

Statistical Analysis of Uniparental Disomy Data Using Hidden Markov Models

H. Zhao^{1*}, J. Li¹ and W. P. Robinson²

¹ Department of Epidemiology and Public Health, Yale University School of
Medicine, New Haven, CT 06520, U.S.A.

² Department of Medical Genetics, University of British Columbia,
Vancouver, B.C., Canada V5Z 4H4

SUMMARY. Genetic studies of uniparental disomy employing many markers have helped geneticists to gain a better understanding of the molecular mechanisms underlying nondisjunction. However, most existing methods cannot simultaneously analyze all genetic markers and consistently incorporate crossover interference, thus fail to make the most use of genetic information in the data. In the present article, we describe a hidden Markov model for multilocus uniparental disomy data. This method is based on the chi-square model for the crossover process and can simultaneously incorporate all marker information including untyped and uninformative markers. We then apply this novel method to analyze a set of UPD 15 data.

1. Introduction

Uniparental disomy (UPD) is the situation where chromosome number is normal but both copies (homologs) of a chromosome pair have originated from

* *Email:* hongyu.zhao@yale.edu

a single parent instead of from both parents. Both Prader-Willi syndrome and Angelman syndrome are often associated with UPD15. Although recent studies have revealed that the recombination patterns among meioses leading to nondisjunction may be different from those among normal meioses, information in the collected data has not been fully utilized by the existing methods as reviewed in the following. Therefore, the objective of this article is to develop a general statistical approach that overcomes the limitations of the existing methods in the analysis of UPD data.

UPD generally arises as a consequence of nondisjunction events (chromosome missegregation) during meiosis, the process that produces gametes having one-half the genetic material of the “parent” cell. At the start of meiosis, two chromosome sets are present with each chromosome having a pairing partner, and such pairs are called *homologous* pairs. During meiosis, homologous chromosomes pair and each of the paired chromosomes duplicates, resulting in a synapsed structure of a bundle of four homologous *chromatids*. Chromatids which are copies of the same chromosome are called *sister chromatids*, while those originating from homologous chromosomes are called *nonsister chromatids*. We call these four homologous chromatids a *four strand bundle*. Crossovers then take place after the formation of the four-strand bundle, with each crossover involving two nonsister chromatids. The number and locations of crossovers vary from chromosome to chromosome for the same meiosis, and from meiosis to meiosis for the same chromosome. For detailed background on the genetic concepts discussed in this article, see (Hawley and Mori, 1999). In meiosis, there are two rounds of chromosome segregation. In meiosis I (MI) the homologous pairs segregate, while the

sister chromatids segregate in meiosis II (MII). After a normal meiosis, each gamete carries one of the four strands from the original four-strand bundle.

A *nondisjunction* event results in segregation of two of the four strands into one gamete. This can result either from the failure of the chromosomes to separate or early separation of the chromosomes that allows random movement of the chromatids to the poles. Different meiotic nondisjunction events are classified as MI nondisjunction if the two chromosome copies within the gamete are homologous, and as MII nondisjunction if the two copies are sister chromatids. In humans, missegregation of chromosomes occurs at a high frequency. Aberrant segregation may occur in more than 10% of female meioses, and increases dramatically in mothers over 35 years old (Orr-Weaver, 1996). It has been found that the nondisjunction event leading to UPD15 is predominantly due to a maternal MI segregation error and there is a maternal age effect (Robinson, Kuchinka, Bernasconi and et al., 1998). In this article, we will apply our method to analyze a subset of the 115 UPDs analyzed in Robinson et al. (1998) consisting of 69 cases due to MI errors. A total of 72 genetic markers were used in this study, although they were not typed on all individuals. For a given marker, if the parent undergoing nondisjunction is heterozygous, i.e. the two homologous chromosomes carry different alleles, we can distinguish two types for the UPD individual at this marker: reduced (homozygous, denoted by R) and nonreduced (heterozygous, denoted by N) genotypes for the two nondisjoined chromosomes. When the parent is homozygous for a given marker or is untyped for this marker, we use U to denote the genotype of the UPD individual. Therefore, each UPD individual can be represented as a character string using R , N , and U , such as

“...NRNUN...”. One objective of UPD studies is to derive estimates of the crossover patterns that are associated with nondisjunction. In particular, geneticists are interested in examining whether genetic distances, defined as the average number of crossovers between two markers on a single strand per meiosis, are altered in such meioses. The unit of genetic distance is Morgan, and two markers are one Morgan apart from each other if, on average, there is one crossover between them per meiosis. Another commonly used unit for genetic distance is centiMorgan, which is abbreviated as cM.

In the analysis of UPD data with many markers, two issues must be appropriately addressed: the use of joint information from multiple markers and a genetics phenomenon called crossover interference. Crossover interference refers to the non-independent occurrence of crossovers on a chromosome, and it has been observed in most organisms, including humans (Broman and Weber, 2000). Although various statistical methods for nondisjunction data have been discussed in the literature, the treatments of these two issues are not entirely satisfactory. First, the method developed by Chakravarti and Slaugenhaupt (1987) can only handle one marker at a time. Shahar and Morton (1986) established relationships between trisomy (another type of nondisjunction data) probabilities and genetic distances, and then derived joint likelihood for multilocus data. However, several assumptions on the crossover process during meiosis were made in their approach and these assumptions are not consistent with each other. Chakravarti, Majumder, Slaugenhaupt and et al. (1989) proposed two multilocus analysis approaches: one was to assume at most three chiasmata across the region under study with at most one chiasma in each marker interval, and the other was to treat

the proximal marker as a pseudocentromere relative to the distal marker. It is apparent that the first approach is not applicable to chromosomes likely to have more than three chiasmata or to studies involving large marker intervals. Although crossover interference was considered for each pair of consecutive markers in the second approach, it implicitly assumes the absence of chiasma interference when treating the proximal marker as a pseudocentromere. Feingold, Brown and Sherman (2000) derived multipoint likelihoods for trisomy data under the assumption of no crossover interference. However, the genetic distance estimates from their approach may be biased because crossover interference does seem to exist during normal human meiosis. We recently derived general relationships between multilocus UPD probabilities and multilocus ordered tetrad probabilities (Zhao, Li and Robinson, 2000). For an arbitrary model for the crossover process, this relationship allows the calculation of the likelihood for the observed UPD data by using all marker information simultaneously, including untyped and uninformative markers. One limitation of this approach is that the amount of computation increases exponentially with the number of markers. Because many markers (20 or more) are often used in nondisjunction studies, alternative approaches are needed to evaluate the likelihood if all markers are considered simultaneously. To develop an alternative approach, some compromise needs to be made between the generalizability of the method and the flexibility of the stochastic model for the crossover process. In this article, assuming that the crossover process follows the chi-square model (Zhao, Speed and Mcpeek, 1995), we describe how to evaluate the likelihood of the observed UPD data through a hidden Markov model (HMM). Using this approach, the amount

of computation increases linearly with the number of markers under study. Therefore, likelihood calculation becomes feasible for UPD data involving many genetic markers and we can incorporate crossover interference in a consistent manner.

2. HMM for UPD data when crossover process is modeled by the chi-square model

The chi-square model for crossovers has a long history (Bailey, 1961). Foss, Lande, Stahl and Steinberg (1993) represented the model in the form of $Cx(Co)^m$ as follows: assume the crossover intermediates (C events) are randomly distributed along the four strand bundle, and every intermediate resolves either as a crossover (Cx) or not (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 model (no interference model) corresponds to $m = 0$. Zhao et al. (1995) found that the chi-square model provides good fit to data from many organisms.

2.1 Hidden Markov model

As discussed in Rabiner (1989), an HMM has the following five components: (1) the set of hidden states: $S = \{S_1, \dots, S_i, \dots, S_L\}$, $1 \leq i \leq L$; (2) the set of distinct observation symbols: $V = \{v_1, \dots, v_k, \dots, v_M\}$, $1 \leq k \leq M$; (3) the state transition probability distribution: $A = \{a_{ij}\}$, where $a_{ij} = P(q_{r+1} = S_j | q_r = S_i)$, $1 \leq i, j \leq L$ and 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 $b_j(v_k) = P(O_r = v_k | q_r = S_j)$, $1 \leq j \leq L, 1 \leq k \leq M$ and O_r de-

notes the observation symbol at time r ; and (5) the initial state distribution: $\pi = \{\pi_i\}$ where $\pi_i = P(q_1 = S_i)$. For given S, V, A, B , and π , the probability of an observation sequence $O = O_1O_2 \cdots O_n$ can be evaluated using either the forward or the backward algorithm (Rabiner (1989)), where n denotes the length of the sequence.

Therefore, if we can identify the S, V, A, B , and π for the HMM that describes the chi-square model for the UPD data, multilocus UPD probabilities can be easily calculated using the forward algorithm. Because meiotic nondisjunction events can be classified as either MI nondisjunction or MII nondisjunction, we first consider MI nondisjunction in our discussion. We discuss the Poisson model and the general chi-square model $Cx(Co)^m$ in turn.

2.2 MI nondisjunction and the no interference model (Poisson model)

Consider n markers in the order of $CEN - \mathcal{A}_1 - \dots - \mathcal{A}_n$ with genetic distance between \mathcal{A}_{r-1} and \mathcal{A}_r being d_r , where $r = 1, 2, \dots, n$ and \mathcal{A}_0 is the centromere CEN . Centromere is the constricted region of a nuclear chromosome, to which the spindle fibers attach during division. For a heterozygous marker \mathcal{A}_r with alleles A_r and a_r in the parent, there are six distinguishable patterns: $S_1 = [A_r, A_r, a_r, a_r]$, $S_2 = [A_r, a_r, A_r, a_r]$, $S_3 = [A_r, a_r, a_r, A_r]$, $S_4 = [a_r, A_r, A_r, a_r]$, $S_5 = [a_r, A_r, a_r, A_r]$ and $S_6 = [a_r, a_r, A_r, A_r]$ for an ordered tetrad. A tetrad is the four haploid product cells from a single meiosis. We define six hidden states corresponding to these six patterns for the HMM. For UPD data, at each heterozygous marker \mathcal{A}_r , there are three observed genotypes A_rA_r , A_ra_r , and a_ra_r for the offspring. As mentioned in the Introduction, we distinguish two types at a heterozygous marker: reduced (R)

and nonreduced (N). Therefore, we need two observation symbols $v_1 = R$ and $v_2 = N$ for a heterozygous marker. If the parent carries two identical alleles or this marker is not typed, this marker is called uninformative or untyped. We denote this situation using the third observation symbol $v_3 = U$. Therefore, $M = 3$ and $v = \{R, N, U\}$. Now consider the four strands during meiosis. For two markers \mathcal{A}_{r-1} and \mathcal{A}_r , we distinguish three tetrad types: parental ditype where all four strands retain the parental type, tetratype where two of the four strands show recombination, and nonparental ditype where all four strands are recombinants. The probabilities for these three types are denoted by p_0 , p_1 , and p_2 . When there is no interference, these three probabilities are: $p_0 = (1 + 2e^{-3d_r} + 3e^{-2d_r})/6$, $p_1 = 2(1 - e^{-3d_r})/3$, and $p_2 = (1 + 2e^{-3d_r} - 3e^{-2d_r})/6$, where d_r is the genetic distance between \mathcal{A}_{r-1} and \mathcal{A}_r (Haldane, 1931). The transition probability distribution between two consecutive markers is given by

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	p_0	$p_1/4$	$p_1/4$	$p_1/4$	$p_1/4$	p_2
S_2	$p_1/4$	p_0	$p_1/4$	$p_1/4$	p_2	$p_1/4$
S_3	$p_1/4$	$p_1/4$	p_0	p_2	$p_1/4$	$p_1/4$
S_4	$p_1/4$	$p_1/4$	p_2	p_0	$p_1/4$	$p_1/4$
S_5	$p_1/4$	p_2	$p_1/4$	$p_1/4$	p_0	$p_1/4$
S_6	p_2	$p_1/4$	$p_1/4$	$p_1/4$	$p_1/4$	p_0

The observation symbol probability distribution in state j depends on which two strands are observed in a UPD individual. Random spindle-centromere attachment assumption (RSCA) assumes that two centromeres have the same chance to go to either pole at the first meiotic division, and the divided centromeres have the same chance to go either pole at the second meiotic division. Under RSCA, we can, without loss of generality, assume

that the first and third strands are observed in the UPD individual with MI error. To define the observation symbol probabilities, we distinguish two cases: the parent being heterozygous or being homozygous/untyped. When the parent is heterozygous, it is impossible to observe U and the observation symbol probability distribution is

$$\begin{array}{cccccc} & S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \\ R & 0 & 1 & 0 & 0 & 1 & 0 \\ N & 1 & 0 & 1 & 1 & 0 & 1 \end{array} .$$

If the parent is homozygous or untyped, the probability of observing R or N is 0 and the probability of observing U is 1. Finally, we need to define the initial state distribution $\pi = \{\pi_i\}$ to complete our HMM specification. Because the first marker is the centromere, the initial distribution is $\pi = \{1/2, 0, 0, 0, 0, 1/2\}$.

2.3 MI nondisjunction and the $Cx(Co)^m$ model

For the $Cx(Co)^m$ model, if we define the hidden state at each marker according to the six patterns listed above, these hidden states no longer form a Markov chain. This is because that the crossover events in adjacent intervals are not independent. Recall that the crossover intermediates (C events) in the chi-square model are independently distributed along the chromosome and they resolve as crossovers in a deterministic fashion. Starting from the centromere, suppose we keep track of how many Co events (C events not resolved as crossovers) have occurred since the last Cx event (crossover) before a genetic marker, and denote the number of such Co events by l . For the $Cx(Co)^m$ model, l can be $0, 1, \dots, m$. Denote the hidden state at a marker by $S_{i,l}$, where $i = 1, \dots, 6$ denotes one of the six patterns and l was just defined, then the $S_{i,l}$ form a Markov chain. In this case there are $L = 6(m+1)$ distinct

hidden states for each marker \mathcal{A}_r . Each hidden state is represented by $S_{i,l}$, where $i = 1, \dots, 6$ denotes one of the six patterns, and l denotes the number of Co events after the last Cx event before marker \mathcal{A}_r . Let $p = m + 1$, and the elements in the $6p \times 6p$ transition matrix $P(q_{r+1} = S_{i_{r+1}, l_{r+1}} | q_r = S_{i_r, l_r})$ can be defined as follows.

First suppose the tetrad type between the two markers is parental ditype. If $l_r \leq l_{r+1}$, the number of C events between the two markers is $kp + (l_{r+1} - l_r)$, where $k \geq 0$. Therefore, $P(q_{r+1} = S_{i_{r+1}, l_{r+1}} | q_r = S_{i_r, l_r}) = \sum_{k=0}^{\infty} P\{C = kp + (l_{r+1} - l_r)\} p_0^k$, where $P\{C = kp + (l_{r+1} - l_r)\} = e^{-y_{r+1}} y_{r+1}^{kp + (l_{r+1} - l_r)} / (kp + (l_{r+1} - l_r)!)^k$, y_{r+1} is the average number of C events between \mathcal{A}_r and \mathcal{A}_{r+1} , and p_0^k is the conditional probability of having parental ditype given there being k chiasmata between two markers. The genetic distance $d_{r+1} = y_{r+1}/2p$. If $l_r > l_{r+1}$, the number of C events between the two markers is $kp + (l_{r+1} - l_r)$, where $k \geq 1$. Therefore, $P(q_{r+1} = S_{i_{r+1}, l_{r+1}} | q_r = S_{i_r, l_r}) = \sum_{k=1}^{\infty} P\{C = kp + (l_{r+1} - l_r)\} p_0^k$. If the tetrad type between the two markers is nonparental ditype, we can simply substitute p_0^k in the above expressions by p_2^k , where p_2^k is the conditional probability of having nonparental ditype given there being k chiasmata between two markers. If the tetrad type between the two markers is tetratype, there are four patterns with equal chance at \mathcal{A}_{r+1} . Therefore, we substitute p_0^k or p_2^k in the above expressions by $p_1^k/4$, where p_1^k is the conditional probability of having tetratype given there being k chiasmata between two markers. To calculate p_0^k , p_1^k , and p_2^k , we note that given k crossovers between two markers, then the probability that the two markers show parental ditype is $p_0^k = \{1/2 + (-1/2)^k\}/3$, and the probability that the two markers show nonparental ditype is $p_2^k = \{1/2 + (-1/2)^k\}/3$ for $k \geq 1$

and $p_2^0 = 0$ (Mather, 1935).

The observation symbol distribution only depends on the first subscript in the hidden state, and it is in the same form as that in Poisson model. The initial hidden state can only be one of the $2(m + 1)$ states $S_{1,l}$ and $S_{6,l}$, where $l = 0, \dots, m$. If we assume that the crossover process is stationary, these $2m + 2$ states have the same probability to be the initial state. These model specifications allow us to evaluate any UPD data probability using the forward algorithm.

2.4 *MII nondisjunction and the $Cx(Co)^m$ model*

Because the hidden states are defined on the four-strands before meiotic divisions, the number of hidden states, the transition probability matrix, and the initial state distribution are the same as those in our discussion of MI nondisjunction. The only difference is that either pair of sister chromatids are observed in a UPD individual with equal probability. The observation symbol distribution for MII disjunction can be easily derived similar to the MI nondisjunction case.

Having defined the five components of an HMM that corresponds to the chi-square model, it is straightforward to evaluate the probability of any UPD individual through standard HMM approaches. Using this framework, genetic distances among genetic markers can be estimated via the maximum likelihood method.

3. **Simulation results**

Consider a set of markers in the order of $CEN - \mathcal{A}_1 - \dots - \mathcal{A}_n$ with genetic distances d_1, \dots, d_n among them. In our simulations, we varied the sample size ($S = 50, 100, \text{ and } 400$), the interference parameter m in the chi-square

model $Cx(Co)^m$ ($m = 0, 1, 2,$ and 3), and the proportion of missing (untyped or uninformative) data ($\mu = 0\%, 25\%,$ and 50%). We also varied the number of markers and the distances among the markers. For each parameter combination, we generated 100 simulated data sets and estimated the distances among the markers and the standard errors for these distance estimates under the correct model. For the simplicity of our discussion, we only summarize our simulation results for the genetic distance estimate in the first interval (d_1) for five equally spaced markers with 20cM between each pair of consecutive markers. The averages of these estimates and the averages of the associated standard errors from these 100 simulated data sets are listed in Table 1. In general, the maximum likelihood estimates were almost unbiased under the simulation models studied, although there seems to be some bias for small sample sizes. The associated standard errors increased with smaller sample sizes, lower interference values, and larger proportion of missing values.

[Table 1 about here.]

We further studied the estimation of the correct order m in the chi-square model using UPD data. We generated 100 data sets of sample size 100 for each model ($m = 0, 1, 2,$ or 3) and fitted them to six chi-square models ($m = 0, 1, \dots, 5$). For the 100 simulated data sets under each model, we calculated the average of the maximized log-likelihoods under each of the six chi-square models. The results are summarized in the upper portion of Table 2. Although the correct model had the largest average likelihood for each model simulated, the differences among the fitted models were small.

For each simulated data set, we estimated the order m in the chi-square model by choosing the model $Cx(Co)^m$ with the largest likelihood. In the lower portion of Table 2, we summarize the number of times that each order m was chosen. Although the correct model was chosen most frequently, the proportion of model order misspecification increased with the true model order m . For example, if the data were generated under the $Cx(Co)^3$ model, the correct model was identified only 25% of the times. Therefore, with a sample size of 100 UPD individuals, there was a high probability that the correct model may not be identified.

[Table 2 about here.]

4. Application to UPD15 data

In this section, we apply our method to analyze a UPD15 data set consisting of 69 cases due to MI errors. The data set analyzed here is a subset of the 115 UPDs analyzed in Robinson et al. (1998), because of the exclusion of both of trisomy cases and of some UPD cases that were typed entirely outside the Robinson lab and that therefore included a largely different set of markers. As markers for the centromere of chromosome 15 are not available, we use markers $D15S541$, $D15S542$, and $D15S543$ to infer the meiotic stage of origin. These markers are the most proximal markers to the centromere. As there is no known crossing over between them, as in Robinson et al. (1998), we treated them as one marker and were able to determine meiotic stage of origin for each case.

4.1 Genetic distance estimates

We analyzed ten markers on chromosome 15 in the order of $CEN - GABRB3 - D15S24 - ACTC - CYP19 - D15S98 - D15S108 - D15S131 -$

$D15S114 - D15S100 - D15S87$. For different interference parameters in the chi-square model, we estimated the genetic distances among these ten markers and the centromere using the maximum likelihood method. For this ten marker data set, there were many cases having untyped and uninformative markers. The proportion of such “missing” markers was 44%. For MI nondisjunction, the estimated genetic distances among these markers and the associated standard errors are shown in Table 3. The total estimated genetic length across these ten markers ranged from 189.4cM to 128.6cM for different m values. As the value of m increases, the estimated genetic distance decreases. The maximized log-likelihoods from different chi-square models were very similar, with the no interference model Cx having the smallest log-likelihood (-158.1) and the $Cx(Co)^6$ model having the largest log-likelihood (-156.1). The results for models $Cx(Co)^7$ and higher are not shown here but they had smaller likelihood values compared to the $Cx(Co)^6$ model. If we construct a confidence set for m through the profile likelihood, this set does not contain $m = 0$. Therefore, there is some indication of crossover interference using this ten-marker data set.

[Table 3 about here.]

4.2 *Comparison with known genetic maps for chromosome 15*

For MI nondisjunction error, our estimated total length between the centromere and the most distant marker $D15S87$ ranged from 128.6cM to 189.4cM. For the no interference model, our estimate is greater than the female map distance 148.0cM estimated from the Center-d’Étude-du-Polymorphisme-Humain (CEPH) pedigrees. For the chi-square model with interference, $m = 1$ or larger, our estimate is smaller than 148.0cM. Previously the CEPH

female chromosome 15 genetic map from D15S11 to D15S87 was estimated as 181cM, whereas the current CEPH/Genethon estimate of chromosome 15 map length is 148cM. Other genetic maps, all based on the same CEPH data and published within the last five years, estimated the female length anywhere between 135cM and 216cM (Robinson et al. (1998)). In our previous study using the same data set, we estimated the total distance to be 101cM under the assumption that there is at most one chiasma in each marker interval (Zhao et al. (2000)). The difference in genetic distance estimates is the result of different assumptions underlying each approach. There is some possibility that double recombinations might occur in some large intervals, and the assumption of at most one exchange in each interval may lead to underestimating the genetic distance.

5. Discussion

Because the chi-square model has been found to fit data well for a number of organisms (Zhao et al. (1995)), we model the crossover process using this class of models in this article. One nice property of this class of models is that the model can be described as an HMM along the chromosome, allowing us to use the efficient numerical algorithm for HMM to evaluate the probability of multilocus UPD data. This modeling approach can utilize all marker information in the data, and also allows the incorporation of crossover interference in a consistent manner. Therefore, this approach represents a consistent and computationally feasible treatment of multilocus UPD data. We have implemented the HMM in a computer program and tested it extensively using simulations. Our simulations showed that genetic distances can be estimated well even when the sample size is as small as 50 and the proportion of un-

typed and uninformative markers is moderate (less than 50%). The computer program can be downloaded at <http://bioinformatics.med.yale.edu>.

The HMM approach can be extended to the Poisson-skip model (Lange, Zhao and Speed (1997)), a generalization of the chi-square model. Although the chi-square model has been found to fit data well from many organisms, it may not accurately describe the crossover process underlying meiotic nondisjunction. 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 this article.

ACKNOWLEDGEMENTS

We thank the editor, the associate editor, and the referee for their constructive comments. This work was supported in part by grant HD36834 from the National Institutes of Health and Research Grant FY98-0752 from the March of Dimes Birth Defects Foundation.

REFERENCES

- Bailey, N. (1961). *An Intruduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.
- Broman, K. and Weber, J. (2000). Characterization of human crossover interference. *American Journal of Human Genetics* **66**, 1911–1926.
- Chakravarti, A., Majumder, P., Slaugenhaupt, S. and et al. (1989). Gene-centromere mapping and the study of nondisjunction in autosomal tri-

- somies and ovarian teratomas. In *Molecular and Cytogenetic Studies of Nondisjunction*, pages 45–79. Alan R. Liss Inc.
- Chakravarti, A. and Slaugenhaupt, S. (1987). Methods for studying recombination on chromosomes that undergo nondisjunction. *Genomics* **1**, 35–42.
- Feingold, E., Brown, A. and Sherman, S. (2000). Multipoint estimation of genetic maps for human trisomies with one parent or other partial data. *American Journal of Human Genetics* **66**, 958–968.
- Foss, E., Lande, R., Stahl, F. and Steinberg, C. (1993). Chiasma interference as a function of genetic distances. *Genetics* **133**, 681–691.
- Haldane, J. (1931). The cytological basis of genetical interference. *Cytologia* **3**, 54–65.
- Hawley, R. and Mori, C. (1999). *The Human Genome: a User's Guide*,. Academic Press, San Diego.
- Lange, K., Zhao, H. and Speed, T. P. (1997). The poisson-skip model of crossing-over. *The Annals of Applied Probability* **7**, 299–313.
- Mather, K. (1935). Reduction and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics* **30**, 53–78.
- Orr-Weaver, T. (1996). Meiotic nondisjunction does the two step. *Nature Genetics* **14**, 374–376.
- Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.
- Robinson, W., Kuchinka, B., Bernasconi, F. and et al. (1998). Maternal meiosis i non-disjunction of chromosome 15: dependence of the maternal age effect on level of recombination. *Human Molecular Genetics* **7**, 1011–1019.

- Shahar, S. and Morton, N. (1986). Origin of tetratomas and twins. *Human Genetics* **74**, 215–218.
- Zhao, H., Li, J. and Robinson, W. P. (2000). Multipoint genetic mapping with uniparental disomy data. *American Journal of Human Genetics* **67**, 851–861.
- Zhao, H., Speed, T. P. and Mcpeek, M. S. (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**, 1045–1056.

[ptb]

Table 1

Simulation results for genetic distance estimates under different sample sizes ($S = 50, 100, \text{ and } 400$), different interference parameter values ($m = 0, 1, 2, \text{ and } 3$), and different proportions of uninformative/untyped data ($\mu = 0\%, 25\%, \text{ and } 50\%$). Five equally spaced markers with 20cM between each pair of consecutive markers were considered in these simulations. For simplicity, we only summarize the results for the estimate of the first interval (d_1) based on 100 simulated data sets. For each simulated data set, we estimated the genetic distance and the standard error for each distance estimate. The results are the averages of these 100 distance estimates and the averages of these 100 standard error estimates.

m	$d_1: \text{cM}$								
	$S = 50$			$S = 100$			$S = 400$		
	$\mu = 0$	$\mu = 0.25$	$\mu = 0.50$	$\mu = 0$	$\mu = 0.25$	$\mu = 0.50$	$\mu = 0$	$\mu = 0.25$	$\mu = 0.50$
0	22(13)	20(10)	18(12)	21(7)	19(7)	19(8)	20(3)	21(3)	20(4)
1	21(9)	22(10)	22(14)	20(5)	22(6)	20(7)	20(2)	21(3)	20(3)
2	20(7)	21(7)	20(10)	21(5)	22(6)	19(6)	21(2)	21(2)	20(3)

[ptb]

Table 2

For each chi-square model ($m = 0, 1, 2, 3$), 100 data sets of sample size 100 were generated with 5 equally spaced markers. The genetic distance between each pair of markers was set at 20cM. The simulated data sets were fitted to different $Cx(Co)^m$ models, where $m = 0, 1, \dots, 5$. The average of the maximized log-likelihoods (L) under each fitted model is presented in the upper portion of this table, the corresponding standard deviation is given in the parentheses. Among the 100 data sets generated under each model, the number of times (C) a certain model had the largest likelihood is summarized in the lower portion of this table.

		χ^2 model(m)					
		0	1	2	3	4	5
L	0	-188 (12)	-189 (12)	-191 (11)	-193 (11)	-195 (11)	-197 (11)
	1	-208 (12)	-207 (11)	-208 (11)	-209 (11)	-210 (11)	-211 (11)
	2	-217 (11)	-219 (10)	-214 (10)	-215 (10)	-216 (10)	-217 (10)
	3	-220 (10)	-218 (10)	-216 (9)	-216 (9)	-216 (9)	-216 (9)
	0	73	19	7	1	0	0
	1	27	40	26	3	3	1
	2	14	20	27	19	16	14
	3	0	10	19	25	21	25

[ptb]

Table 3

The estimated genetic distances among the genetic markers from the UPD15 data under different chi-square models. In the table, the d_i are the estimated genetic distance between the $i - 1$ th and the i th markers, the m is the order of the chi-square model, and L is the maximized log-likelihood under a given chi-square model. The column “soton” gives the estimated genetic distances among these markers from normal meioses.

	chi-square models (m)							Soton
	0	1	2	3	4	5	6	
d_1	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	3.8
d_2	6.0 (2.5)	5.3 (1.7)	5.4 (1.7)	5.5 (0.8)	5.6 (3.3)	5.9 (0.8)	5.7 (0.8)	11.9
d_3	7.9 (5.0)	6.8 (3.3)	6.5 (3.3)	6.6 (0.8)	6.2 (0.8)	6.1 (0.8)	6.2 (3.3)	12.5
d_4	44.0 (12.5)	33.3 (5.0)	31.9 (5.8)	31.0 (5.0)	31.7 (5.0)	32.0 (4.2)	32.1 (4.2)	17.4
d_5	7.6 (6.6)	7.2 (5.0)	7.7 (5.0)	7.9 (4.2)	8.6 (4.2)	8.8 (4.2)	9.1 (3.3)	15.9
d_6	8.0 (5.0)	7.3 (4.2)	7.1 (0.8)	7.2 (1.7)	6.8 (1.7)	6.8 (0.8)	6.7 (1.7)	11.5
d_7	50.4 (10.0)	33.6 (7.5)	28.8 (4.2)	27.9 (1.7)	26.3 (1.7)	25.8 (3.3)	25.5 (3.3)	19.4
d_8	6.7 (6.6)	5.2 (3.3)	4.8 (2.5)	4.8 (2.5)	4.7 (2.5)	4.6 (2.5)	4.7 (1.7)	4.2
d_9	29.8 (8.3)	20.4 (6.6)	18.5 (6.6)	18.2 (1.7)	18.1 (1.7)	17.9 (1.7)	18.0 (1.7)	38.4
d_{10}	29.0 (7.5)	22.2 (5.8)	20.9 (4.2)	20.5 (1.7)	20.6 (1.7)	20.7 (2.5)	20.6 (1.7)	13.0
$\sum d_i$	189.4	141.3	131.6	129.6	128.9	128.6	128.6	148.1
L	-158.1	-157.2	-156.7	-156.4	-156.2	-156.1	-156.1	