele different from those of the NDJP, and this happens with probability q_G; (2) the CDJP contributes an allele that is the same as the one contributed by the NDJP, and this happens with probability $(1 - q_G)/2$. On the other hand, we observe an X if and only if the CDJP contributes an allele which is carried but not contributed by the NDJP. This happens with probability $(1 - q_G)/2$. In addition, the chance that we observe an N is 0. From the above discussion, we obtain the following observation symbol probability distribution:

$$\begin{array}{cccccccc} & S_{1,l} & & S_{2,l} & S_{3,l} & S_{4,l} & & S_{5,l} & S_{6,l} \\ R & 0 & & \frac{1}{2}(1+q_G) & 0 & 0 & & \frac{1}{2}(1+q_G) & 0 \\ N & q_G & & 0 & q_G & q_G & & 0 & q_G \\ X & (1-q_G) & \frac{1}{2}(1-q_G) & & (1-q_G) & (1-q_G) & \frac{1}{2}(1-q_G) & & (1-q_G) \end{array}.$$

If the marker in the NDJP is either homozygous or untyped, $P(U|S_{i,l}) = 1$.

MII Nondisjunction, Both Parents Available

For MII nondisjunction events, the two chromosomes inherited from the NDJP are sister chromatids. We can obtain the observation symbol probabilities as in the MI nondisjunction case. For a completely informative marker, we have the following observation symbol probabilities for the first four mating types defined in the Notation for Multilocus Trisomy Data section:

$$\begin{array}{cccccccc} & S_{1,l} & S_{2,l} & S_{3,l} & S_{4,l} & S_{5,l} & S_{6,l} \\ R & 1 & 0 & 0 & 0 & 0 & 1 \\ N & 0 & 1 & 1 & 1 & 1 & 0 \end{array}.$$

For a marker at which the parental mating type is an intercross, we have the following observation symbol probabilities:

$$\begin{array}{cccccccc} & S_{1,l} & S_{2,l} & S_{3,l} & S_{4,l} & S_{5,l} & S_{6,l} \\ R & 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ N & 0 & 0 & 0 & 0 & 0 & 0 \\ X & 0.5 & 1 & 1 & 1 & 1 & 0.5 \end{array}.$$

When the marker is uninformative or untyped, $P(U|S_{i,l}) = 1$.

MII Nondisjunction When Only the NDJP Is Available

If the NDJP is heterozygous at a marker, we have the following observation symbol probability distribution

$$\begin{array}{cccccccc} & S_{1,l} & & S_{2,l} & S_{3,l} & S_{4,l} & & S_{5,l} & S_{6,l} \\ R & \frac{1}{2}(1+q_G) & & 0 & 0 & 0 & & 0 & \frac{1}{2}(1+q_G) \\ N & 0 & & q_G & q_G & q_G & & q_G & 0 \\ X & \frac{1}{2}(1-q_G) & 1-q_G & & 1-q_G & 1-q_G & 1-q_G & & \frac{1}{2}(1-q_G) \end{array}.$$

If the marker in the NDJP is either homozygous or untyped, $P(U|S_{i,l}) = 1$.

Electronic-Database Information

The URLs for data in this article are as follows:

Authors' Web site, http://bioinformatics.med.yale.edu
Genetic Location Database, http://cedar.genetics.soton.ac.uk/public_html/ldb.html

References

Bailey NTJ (1961) An introduction to the mathematical theory of genetic linkage. Oxford University Press, London

Broman KW, Weber JL (2000) Characterization of human crossover interference. Am J Hum Genet 66:1911–1926

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

Chakravarti A, Majumder PP, Slaugenhaupt SA, Deka R, Warren AC, Surti U, Ferrell RE, Antonarakis SE (1989) Gene-centromere mapping and the study of nondisjunction in autosomal trisomies and ovarian teratomas. In: Molecular and cytogenetic studies of nondisjunction. Alan R. Liss, New York, pp 35–42

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–37

Feingold E, Brown AS and Sherman SL (2000) Multipoint estimation of genetic maps for human trisomies with one parent or other partial data. Am J Hum Genet 66:958–968

Foss E, Lande R, Stahl FW, Steinberg CM (1993) Chiasma interference as a function of genetic distances. Genetics 133: 681–691

Goldstein DR, Zhao H, Speed TP (1997) The effects of genotyping errors and interference on estimation of genetic distance. Hum Hered 47:86–100

Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM (1996) An introduction to genetic analysis, 6th ed. W. H. Freeman, New York

Hassold TJ, Jacobs PA (1984) Trisomy in man. Ann Rev Genet 18:69–97

Hulten MA (1974) Chiasma distribution at diakinesis in the normal human male. Hereditas 76:55–78

Lamb NE, Freeman SB, Savage-Austin A, Pettay D, Taft L, Hersey J, Gu Y, Shen J, Saker D, May KM, Avramopoulos D, Petersen MB, Hallberg A, Mikkelsen M, Hassold TJ, Sherman L (1996) Susceptible chiasmate configuration of chromosome 21 predisposes to non-disjunction in both maternal meiosis I and meiosis II. Nat Genet 14:400–405

Lamb NE, Feingold E, Savage A, Avramopoulos D, Freeman S, Gu Y, Hallberg A, Hersey J, Karadima G, Pettay D, Saker D, Shen J, Taft L, Mikkelsen M, Petersen MB, Hassold T, Sherman SL (1997) Characterization of susceptible chiasma configurations increase the risk for maternal nondisjunction of chromosome 21. Hum Mol Genet 9:1391–1399

Lange K, Zhao H, Speed TP (1997) The Poisson-skip model of crossing-over. Ann Appl Prob 7:299–313

Orr-Weaver T (1996) Meiotic nondisjunction does the two step. Nat Genet 14:374–376

Ott J, Linder D, McCaw BK, Lovrien EW, Hecht F (1976) Estimating distances from centromere by means of benign ovarian teratomas in man. Ann Hum Genet 40:191–196

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77: 257–286

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

Sherman SL, Petersen MB, Freeman SB, Hersey J, Pettay D, Taft L, Frantzen M, Mikkelsen M, Hassold TJ (1994) Nondisjunction of chromosome 21 in maternal meiosis I: evidence for a maternal age-dependent mechanism involving reduced recombination. Hum Mol Genet 3:1529–1535

Yu K, Feingold E (2001) Estimating the frequency distribution of crossovers during meiosis from recombination data. Biometrics 57:427–434

Zhao H, Li J, Robinson WP (2000) Multipoint genetic mapping with uniparental disomy data. Am J Hum Genet 67: 851–861

——— Statistical analysis of uniparental disomy data using hidden Markov models. Biometrics (in press)

Zhao H, Speed TP, Mcpeek MS (1995) Statistical analysis of crossover interference using the chi-square model. Genetics 139:1045–1056