ARTICLE

Accommodating Chromosome Inversions in Linkage Analysis

Gary K. Chen, Erin Slaten, Roel A. Ophoff, and Kenneth Lange

This work develops a population-genetics model for polymorphic chromosome inversions. The model precisely describes how an inversion changes the nature of and approach to linkage equilibrium. The work also describes algorithms and software for allele-frequency estimation and linkage analysis in the presence of an inversion. The linkage algorithms implemented in the software package Mendel estimate recombination parameters and calculate the posterior probability that each pedigree member carries the inversion. Application of Mendel to eight Centre d'Étude du Polymorphisme Humain pedigrees in a region containing a common inversion on 8p23 illustrates its potential for providing more-precise estimates of the location of an unmapped marker or trait gene. Our expanded cytogenetic analysis of these families further identifies inversion carriers and increases the evidence of linkage.

Recent reports suggest that chromosomal rearrangements such as insertions, deletions, and inversions are more common in humans than previously believed.[1–3] These rearrangements pose a problem when mapping disease genes, and more-subtle models and statistical methods are sorely needed to deal with them. In the present study, we discuss the issue of inversion polymorphisms in gene mapping and how this barrier to proper statistical inference can be surmounted through a modified algorithm for linkage analysis.

Various large-scale rearrangements have been characterized elsewhere by statistical analysis of genotype data.[4,5] Genomic rearrangements include duplications, deletions, insertions, and inversions in stretches of DNA with a range of <1 kb to >5 Mb.[6] The substrates for these common rearrangements are generally highly homologous sequences of low complexity, known as "low-copy repeats" (LCRs).[7] The LCRs extend ~10–400 kb and flank the rearranged genomic segment.[8] Different orientations of the LCRs can lead to aberrant recombination events. With the same orientation of repeat sequences upstream and downstream of a given region, misregistered pairing of homologues can occur during meiosis, leading to loss (deletion) or gain (insertion) of genetic material in the two resulting gametes. If the repeat sequences flanking a stretch of DNA are inverted with respect to one another and the region bends into a loop structure during meiosis, then an intrachromosomal recombination event may occur and cause an inversion of the DNA segment. Inversions on 4p16 and 8p23 are flanked by clusters of olfactory-receptor genes, which are likely the substrates for these intrachromosomal rearrangements.[9]

Several characterized inversions are associated with deleterious phenotypes. Disruption of critical regulatory or coding sequences via such rearrangements has been implicated in certain rare diseases known as "genomic disorders."[6,10] Notable examples are hemophilia A,[11,12] Prader-Willi or Angelman syndrome,[13,14] Williams-Beuren syndrome,[15,16] and Hunter syndrome.[17] Other inversions are believed to be selectively neutral or advantageous. Cytogenetic analyses of unaffected individuals have uncovered common neutral inversions on chromosome 9,[18] 4p16,[9] and 8p23.[19] Stefansson et al.[4] propose that an inversion on 17q21 with a frequency of ~21% in Europeans provides a selective reproductive advantage. A comparison between chimpanzee and human maps reveals polymorphic inversions on 7p22, 7q11, and 16q24, with minor-allele frequencies in the range of 5%–48%. These inversions may be a driving force in primate evolution.[20]

Taking into account genomic structural variation is crucial in linkage studies of human diseases. When a fixed marker order is assumed for all individuals in an inverted region, one tends to see spurious recombination events among inversion carriers. The traditional reaction has been to inflate map distances. Because this involves so many internal contradictions, it is better to invoke genotyping error[5] and discard some observations. In a recent study that compared genetic map distances across populations, genotyping errors could explain discrepancies in map distances in some regions but not in the large 8p23 inversion.[21] Many investigators now exclude markers within the problematic 8p23 region when conducting genome screens for complex traits.[22–24] Although such caution is understandable, it is bound to result in failure if the disease gene falls within the inversion.

In this work, we present a mathematical model, statistical methods, and likelihood algorithms for dealing with chromosome inversions. These methods take as known the population frequency of the inverted chromosome and the boundaries of the inversion. To validate our theory and methods, we implemented them in the software package Mendel (UCLA Human Genetics Web site) and per-

From the Departments of Human Genetics (G.K.C.; R.A.O.; K.L.) and Biomathematics (K.L.) and the Center for Neurobehavioral Genetics (E.S.; R.A.O.), David Geffen School of Medicine at the University of California at Los Angeles (UCLA), Los Angeles

Received February 8, 2006; accepted for publication April 25, 2006; electronically published June 6, 2006.

Address for correspondence and reprints: Dr. Kenneth Lange, Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-7088. E-mail: klange@mednet.ucla.edu

Am. J. Hum. Genet. 2006;79:238–251. 0002-9297/2006/7902-0007$15.00

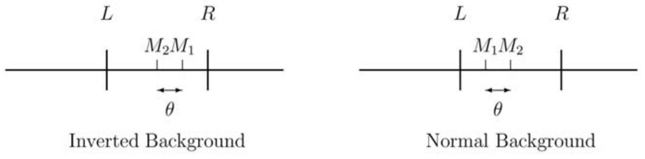

Figure 1. Chromosome-inversion diagram with both markers inside the inversion.

formed an analysis of eight CEPH families over an 18-cM region on chromosome 8p23. On the basis of the same eight CEPH families, Broman et al.[5] demonstrated that the region harbors an inversion spanning ~12 cM in females and ~2 cM in males. Our analysis confirms these estimates and shows that one can map candidate genes more precisely in an inverted region by using modified versions of standard software.

Before discussing the results of our data analysis, we turn to a mathematical model that sheds light on the nature of genetic equilibrium in the presence of inversions and the rate at which equilibrium is approached. Our mathematical model shows that allele frequencies and haplotype frequencies specific to normal and inverted chromosomes decay over time to predictable equilibrium values. Suppression of recombination between normal and inverted chromosomes slows the approach to linkage equilibrium. Because of the complexity of the model, our most demanding mathematical derivations will be relegated to the appendixes.

Methods

Linkage Equilibrium

Under the population assumptions typically invoked in modeling Hardy-Weinberg and linkage equilibrium, convergence to equilibrium within an inversion tends to be slower than convergence to equilibrium in comparable regions outside the inversion. Suppression of recombination between an inversion and its normal counterpart, effectively demonstrated in balancer chromosome technology,[25] explains this phenomenon.

To explore the population dynamics of a pair of markers, we first must understand the dynamics at a single marker. An allele a of a marker can exist on an inverted or a normal chromosome background. Let $p^m_{a|i}$ denote its frequency (conditional probability) at generation m on the inverted background, and let $p^m_{a|n}$ denote its frequency on a normal background. If the marker falls inside the inversion, then these frequencies do not change over time because of suppression of recombination. In symbols,

$$p^m_{a|i} = p_{a|i} \text{ and } p^m_{a|n} = p_{a|n} \; . \tag{1}$$

This result is in striking contrast to the case where the marker falls outside the inversion. Recombination can transfer an allele from an inverted background to a normal background and vice versa. The dynamics of the process involve the population frequency q of the inversion and the recombination fraction θ separating the marker and the nearer inversion boundary. Assuming that the production of the next generation occurs by random union of gametes, we can write the recurrence

$$qp^{m+1}_{a|i} = q^2p^m_{a|i} + 2q(1-q)\frac{1}{2}\left[(1-\theta)p^m_{a|i} + \theta p^m_{a|n}\right] \tag{2}$$

for the frequency (joint probability) of a gamete bearing allele a and the inversion. The first term, $q^2p^m_{a|i}$, on the right of the recurrence (eq. [2]) is the probability that the parent of the gamete possesses two inverted chromosomes and passes allele a at the designated marker. The second term on the right is the probability that the parent possesses one inverted and one normal chromosome and passes allele a. Dividing equation (2) by q gives the textbook recurrence

$$p^{m+1}_{a|i} = qp^m_{a|i} + (1-q)\left[(1-\theta)p^m_{a|i} + \theta p^m_{a|n}\right] \tag{3}$$

for convergence to linkage equilibrium. This is hardly surprising, because the nearer inversion boundary can be thought of as a second marker with the two alleles i and n. The normal chromosome counterpart of recurrence (eq. [3]) is

$$p^{m+1}_{a|n} = (1-q)p^m_{a|n} + q\left[(1-\theta)p^m_{a|n} + \theta p^m_{a|i}\right] \; . \tag{4}$$

We present this classic derivation in detail because other arguments to come are patterned on it. Standard transformation of recurrences (3) and (4) yield the recurrences

$$p^{m+1}_{a|i} - p_a = (1-\theta)(p^m_{a|i} - p_a)$$

and

$$p^{m+1}_{a|n} - p_a = (1-\theta)(p^m_{a|n} - p_a) \; , \tag{5}$$

with $p_a = qp^m_{a|i} + (1-q)p^m_{a|n}$ representing the constant overall frequency of allele a. These recurrences show that equilibrium is approached at the geometric rate $(1-\theta)$. Again we stress that this result is extremely well known.[26]

We now examine convergence to equilibrium with two markers. Denote the population frequency of an allele b at the second locus by either $p^m_{b|i}$ or $p^m_{b|n}$, and denote the frequency of a haplotype with allele a at the first marker and allele b at the second marker by either $p^m_{ab|i}$ or $p^m_{ab|n}$. As before, these conditional probabilities depend on the generation number m and the inverted background i or the normal background n. There are four cases to consider, depending on the location of the markers relative to the inversion boundaries.

The simplest case involves two markers located inside the inversion with a recombination fraction θ separating them (see fig.

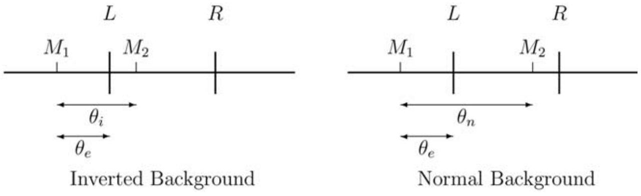

Figure 2. Chromosome-inversion diagram with one marker flanking the inversion and the other located within it.

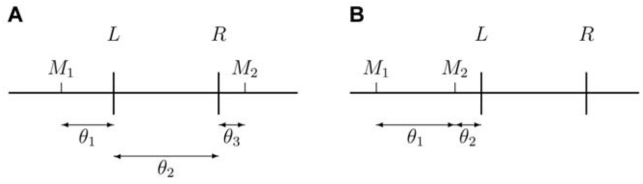

Figure 3. Chromosome-inversion diagram with both markers flanking the inversion. *A,* Markers on opposite sides. *B,* Markers on the same side.

1). The random-union-of-gametes argument now leads to the haplotype recurrence

$$qp_{ab|i}^{m+1} = q^2(1-\theta)p_{ab|i}^m + q^2\theta p_{a|i}^m p_{b|i}^m + 2q(1-q)\frac{1}{2}p_{ab|i}^m \ .$$

Dividing this by q and subtracting the equilibrium frequency $p_{a|i}p_{b|i}$ gives the amended recurrence

$$p_{ab|i}^{m+1} - p_{a|i}p_{b|i} = (1-q\theta)(p_{ab|i}^m - p_{a|i}p_{b|i}) \ . \tag{6}$$

The analogous recurrence for haplotypes on the normal background is

$$p_{ab|n}^{m+1} - p_{a|n}p_{b|n} = \left[1-(1-q)\theta\right](p_{ab|n}^m - p_{a|n}p_{b|n}) \ . \tag{7}$$

The recurrences (6) and (7) show that equilibrium is approached at the geometric rates $1-q\theta$ and $1-(1-q)\theta$, respectively. These resemble the familiar rate of $1-\theta$ in the absence of an inversion, except for the factors of q and $1-q$ modifying θ. Thus, the approach to equilibrium is slowed by the presence of the inversion.

Figure 2 depicts the situation in which one marker flanks the inversion and the other falls inside the inversion. Here, θ_e denotes the recombination fraction between marker M_1 (the external marker), the left inversion boundary θ_i denotes the recombination fraction between M_1 and M_2 on the inverted background, and θ_n denotes the recombination fraction between M_1 and M_2 on the normal background. In this notation, it is straightforward to derive the recurrences

$$qp_{ab|i}^{m+1} = q^2(1-\theta_i)p_{ab|i}^m + q^2\theta_i p_{a|i}^m p_{b|i} + 2q(1-q)\frac{1}{2}(1-\theta_e)p_{ab|i}^m$$
$$+2q(1-q)\frac{1}{2}\theta_e p_{a|n}^m p_{b|i}$$

and

$$(1-q)p_{ab|n}^{m+1} = (1-q)^2(1-\theta_n)p_{ab|n}^m + (1-q)^2\theta_n p_{a|n}^m p_{b|n}$$
$$+2q(1-q)\frac{1}{2}(1-\theta_e)\,p_{ab|n}^m + 2q(1-q)\frac{1}{2}\theta_e p_{a|i}^m p_{b|n}$$

with use of the relationships $p_{b|i}^m = p_{b|i}$ and $p_{b|n}^m = p_{b|n}$ for the marker inside the inversion. Dividing these two recurrences by q and $1-q$, respectively, leads to the vector recurrence

$$\begin{pmatrix} p_{ab|i}^{m+1} \\ p_{ab|n}^{m+1} \end{pmatrix}$$
$$= \begin{bmatrix} q(1-\theta_i)+(1-q)(1-\theta_e) & 0 \\ 0 & (1-q)(1-\theta_n)+q(1-\theta_e) \end{bmatrix}$$
$$\times \begin{pmatrix} p_{ab|i}^m \\ p_{ab|n}^m \end{pmatrix} + \begin{bmatrix} q\theta_i p_{b|i} & (1-q)\theta_e p_{b|i} \\ q\theta_e p_{b|n} & (1-q)\theta_n p_{b|n} \end{bmatrix} \begin{pmatrix} p_{a|i}^m \\ p_{a|n}^m \end{pmatrix} \ . \tag{8}$$

One can easily check that the choices $p_{ab|i} = p_a p_{b|i}$ and $p_{ab|n} = p_a p_{b|n}$ furnish a fixed point of recurrence (8) when the single-marker equilibria $p_{a|i}^m = p_{a|n}^m = p_a$ are in force for the marker flanking the inversion. Appendix A undertakes a thorough mathematical analysis of this interesting dynamic system.

In figure 3*A,* the markers flank opposite sides of the inversion. In addition to the recombination fractions θ_1, θ_2, and θ_3 appearing in figure 3*A,* we need the recombination fraction θ_{12} separating M_1 and M_2 on a normal chromosome background. If we postulate no chiasma interference, then Trow's formula[27] determines θ_{12} according to

$$1-2\theta_{12} = (1-2\theta_1)(1-2\theta_2)(1-2\theta_3) \ .$$

In this notation, the logic already invoked leads to the vector recurrence

$$\begin{pmatrix} p_{ab|i}^{m+1} \\ p_{ab|n}^{m+1} \end{pmatrix} = M\begin{pmatrix} p_{ab|i}^m \\ p_{ab|n}^m \end{pmatrix} + N\begin{pmatrix} p_{a|i}^m p_{b|i}^m \\ p_{a|i}^m p_{b|n}^m \\ p_{a|n}^m p_{b|i}^m \\ p_{a|n}^m p_{b|n}^m \end{pmatrix} \ , \tag{9}$$

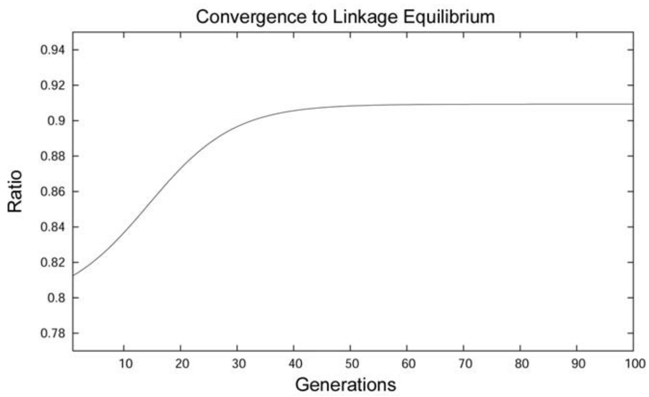

Figure 4. The ratio (12) summarizing convergence to linkage equilibrium as a function of generation number, *m.*

Table 1. Parameter Values for the Demonstration of Convergence to Linkage Equilibrium

Parameter[a]	Parameter Value
Initial:	
$p_{ab\|i}$	.30
$p_{a\|i}$	.30
$p_{b\|i}$	.60
$p_{a\|n}$	.25
$p_{b\|n}$	.60
Equilibrium:	
p_a	.28
p_b	.58
Inversion:	
q	.60

NOTE.—Genetic distances are based on Haldane's function. For $\theta_{12} = .23$, the distance is 30 cM (the total distance between the two markers as computed from Trow's formula); for $\theta_1 = \theta_2 = \theta_3 = .09$, the distance is 10 cM.

[a] Parameter names follow the conventions used in the "Linkage Equilibrium" subsection of the "Methods" section.

where M is the matrix

$$q\begin{bmatrix} 1-\theta_{12} & 0 \\ \theta_1\theta_3 & (1-\theta_1)(1-\theta_3) \end{bmatrix} + (1-q)\begin{bmatrix} (1-\theta_1)(1-\theta_3) & \theta_1\theta_3 \\ 0 & 1-\theta_{12} \end{bmatrix} \tag{10}$$

and N is the matrix

$$q\begin{bmatrix} \theta_{12} & 0 & 0 & 0 \\ 0 & \theta_1(1-\theta_3) & (1-\theta_1)\theta_3 & 0 \end{bmatrix} + (1-q)\begin{bmatrix} 0 & (1-\theta_1)\theta_3 & \theta_1(1-\theta_3) & 0 \\ 0 & 0 & 0 & \theta_{12} \end{bmatrix} .$$

At the single-marker equilibria, the choices $p_{ab|i} = p_{ab|n} = p_a p_b$ furnish a fixed point of the recurrence (9).

Finally, figure 3*B* depicts the case in which the markers flank the same side of the inversion. On the basis of the recombination fractions indicated in the figure, the conditional haplotype frequencies also obey the vector recurrence (9) but with matrix M defined as

$$q\begin{bmatrix} 1-\theta_1 & 0 \\ (1-\theta_1)\theta_2 & (1-\theta_1)(1-\theta_2) \end{bmatrix} + (1-q)\begin{bmatrix} (1-\theta_1)(1-\theta_2) & (1-\theta_1)\theta_2 \\ 0 & 1-\theta_1 \end{bmatrix} \tag{11}$$

and matrix N defined as

$$q\begin{bmatrix} \theta_1 & 0 & 0 & 0 \\ 0 & \theta_1(1-\theta_2) & \theta_1\theta_2 & 0 \end{bmatrix} + (1-q)\begin{bmatrix} 0 & \theta_1\theta_2 & \theta_1(1-\theta_2) & 0 \\ 0 & 0 & 0 & \theta_1 \end{bmatrix} .$$

At the single-marker equilibria, the choices $p_{ab|i} = p_{ab|n} = p_a p_b$ again furnish a fixed point of the recurrence (9).

It is instructive to consider the deceleration of convergence to linkage equilibrium when the markers flank the inversion as shown in figure 3*A*. To assess the rate of convergence, figure 4 plots the ratio

$$\frac{|p_{ab|i}^{m+1} - p_a p_b|}{|p_{ab|i}^{m} - p_a p_b|} \tag{12}$$

as a function of *m,* starting at the initial parameter values in table 1. The ratio (12) stabilizes after 40 generations at ~0.91. By contrast, for a similar chromosome segment without an inversion, the ratio

$$\frac{|p_{ab}^{m+1} - p_a p_b|}{|p_{ab}^{m} - p_a p_b|} = 1 - \theta \tag{13}$$

remains constant at the value 0.77. A higher ratio is indicative of a slower rate of convergence to linkage equilibrium. Although it is intuitively obvious that the limit of the ratio (12) should exceed the ratio (13), the mathematical theory of appendix A precisely predicts the difference.

Allele-Frequency Estimation

As already noted, allele frequencies may differ on normal and inverted backgrounds for a marker in an inverted region. This raises the question of how to estimate their population frequencies from a random sample of individuals. The problem is more subtle than it first appears. Consider the genotype frequencies in table 2, which are based on codominant alleles, where q is the known frequency of the inversion and $p_{a|n}$ and $p_{a|i}$ are the frequencies of allele a on normal and inverted backgrounds, respectively. If the random sample contains no information on

Table 2. Population Frequencies of Various Genotypes

No. of Inverted	Population Frequency by Genotype	
Chromosomes	Homozygous *a/a*	Heterozygous *a/b*
Unknown	$[(1-q)p_{a\|n} + qp_{a\|i}]^2$	$2[(1-q)p_{a\|n} + qp_{a\|i}][(1-q)p_{b\|n} + qp_{b\|i}]$
0	$(1-q)^2 p_{a\|n}^2$	$(1-q)^2 2p_{a\|n}p_{b\|n}$
1	$2q(1-q)p_{a\|n}p_{a\|i}$	$2q(1-q)[p_{a\|n}p_{b\|i} + p_{a\|i}p_{b\|n}]$
2	$q^2 p_{a\|i}^2$	$q^2 2p_{a\|i}p_{b\|i}$

Table 3. Genetic Distances for the CEPH Data

	Map Coordinates (cM)	
Marker	Female	Male
D8S1706	6.51	14.44
Left boundary[a]	9.68	16.07
D8S351	12.85	17.69
D8S1130	25.00	19.84
Right boundary[a]	27.75	20.92
D8S552	30.50	22.00
D8S1754	33.00	22.00

[a] Assumed inversion boundaries according to Broman et al.[5]

inversion status, then the likelihood contains only factors such as

$$[(1-q)p_{a|n}+qp_{a|i}]^2$$

for homozygotes and

$$2[(1-q)p_{a|n}+qp_{a|i}][(1-q)p_{b|n}+qp_{b|i}]$$

for heterozygotes. This makes it clear that only the convex combinations—$(1-q)p_{a|n}+qp_{a|i}$—can be estimated. For a given allele *a,* neither $p_{a|n}$ nor $p_{a|i}$ by itself is identifiable. This situation radically changes when information on inversion status is available.

Because of the expense of inversion assignment, data from mixed samples of assigned and unassigned individuals are most likely to be used in allele-frequency estimation. Appendix B derives an expectation-maximization (EM) algorithm pertinent to this situation. In the "Results" section, we illustrate the algorithm in action on genotypes gleaned from two different data sets.

Algorithms for Linkage Analysis

We now turn to gene-mapping issues raised by our population-genetics model. To keep the algorithms as simple as possible, we assume linkage equilibrium in the extended sense just discussed as well as Haldane's model of recombination. Because of the demands of multipoint mapping, we must deal with linkage equilibrium involving more than two markers. The obvious rule in computing haplotype frequencies in this context is to take the product of ordinary allele frequencies outside the inversion and allele frequencies specific to inverted or normal chromosomes inside the inversion. Because of the inevitable mathematical complexities, we did not formally investigate the rate of convergence to equilibrium for three or more markers, but there can be little doubt that convergence occurs given enough time. Furthermore, the rate of convergence is almost certainly slower than it would be in the absence of the inversion, because of recombinations suppression.

For a marker inside the inversion, a user of Mendel has the option of entering two sets of allele frequencies in the locus file: one set pertinent to the normal background and one set pertinent to the inverted background. If genotyping data on individuals with known inversion status are unavailable, then one can ignore the inversion in estimating allele frequencies and assume identical allele frequencies on normal and inverted backgrounds. This decision entails some risk, but the alternative of refraining from linkage analysis altogether is even less appealing.

The likelihood *L* of a pedigree of *n* people is usually written as

$$L=\sum_{G_1}\cdots\sum_{G_n}\prod_i \mathrm{Pen}(X_i\mid G_i)\prod_j \mathrm{Prior}(G_j)\prod_{\{k,l,m\}}\mathrm{Tran}(G_m\mid G_k,G_l)\ , \tag{14}$$

where $\mathrm{Pen}(X_i\mid G_i)$ is the penetrance of the phenotype X_i of person *i* given his or her genotype G_i, $\mathrm{Prior}(G_j)$ is the prior probability of genotype G_j of founder *j,* and $\mathrm{Tran}(G_m\mid G_k,G_l)$ is the probability that parents *k* and *l* with genotypes G_k and G_l transmit genotype G_m to their child *m.*[28] There are two competing deterministic algorithms for evaluating the likelihood (14). The Lander-Green-Kruglyak algorithm[29–31] considers all pedigree members collectively as it marches from one locus to the next. It is difficult for that algorithm to accommodate both inverted and normal chromosomes, since the physical order of the loci is no longer rigidly maintained. The classic Elston-Stewart[32] algorithm takes all loci simultaneously, but one person at a time. Thus, it is flexible enough to consider alternative locus orders and alternative interlocus distances.

Because the Elston-Stewart algorithm scales exponentially in the number of loci, it is impossible to handle more than a few loci within any likelihood evaluation or maximum-likelihood run. In computing location scores, Mendel operates by considering a window of markers centered on a putative trait location. The size of the window is adjusted by the user. The genotypes that Mendel hands to its internal likelihood-evaluation routines are multilocus, with phased maternal and paternal haplotypes. Mendel assumes penetrances $\mathrm{Pen}(X_i\mid G_i)$ that factor into locus-specific penetrances. In practice, most of these are purely qualitative, with values limited to 0 or 1. More-complicated penetrances are possible, particularly at the trait locus. The prior probability $\mathrm{Prior}(G_j)$ of genotype G_j factors into maternal and paternal haplotype frequencies computed under linkage equilibrium by the product rule just described.

To distinguish the inversion background of a haplotype and whether markers fall inside or outside the inversion, left and right boundary markers must be included in each Mendel run. The left

Table 4. Inversion Status of the Parents of Eight CEPH Families on the Basis of Two-Color FISH

	Inversion Status[a]	
Family	Father	Mother
102[b]	I/I	N/I
1331[c]	N/I	NA
1332	N/N	N/I
1347	N/I	N/I
1362	N/I	I/I
1413	N/N	I/I
1416	N/I	I/I
884	I/I	N/I

[a] N = normal background; I = inverted background.

[b] Although individual 102 was scored, she was omitted from statistical analysis.

[c] Individual 1331 was not scored because of technical difficulties with the cell line. NA = not applicable.

Table 5. Estimated Allele Frequencies for *D8S1130* on Each Background

Allele Frequency by Background		
Uniform[a]	Normal	Inverted
.1994	.1805	.2120
.2061	.3505	.1095
.1935	.0869	.2647
.2173	.1408	.2684
.0871	.0001	.1452
.0967	.2412	.0001

[a] Frequencies assumed equal on the two backgrounds.

boundary marker is biallelic, with allele 1 flagging normal chromosomes and allele 2 flagging inverted chromosomes. The right boundary marker is monoallelic. Ordinarily, genotypes at the boundary markers are recorded as blank in the Mendel input pedigree file. If cytogenetic information is available to support a specific assignment at the left boundary marker, then the assignment can be entered for the corresponding person in the pedigree file. Otherwise, all four combinations of inverted and normal chromosomes are considered possible for the person.

The transmission probabilities $\mathrm{Tran}(G_m \mid G_k, G_l)$ are the most complicated feature encountered in computing pedigree likelihoods. The background information provided by the alleles at the left boundary marker determine the order of loci inside the inversion and whether recombination is suppressed. With two parental haplotypes of normal background, no modifications of the input map data are necessary. If the two haplotypes have opposite backgrounds, then recombination is suppressed in the inverted region. When both haplotypes occur on inverted backgrounds, the order of the internal loci is reversed. For example, if the boundaries and internal loci occur in the order $L - M_1 - M_2 - M_3 - R$ on a normal background, then they occur in the order $L - M_3 - M_2 - M_1 - R$ on the inverted background. Reversal switches the nearest internal loci (M_1 and M_3) neighboring the left and right boundary markers L and R as well as the map distances separating these boundaries and their nearest neighbors. Map distances over adjacent intervals are additive, and recombination fractions θ can be computed from map distances d by Haldane's formula

$$\theta = \frac{1}{2}(1 - e^{-2d}) \ .$$

Equivalently, recombination fractions over adjacent intervals can be combined via Trow's formula. Regardless of whether Haldane's or Trow's method is used, all transmission probabilities are computable with careful bookkeeping.

In addition to asking where a marker or trait likely maps, many researchers will be interested in which family members carry inverted chromosomes. The best statistical approach to this question is to compute the posterior probability of each possible chromosome background for each family member not subjected to cytogenetic analysis with FISH. Taking phase into account, there are four such posterior probabilities per person. If we let X denote the phenotype vector for the entire pedigree and $G_{il} = a/b$ denote the event that person i has genotype a/b at the left boundary marker l, then the relevant conditional probability is

$$\Pr(G_{il} = a/b \mid X) = \frac{\Pr(G_{il} = a/b,\ X)}{\Pr(X)} \ .$$

Each of the joint likelihoods $\Pr(G_{il} = a/b,\ X)$ can be computed as an ordinary likelihood if we force all of i's genotypes except a/b at the left boundary marker to have zero or near-zero penetrance.

Marker Mapping of Eight CEPH Families

To evaluate the effectiveness of our gene-mapping software, we analyzed data from the eight CEPH families deposited at the Marshfield Clinic Research Foundation Web site. Broman et al.[5] observed certain individuals in these families with an abnormally high number of apparent crossovers in a short interval on chromosome region 8p23. The most plausible explanation for this phenomenon is that the individuals carry a common chromosome inversion. Subsequent FISH experiments by the same authors confirmed this hypothesis and suggested a population frequency of 21% for the inversion, on the basis of data from 50 unrelated individuals.[5]

Given the wealth of genetic data on these CEPH families and the experimentally verified presence of inversions, we decided to test our software on the families. Table 3 summarizes a stripped-down marker map of the region taken from the study by Broman et al.,[5] with use of genetic distances provided by the Marshfield Web site. The markers in this map are selected for informativeness and are sufficiently well spaced to give a hint of the increased power of our software to map a trait relative to an established marker map. Here, we take marker *D8S351* as the trait in a location-score analysis.

Haplotyping the families with the haplotyping option of Mendel detects two families with possible triple-recombination events under the normal map. These occur in the inversion interval in individuals 10 and 11 of family 1362 and in individuals 3 and 9 of family 1413. The remaining six families do not show any unusual recombination patterns.

Allele frequencies for the trait (*D8S351*) and all markers except

Table 6. Maximum Location Scores When Marker *D8S351* Is Positioned in the Interval between the Left Boundary and Marker *D8S1130*

	Maximum Location Scoreby Method		
Family	Standard	Inversion	FISH[a]
All	27.8	32.7	34.2
1362	1.7	5.8	5.8
1413	3.8	4.0	5.4
1347	2.7	2.7	2.7
1416	1.8	1.8	1.8
884	6.0	6.0	6.0
102	6.8	6.8	6.8
1331	3.3	3.3	3.3
1332	3.0	3.0	3.0

[a] Cytogenetic data are included with inversion-specific mapping option.

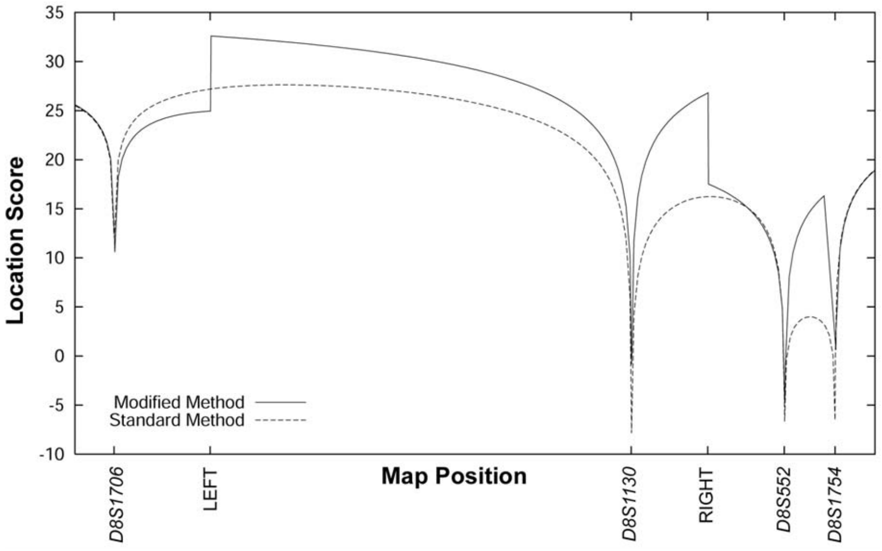

Figure 5. Location-score curves for all families with use of the standard method (*dashed curve*) versus the inversion-specific method (*solid curve*). LEFT = left boundary; RIGHT = right boundary.

D8S1130 were estimated by Mendel. Although these estimates assume the same frequencies on normal and inverted backgrounds, they do take full account of the correlations between family members.[33] Genotypes on the intrainversion marker *D8S1130* are also available for 656 parents of the nuclear families ascertained for autism that are available from the Autism Genetics Resource Exchange.[34] As a test case for the EM algorithm, we combined these 656 genotypes with the genotypes of 16 parents from the CEPH families, 14 of which have inversion status determined. Because alleles are scored differently in the two data sets, we used the program MicroMerge[35] to find a suitable alignment between alleles of the data sets. To improve parameter estimates in the EM algorithm, we combined alleles via Mendel, so that the minimum frequency for any allele was at least 5%.

It is a priori obvious that including cytogenetic status should increase the power to detect linkage. In the study of Broman et al.,[5] the mothers of families 1413 and 1362 were found by two-color FISH to be homozygous for the inversion. To increase the information available for allele-frequency estimation and linkage analysis, we replicated the earlier inversion assignments and attempted to determine the status of the remaining parents in the eight CEPH families. FISH was performed on metaphase chromosomes from lymphoblastoid cell lines (Coriell Cell Repositories) from all parents except the mother of family 1331. BAC clone

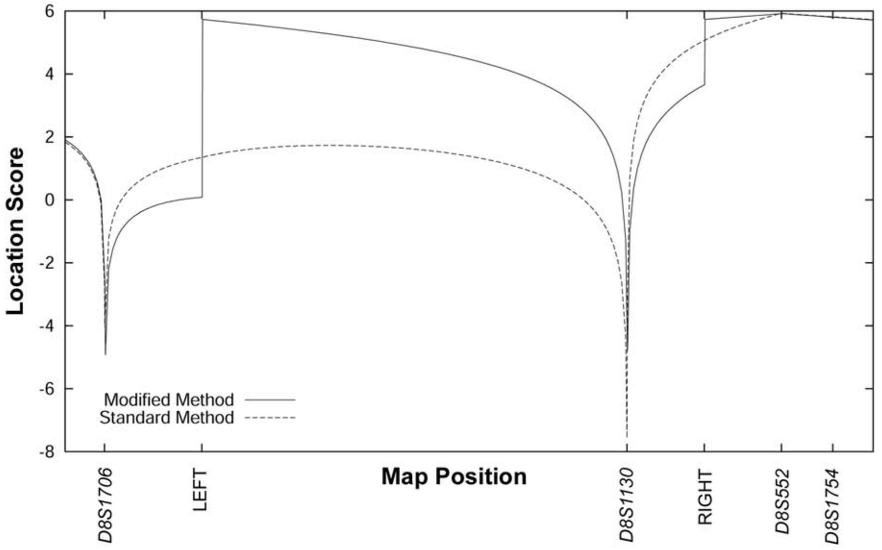

Figure 6. Location-score curves for family 1362 with use of the standard method (*dashed curve*) versus the inversion-specific method (*solid curve*). LEFT = left boundary; RIGHT = right boundary.

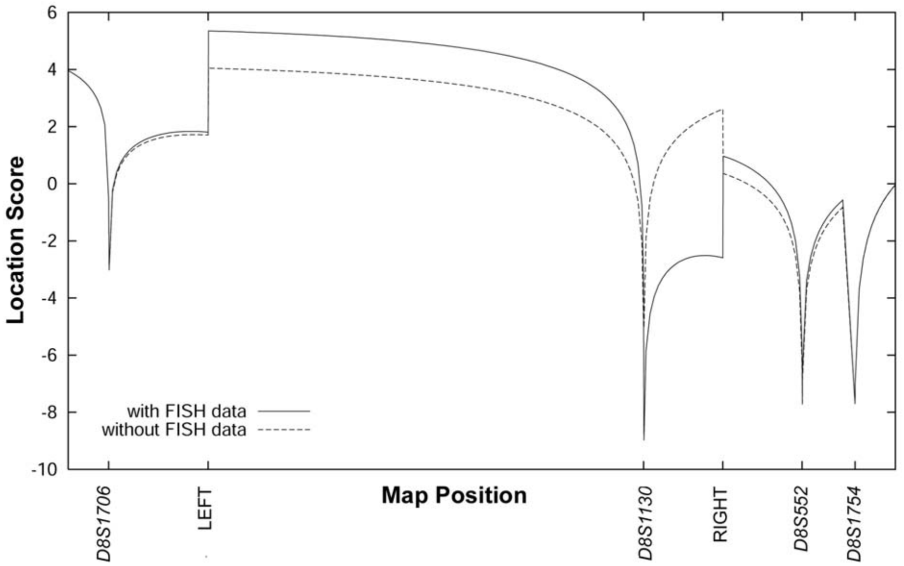

Figure 7. Location-score curves for family 1413 with the mother unkaryotyped (*dashed curve*) versus with the mother karyotyped (*solid curve*). LEFT = left boundary; RIGHT = right boundary.

GS173O4, located at the distal boundary of the 8p23 inversion,[9] and BAC clone RP11-235O5, located near the proximal boundary of the 8p23 inversion at positions 10593770-10642582, were used as probes. UCSC build hg17 indicates that both probes are located inside the inversion. Probe and slide preparation, DNA hybridization, and analysis were performed by conventional methods. At least 20 cells per case subject were analyzed by direct microscopic visualization and digital-imaging analysis.

Results

Results of the cytogenetic experiments are shown in table 4. The mother of family 102 has a questionable inversion status. She appears to carry one normal and one inverted chromosome, contradicting her involvement in an obligate recombination event in the inversion interval. Thus, we excluded her inversion status in statistical analysis. From the remaining 14 FISH scores, we estimated an inversion frequency of 60% in the CEPH families. For marker *D8S1180,* the EM algorithm converges to the background-specific allele frequencies shown in table 5. The table also lists, for comparison, estimates that ignore chromosome background. To assess the significance of these findings, one can compute a likelihood-ratio test comparing the null hypothesis of equal allele frequencies to the alternative hypothesis of background-specific frequencies. Twice the difference of maximum log-likelihoods yields a χ^2 statistic of 3.5 (5 df) for the six alleles of *D8S1180.* This corresponds to a *P* value of 0.62, if we ignore the failure of some estimates to remain on the interior of the parameter space. In any case, there is little reason to reject the null hypothesis, and we feel justified in invoking it in further analysis.

We computed location scores on the eight families both jointly and separately, assuming an inversion frequency of 60%. In the first set of analyses reported here, we designated the inversion status of all pedigree members as "unknown." To assess whether adding cytogenetic information would improve location-score estimates, we reanalyzed the eight families in accordance with Broman et al.[5] and our cytogenetic findings.

Table 6 presents the best location scores for the families analyzed with use of the standard location-scores option of Mendel and the new inversion-specific location-scores option. The table also lists location scores under the inversion-specific option when the previous and new cytogenetic data are included. Location scores >3.3 are genomewide significant, according to the Lander and Kruglyak criterion.[36] When all families are analyzed jointly, the standard Mendel option yields a maximum location score of 27.8 with the trait *D8S351* positioned between the left boundary and marker *D8S1130,* an interval that is consistent with its known location on the Marshfield map. Under the inversion-specific option, the maximum location score increases to 32.7 and positions trait *D8S351* in the same interval. The location score increases to 34.2 when cytogenetic data are included. The single-family analyses are also revealing, particularly for families 1362 and 1413. The evidence from these two families alone is insufficient to map *D8S351* very precisely, but, in combination with the evidence from the other six families, it places *D8S351* close to its known location.

Mendel has the capacity either to search a likelihood surface or to evaluate it over a grid of points. The maximum-likelihood estimates in table 6 reflect Mendel's search mode applied to the region between the left boundary and marker *D8S1130.* Some of the most salient results from

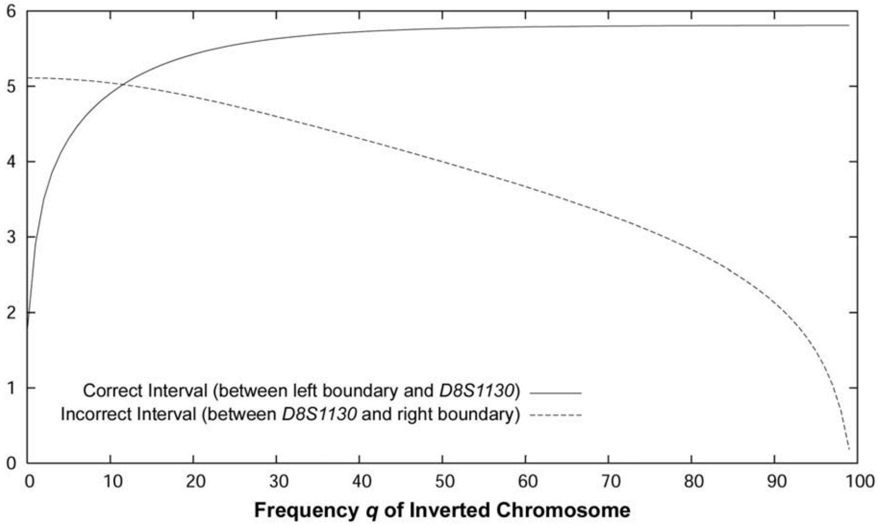

Figure 8. Maximum location scores for the placement of marker *D8S351* as a function of the inversion frequency *q*

table 6 are reinforced in figures 5 and 6, which plot the output of Mendel's grid mode. The location-score curves depicted in these figures show abrupt changes at inversion boundaries. There are two explanations for this odd behavior. First, the log-likelihood plunges downward at a marker whenever a family manifests an obligate crossover between the trait and the marker. We have repeatedly emphasized that a wrong map can produce apparent obligate crossovers. Second, the set of markers employed in the linkage computations changes as we pass from interval to interval. Because the markers are not infinitely polymorphic, the information content varies from set to set.

In the absence of cytogenetic data, our statistical analysis reveals that the mothers in both families have at least a 99% posterior probability of carrying two inverted chromosomes, a prediction confirmed by Broman et al.[5] and our cytogenetic findings. Explicitly coding the inversion status for these two individuals modestly inflates their families' location scores. In family 1413, the increase from 4.0 to 5.4 is less gratifying than the drop in linkage evidence on the presumably incorrect interval between *D8S1130* and the right boundary shown in figure 7.

We were concerned about whether our likelihood analysis would perform robustly over a range of possible inversion frequencies. Family 1362 is of particular interest, because the location-score curve peaks in standard linkage analysis on the interval between *D8S1130* and the right boundary, inconsistent with the known position of *D8S351*. The inversion method maps *D8S351* correctly between the left inversion boundary and *D8S1130* at the estimated inversion frequency of 60%. We reran Mendel's inversion option with inversion frequencies ranging from 0% to 100% for family 1362. As figure 8 indicates, once the inversion frequency exceeds 10%, the inversion method maps *D8S351* to the correct interval.

Discussion

We have demonstrated the intuitively obvious fact that chromosome inversions change the nature of genetic equilibrium and slow the rate of convergence to equilibrium. The virtue of a mathematical model is that it permits precise quantification of this phenomenon. Appendix A shows how the rate of convergence to equilibrium depends on the population frequency of the inverted chromosome and the geometry of the inversion and the markers in its vicinity.

Our model does not consider some subtleties that may be biologically relevant. For instance, it is plausible that markers outside an inversion but close to either boundary may be less likely to recombine than markers farther from a boundary. Here, we have in mind steric hindrances caused by aberrant geometries such as DNA looping. Another complication is recurrent inversion events. These might entail persistent stochastic variation in allele and haplotype frequencies. Because both objections involve second-order effects, we are comfortable in asserting that our model is reasonable given a large enough inversion, adequately spaced markers, and a relatively homogeneous population. In practice, analysis of smaller inversions may not lead to the definitive results seen with 8p23. The ability to detect inversions through aberrant recombination events diminishes as inversion length decreases. Furthermore, if phenomena such as steric hindrance occur at inversion boundaries, then these are apt to be more important with small inversions.

Our analyses successfully demonstrate that taking proper account of inversion status can substantially enhance linkage peaks in large genomic regions with a simple architecture. Broman et al.[5] cytogenetically confirmed inversions in two CEPH families that harbor individuals with highly suspect triple recombinations within a 2.5-Mb inversion in the 8p23 region between markers *D8S1705* and *D8S552*. In the case of family 1362, our modified linkage method yields a location score of 5.8 in the interval between the left boundary and *D8S1130,* in contrast to the standard method, which gives a nonsignificant location score of 1.7. More importantly, the new method correctly maps *D8S351* to the interval between the left boundary and *D8S1130*; the standard linkage method maps *D8S351* between *D8S1130* and the right boundary. Improvements in location scores were more modest for family 1413, perhaps because of less informative markers flanking the inversion boundaries. There were no changes in location scores for the remaining families, which display no recombination events within the inversion region. Cytogenetic assignment of key people is certainly helpful but not absolutely necessary for gene mapping. Including inversion status when analyzing family 1413 narrows the best region for *D8S351* from the two intervals flanking *D8S1130* to only the interval to the left of *D8S1130*. Mapping conclusions are fairly robust to misspecified inversion frequencies.

Computation of posterior probabilities of inversion status can guide cytogenetic analysis of those suspected of carrying inversions, as suggested by the concordance between high posterior probabilities (99%) and FISH results for the mothers in CEPH families 1362 and 1413. Given the expense involved, careful targeting of key individuals for FISH analysis is advisable. There is little point in subjecting individuals to FISH if they have a low posterior probability of carrying an inversion.

When a marker falls inside an inversion, the EM algorithm for estimating its allele frequencies can converge to an inferior mode in small data sets. One remedy is to start the algorithm from multiple random points and record the best mode found. Convergence can also be excruciatingly slow, taking literally thousands of iterations. Fortunately, each iteration is very fast to compute. We have demonstrated that some individuals in a random sample must have experimentally determined inversion status to estimate different allele frequencies on normal and inverted backgrounds. For data sets as small as the CEPH data featured here, it is hard to estimate allele frequencies with much precision. It is probably better to use published estimates of allele frequencies and assume equal frequencies on normal and inverted backgrounds than to rely on poor estimates from just a handful of families. Our likelihood-ratio test on a merged data set for marker *D8S1130* was unable to reject the null hypothesis of equal allele frequencies on the two backgrounds, even given a large inversion frequency.

The increased resolution afforded by a good inversion model comes at a computational price. The Lander-Green-Kruglyak algorithm is no longer viable in linkage analysis, and the imposition of a biallelic left boundary locus makes the Elston-Stewart algorithm less efficient, especially in the presence of large numbers of untyped people. Thus, users should be cautious when selecting a marker panel for analysis. Rare alleles should be combined, and markers should be chosen on the basis of their ability to illuminate dubious recombination events. We recommend SNP markers to alleviate computational burdens and to achieve dense coverage. Mendel has the capacity to combine very close SNPs into supermarkers. Markers closely flanking inversion boundaries are helpful. Finally, any tactic that reduces genotyping errors should be fully exploited. Such errors corrupt genetic maps and create serious distractions in discerning inversions.[21]

Acknowledgments

We thank Janet Sinsheimer for checking our mathematical derivations, Angela Presson for providing guidance on using MicroMerge, and Harry Hahn for his help in plotting location scores. This research was supported in part by National Institutes of Health grants GM068875 (to R.A.O.), GM53275 (to K.L.), MH59490 (to K.L.), and National Alliance for Research on Schizophrenia and Depression (NARSAD) grant T32 HG02536 (to G.K.C.). R.A.O. gratefully acknowledges his NARSAD Young Investigator Award.

Appendix A

Convergence of Haplotype Frequencies

Since we have fully described single-marker dynamics, we focus here on haplotype dynamics. For notational convenience, let us define the haplotype and allele frequency column vectors $p = (p_{ab|i}, p_{ab|n})^t$ and $r = (p_{a|i}, p_{a|n}, p_{b|i}, p_{b|n})^t$.

At each generation *m,* we update *p* by

$$p^{m+1} = f(p^m) + g(r^m) = Mp^m + g(r^m) \ , \tag{A1}$$

where $f(p) = Mp$ is linear in *p* and $g(r)$ is quadratic in *r.* If $g(r)$ were a linear function of *r,* then our convergence theory would be simple. As things stand, we exploit the fact that r^m converges to its equilibrium value r^∞ at a known geometric rate.

The magnitude $|\lambda|$ of the dominant eigenvalue λ of the matrix $M = (m_{ij})$ is one of the keys to understanding the dynamical system (A1). This magnitude is called "the spectral radius of *M*" and determines whether *M* is contractive. When one marker flanks the inversion and the other falls inside it, the matrix *M* that appears in recurrence (8) is diagonal. Its eigenvalues therefore coincide with its diagonal entries. Each diagonal entry is a convex combination of two numbers from the open interval (0,1) and therefore belongs to this interval as well. It follows that M has spectral norm strictly <1 and is contractive.

Proving that M is contractive for the cases displayed in figure 3 is more complicated, because M is no longer diagonal. Fortunately, we can bound the spectral radius above by any induced matrix norm of M (proposition 6.3.2 of "Numerical Analysis for Statisticians"[37]). One of the simplest induced matrix norms to apply is the ℓ_∞ norm $\| M \| = \max_i \sum_j |m_{ij}|$. We will show that $\| M \| < 1$ by writing $M = qA + (1-q)B$ and by exploiting the fact that $\| M \| \leqslant q \| A \| + (1-q) \| B \|$. When the markers flank the same side of the inversion, then the representation (11) entails

$$\| A \| = \| B \| = \max\left[1-\theta_1, (1-\theta_1)\theta_2 + (1-\theta_1)(1-\theta_2)\right] = 1-\theta_1 \ ,$$

which clearly disposes of this case.

When the markers flank opposite sides of the inversion, then the representation (10) yields

$$\| A \| = \| B \| = \max\left[1-\theta_{12}, \theta_1\theta_3 + (1-\theta_1)(1-\theta_3)\right] \ .$$

Since $1-\theta_{12}$ belongs to (0,1), we focus on the second term entering the maximum. Now the inequality

$$\theta_1\theta_3 + (1-\theta_1)(1-\theta_3) < 1$$

is equivalent to the inequality

$$\theta_1\theta_3 < \frac{\theta_1+\theta_3}{2} \ .$$

But this latter inequality is a consequence of the arithmetic-geometric mean inequality

$$\sqrt{\theta_1\theta_3} \leqslant \frac{\theta_1+\theta_3}{2}$$

and the inequality $\theta_1\theta_3 < \sqrt{\theta_1\theta_3}$.

The solution to the dynamical system (A1) turns out to be

$$p^m = M^m p^0 + \sum_{k=0}^{m-1} M^k g(r^{m-k-1}) \ . \tag{A2}$$

This representation is true, by definition, when $m = 1$ and can be verified in general by mathematical induction. If we let m tend to infinity and assume that all limit operations are valid, then the long-run equilibrium

$$p^\infty = \sum_{k=0}^{\infty} M^k g(r^\infty) \tag{A3}$$

emerges. Our final objective is to give a rigorous proof of this result, with explicit bounds on the rate of convergence to equilibrium.

Because we have already introduced the ℓ_∞ matrix norm, we will use the ℓ_∞ compatible vector norm

$$\| v \| = \max_i |v_i|$$

in our proof. At a crucial stage in our argument, we will need to use the fact that the quadratic function $g(r)$ satisfies a Lipschitz condition $\| g(r) - g(s) \| \leqslant c \| r - s \|$ for all r and s. The constant c can be derived from the multivariate mean-value inequality

$$\| g(r) - g(s) \| \leqslant \int_0^1 \| dg[tr + (1-t)s] \| \, dt \, \| r - s \| \ ,$$

where $dg(u)$ is the Jacobian of the function $g(u)$. Because the entries of $dg(u)$ are linear functions of u, the supremum of the norm $\| dg(u) \|$ is attained on its compact (closed and bounded) domain. This supremum serves as c.

We will also need a bound on the deviation $\| r^m - r^\infty \|$ of r^m from r^∞. The identities (eqs. [1] and [5]) show that $r^m -$

$r^\infty = Q^m(r^0 - r^\infty)$ for a matrix Q with ℓ_∞ norm $\| Q \| < 1$. The bound $\| r^m - r^\infty \| \leqslant \| Q^m \| \| r^0 - r^\infty \|$ follows directly from the matrix norm property $\| Q^m \| \leqslant \| Q^m \|$.

These preliminaries put us in position to bound the norm of the difference between equations (A2) and (A3) by

$$\begin{aligned}
\| p^m - p^\infty \| &\leqslant \| M^m \| \| p^0 \| + \sum_{k=0}^{m-1} \| M^k \| \| g(r^{m-k-1}) - g(r^\infty) \| + \sum_{k=m}^{\infty} \| M^k \| \| g(r^\infty) \| \\
&\leqslant \| M^m \| \left[\| p^0 \| + \frac{\| g(r^\infty) \|}{1 - \| M \|} \right] + \sum_{k=0}^{m-1} \| M^k \| c \| r^{m-k-1} - r^\infty \| \\
&\leqslant \| M^m \| \left[\| p^0 \| + \frac{\| g(r^\infty) \|}{1 - \| M \|} \right] + \sum_{k=0}^{m-1} \| M^k \| c \| Q^{m-k-1} \| \| r^0 - r^\infty \| \\
&\leqslant \| M^m \| \left[\| p^0 \| + \frac{\| g(r^\infty) \|}{1 - \| M \|} \right] + \sum_{k=0}^{m-1} c d^m \| r^0 - r^\infty \| \\
&\leqslant \| M^m \| \left[\| p^0 \| + \frac{\| g(r^\infty) \|}{1 - \| M \|} \right] + c m d^m \| r^0 - r^\infty \| \quad ,
\end{aligned}$$

where $d = \max\{\| M \| , \| Q \|\}$. Because $md^m = O(t^m)$ for any $t > d$, it follows that $\| p^m - p^\infty \|$ is also $O(t^m)$. In other words, p^m approaches p^∞ at geometric rate t regardless of the starting values p^0 and r^0.

The same local conclusions can be reached by linearizing the iteration map $h(r,p)$ around the equilibrium point (r^∞,p^∞). Because the differential

$$dh(r,p) = \begin{bmatrix} Q & 0 \\ dg(r) & M \end{bmatrix}$$

of $h(r,p)$ is block lower triangular, its eigenvalues coincide with the eigenvalues of the blocks Q and M. This forces the spectral radius of $dh(r^\infty,p^\infty)$ to be the maximum of the spectral radii of Q and M. For example, when the markers flank opposite sides of the inversion as in figure 3*A*, the spectral radius of Q is $\max\{1-\theta_1, 1-\theta_3\}$. The spectral radius of M is the larger root of a complicated quadratic in θ_1, θ_2, and θ_3. It seems obvious that the spectral radius of $dh(r^\infty,p^\infty)$ should exceed $1 - \theta_{12}$, the rate of convergence to linkage equilibrium in the absence of the inversion. Our numerical example confirms this suspicion.

Appendix B

Estimation of Allele Frequencies

We now describe an EM algorithm for estimating allele frequencies at an autosomal locus from a random sample of individuals. For the sake of variety, we will derive the EM algorithm from the perspective of the MM algorithm, a generalization of the EM algorithm that typically uses convexity rather than missing data to construct a minorization of the log-likelihood of the observed data.[38,39] The MM algorithm alternates minorization (first M) with maximization (second M) of the minorizing function. Each iteration of the MM algorithm drives the log-likelihood uphill. Convergence is declared when the log-likelihood stabilizes.

Each observed genotype contributes an additive term to the log-likelihood. For example according to table 2, a homozygous individual *a*/*a* of unknown inversion status contributes the term

$$\ln\left[(1-q)p_{a|n} + qp_{a|i}\right]^2 = 2\ln\left[(1-q)p_{a|n} + qp_{a|i}\right] \qquad \text{(B1)}$$

to the log-likelihood. Unfortunately, the function (B1) involves a convex combination of $p_{a|n}$ and $p_{a|i}$ under the logarithm sign. If these contributions were separated, then maximization would be easy. To devise an MM algorithm that separates the contributions, we use the concavity of the function $f(u) = \ln u$. Concavity entails the inequality

$$\ln(w_1u_1 + w_2u_2) \geqslant w_1 \ln(u_1) + w_2 \ln(u_2)$$

for every combination of weights w_1 and $w_2 = 1 - w_1$ chosen from [0,1] and positive arguments u_1 and u_2.

In particular, if $p^m_{a|n}$ and $p^m_{a|i}$ are the current estimates of $p_{a|n}$ and $p_{a|i}$, then we have

$$\begin{aligned}\ln[(1-q)p_{a|n}+qp_{a|i}] &\geqslant \frac{(1-q)p^m_{a|n}}{(1-q)p^m_{a|n}+qp^m_{a|i}}\ln\left[\frac{(1-q)p^m_{a|n}+qp^m_{a|i}}{p^m_{a|n}}p_{a|n}\right]\\ &\quad+\frac{qp^m_{a|i}}{(1-q)p^m_{a|n}+qp^m_{a|i}}\ln\left[\frac{(1-q)p^m_{a|n}+qp^m_{a|i}}{p^m_{a|i}}p_{a|i}\right]\\ &= w_1\ln p_{a|n}+w_2\ln p_{a|i}+c \qquad \text{(B2)}\end{aligned}$$

in obvious notation. Equality occurs in inequality (B2) when $p_{a|n}=p^m_{a|n}$ and $p_{a|i}=p^m_{a|i}$. This construction furnishes us with a function, $w_1\ln p_{a|n}+w_2\ln p_{a|i}+c$, that is tangent to the function $\ln\left[(1-q)p_{a|n}+qp_{a|i}\right]$ and lies below it for all values of $p_{a|n}$ and $p_{a|i}$. This is the essence of minorization. Similar considerations apply to the heterozygous genotype *a/b,* because

$$\begin{aligned}&\ln\left\{2\left[(1-q)p_{a|n}+qp_{a|i}\right]\left[(1-q)p_{b|n}+qp_{b|i}\right]\right\}\\ &\quad=\ln 2+\ln\left[(1-q)p_{a|n}+qp_{a|i}\right]+\ln\left[(1-q)p_{b|n}+qp_{b|i}\right]\ .\end{aligned}$$

When inversion status is known, most of the likelihoods split under the application of logarithm. The heterozygous genotype *a/b* with one normal and one inverted chromosome is the sole exception, but it can be handled by the minorization

$$\begin{aligned}\ln[p_{a|n}p_{b|i}+p_{b|n}p_{a|i}] &\geqslant \frac{p^m_{a|n}p^m_{b|i}}{p^m_{a|n}p^m_{b|i}+p^m_{b|n}p^m_{a|i}}\ln\left[\frac{p^m_{a|n}p^m_{b|i}+p^m_{b|n}p^m_{a|i}}{p^m_{a|n}p^m_{b|i}}p_{a|n}p_{b|i}\right]\\ &\quad+\frac{p^m_{b|n}p^m_{a|i}}{p^m_{a|n}p^m_{b|i}+p^m_{b|n}p^m_{a|i}}\ln\left[\frac{p^m_{a|n}p^m_{b|i}+p^m_{b|n}p^m_{a|i}}{p^m_{b|n}p^m_{a|i}}p_{b|n}p_{a|i}\right]\\ &= w_1\ln p_{a|n}+w_1\ln p_{b|i}+w_2\ln p_{b|n}+w_2\ln p_{a|i}+c\ ,\end{aligned}$$

where, again, equality occurs when $p_{a|n}=p^m_{a|n}$, $p_{b|n}=p^m_{b|n}$, $p_{a|i}=p^m_{a|i}$, and $p_{b|i}=p^m_{b|i}$.

Because minorization is preserved under summation, we can minorize the log-likelihood $L(p)$ of all observations by a function of the form

$$h(p\mid p^m)=\sum_j d_j\ln p_{j|n}+\sum_j e_j\ln p_{j|i}$$

for positive constants d_j and e_j that depend on the current parameter estimates $p^m_{j|n}$ and $p^m_{j|i}$. Maximization of $L(p)$ subject to the constraints $\sum_j p_{j|n}=1$ and $\sum_j p_{j|i}=1$ is a standard exercise in Lagrange multipliers. The solution

$$p^{m+1}_{j|n}=\frac{d_j}{\sum\limits_k d_k}\quad\text{and}\quad p^{m+1}_{j|i}=\frac{e_j}{\sum\limits_k e_k}$$

defines the next iterate of the MM algorithm. We leave it to the reader to show that this is exactly the EM algorithm derived when the data are viewed as a sequence of hidden multinomial trials.[40] The EM algorithm is straightforward to implement with careful bookkeeping. Extension of the algorithm to noncodominant markers is possible by the methods sketched above, but the details are more complicated.

Web Resources

The URLs for data presented herein are as follows:

Autism Genetics Resource Exchange, http://www.agre.org/ (for genotype and pedigree data of families ascertained for autism)

Marshfield Clinic Research Foundation, http://www.marshfieldclinic.org/research/pages/index.aspx (for genotype and pedigree data for the CEPH families)

UCLA Human Genetics, http://www.genetics.ucla.edu/software/ (for the latest version of Mendel)

References

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528
2. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951
3. Mehan MR, Freimer NB, Ophoff RA (2004) A genome-wide

survey of segmental duplications that mediate common human genetic variation of chromosomal architecture. Hum Genomics 1:335–344

4. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, et al (2005) A common inversion under selection in Europeans. Nat Genet 37:129–137
5. Broman K, Matsumoto N, Giglio S, Martin C, Roseberry J, Zuffardi O, Ledbetter D, Weber JL (2003) Common long human inversion polymorphism on chromosome 8p. In: Goldstein DR (ed) Science and statistics: a festschrift for Terry Speed. Institute of Mathematical Statistics Lecture Notes Monograph Series. Vol 40. Institute of Mathematical Statistics, Bethesda, pp 237–245
6. Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14:417–422
7. Shaw CJ, Lupski JR (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet Spec 13:57–64
8. Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. Trends Genet 18:74–82
9. Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A, Stumm M, Tonnies H, Ventura M, Zollino M, Neri G, Barber J, Wieczorek D, Rocchi M, Zuffardi O (2002) Heterozygous submicroscopic inversions involving olfactory receptor gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. Am J Hum Genet 71: 276–285
10. Ji Y, Eichler EE, Schwartz S, Nicholls RD (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. Genome Res 10:597–610
11. Lakich D, Kazazian HHJ, Antonarakis SE, Gitschier J (1993) Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet 5:236–241
12. Soucie JM, Evatt B, Jackson D, Hemophilia Surveillance System Project Investigators (1998) Occurrence of hemophilia in the United States. Am J Hematol 59:288–294
13. Blennow E, Nielsen KB, Telenius H, Carter NP, Kristoffersson U, Holmberg E, Gillberg C, Nordenskjold M (1995) Fifty probands with extra structurally abnormal chromosomes characterized by fluorescence in situ hybridization. Am J Med Genet 55:85–94
14. Nicholls RD, Saitoh S, Horsthemke B (1998) Imprinting in Prader-Willi and Angelman syndromes. Trends Genet 14:194–200
15. Bayés M, Magano LF, Rivera N, Flores R, Pérez Jurado LA (2003) Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet 73:131–151
16. Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC, Scherer SW (2001) A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet 29:321–325
17. Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tonnesen T, Carlberg BM, Pettersson U (1995) Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. Hum Mol Genet 4:615–621
18. Kaiser P (1984) Pericentric inversions: problems and significance for clinical genetics. Hum Genet 68:1–47
19. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, Weber JL, Ledbetter DH, Zuffardi O (2001) Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 68:874–883
20. Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet 1:e56
21. Jorgenson E, Tang H, Gadde M, Province M, Leppert M, Kardia S, Schork N, Cooper R, Rao DC, Boerwinkle E, Risch N (2005) Ethnicity and human genetic linkage maps. Am J Hum Genet 76:276–290
22. Hicks AA, Petursson H, Jonsson T, Stefansson H, Johannsdottir HS, Sainz J, Frigge ML, Kong A, Gulcher JR, Stefansson K, Sveinbjörnsdottir S (2002) A susceptibility gene for late-onset idiopathic Parkinson's disease. Ann Neurol 52:549–555
23. Gretarsdottir S, Sveinbjörnsdottir S, Jonsson HH, Jakobsson F, Einarsdottir E, Agnarsson U, Shkolny D, et al (2002) Localization of a susceptibility gene for common forms of stroke to 5q12. Am J Hum Genet 70:593–603
24. Kristjansson K, Manolescu A, Kristinsson A, Hardarson T, Knudsen H, Ingason S, Thorleifsson G, Frigge ML, Kong A, Gulcher JR, Stefansson K (2002) Linkage of essential hypertension to chromosome 18q. Hypertension 39:1044–1049
25. Muller HJ (1918) Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. Genetics 3:422–499
26. Crow J, Kimura M (1970) An introduction to population genetics. Harper and Row, New York
27. Trow A (1913) Forms of reproduction: primary and secondary. J Genet 2:313–324
28. Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597
29. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367
30. Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519–527
31. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
32. Elston R, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542
33. Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25
34. Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ, AGRE Steering Committee (2001) The Autism Genetic Resource Exchange: a resource for the study of autism and related neuropsychiatric conditions. Am J Hum Genet 69:463–466
35. Presson A, Sobel E, Lange K, Papp J (2006) Merging microsatellite data. J Comput Biol (in press)
36. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247
37. Lange K (1999) Numerical analysis for statisticians. Springer Verlag, New York
38. Hunter D, Lange K (2004) A tutorial on MM algorithms. Am Statistician 58:30–37
39. McLachlan G, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York
40. Lange K (2002) Mathematical and statistical methods for genetic analysis. Springer Verlag, New York