Spline Methods for the Comparison of Physical and Genetic Maps

Natalia Berloff[1]
Markus Perola[2]
Kenneth Lange[3]

Departments of Biomathematics[3], Human Genetics[2,3], Mathematics[1], and Statistics[3]
University of California
Los Angeles, CA 90095

Address all correspondence to:
Kenneth Lange
Department of Biomathematics
School of Medicine
University of California
Los Angeles, CA 90095-1766
e-mail: klange@ucla.edu
phone: 310-206-8076

Research supported in part by
the American Heart Association[2] and
USPHS grant GM 53275[3].

July 12, 2001

Abstract

The first genetic maps were constructed by linkage analysis. Physical mapping techniques such as radiation hybrids and complete sequencing produce a different picture. For the purposes of population genetics, clinical genetics, and genetic epidemiology, it is helpful to harmonize and amalgamate existing genetic and physical maps. The current paper presents methods for estimating recombination intensity as a function of physical distance along a chromosome. Genetic distance is the integral of intensity. We derive fast reliable estimation algorithms based on a Poisson process model, penalized likelihoods, and cubic spline interpolation. We then apply these algorithms to published recombination data on CEPH families and the complete sequences of chromosomes 21 and 22. Our results are in good agreement with previous studies and the biological data. When final drafts of the other human chromosomes become available, we intend to analyze these chromosomes as well.

Key words and phrases: Genetic recombination, Poisson process, interpolation, penalized likelihood, Newton's method

Short title: Comparison of Physical and Genetic Maps

Introduction

Despite the availability of whole genome sequences, linkage mapping has not lost it relevance. For disease genes, there is little choice but to fall back on either linkage analysis or association testing. Both methods depend on meiotic recombination and are adversely affected by inaccuracies in marker order and spacing. Unfortunately, crossover events are distributed unevenly among and along the human chromosomes. This creates nonlinearities in converting physical distance, measured in base pairs, into genetic distance, measured in expected number of crossovers per gamete (Petes, 2001). There are also sizeable differences in recombination intensities between females and males. A more complete understanding of the relationship between physical distance and genetic distance will provide better parameters for linkage analysis and better estimates of the number of SNPs (single nucleotide polymorphisms) needed for association studies.

The completion of a first draft of the human genome sequence (Lander et al., 2001) has provided geneticists with an unprecedented opportunity to compare physical and genetic distance in humans. Recently for example, Yu et al. (2001) have estimated local recombination intensities ranging from 0 to at least 9 centiMorgans per megabase using the human draft data. This bears out earlier studies suggesting large fluctuations in recombination intensities (Broman et al., 1998). Yu et al. (2001) call regions of high recombination "jungles" and regions of low recombination "deserts." Linkage disequilibrium extends over longer physical distances in deserts.

In the current paper, we tackle the problem of estimating the local intensity of crossing over from available linkage data. One can formalize the problem in the framework of Poisson process and derive fast, reliable estimation algorithms based on splines (Snyder, 1975; Gu and Qiu, 1993; Kingman, 1993). The techniques, both

statistical and numerical, borrow heavily from probability density estimation and interpolation theory. Intensity estimation like density estimation seeks to balance function accuracy against function smoothness. To illustrate the algorithms in action, we consider data on human chromosomes 21 and 22, the two chromosomes with the most complete physical maps.

Poisson Process Model

In Haldane's model of recombination, crossovers occur according to a Poisson process along each chromosome (Lange, 1997). Although this model neglects the phenomenon of chiasma interference, we adopt it here for data analysis. One justification for doing so is that we actually analyze pooled outcomes from many different gametes. It is well known that the superposition of many independent point processes of comparable intensities yields a combined point process that is approximately Poisson (Kingman, 1993). The intensity or rate function $f(x)$ for a Poisson process defined on a region R of a Euclidean space provides the mean number of points occurring in each subregion $S \subset R$ through the integral $\int_S f(x)dx$. In our case, R is an interval $[a, b]$, and the data consist of the random number N of crossovers on $[a, b]$ and their random locations $X_1, \ldots, X_N$. To calculate the density function of the data, we note that N has Poisson density with mean $\lambda = \int_a^b f(x)dx$ and that $X_1, \ldots, X_N$ are independently distributed over $[a, b]$ with common density $\lambda^{-1} f(x)$ (Kingman, 1993). These considerations yield the formula

$$e^{-\int_a^b f(x)\,dx} \frac{\left[\int_a^b f(x)dx\right]^n}{n!} \prod_{i=1}^n \frac{f(x_i)}{\int_a^b f(x)\,dx} \;=\; e^{-\int_a^b f(x)\,dx} \frac{1}{n!} \prod_{i=1}^n f(x_i) \qquad (1)$$

for the joint density at the realization $N = n$ and $X_1 = x_1, \ldots, X_n = x_n$ (Cox and Hinkley, 1974). If we replace $x_1, \ldots, x_m$ by their order statistics, then the $n!$ divisor

drops out of formula (1). Taking logarithms and abbreviating $y_i = f(x_i)$ give the loglikelihood

$$L_1 = \sum_{i=1}^{n} \ln y_i - \int_a^b f(x)\, dx.$$

To control the roughness of $f(x)$, we modify the loglikelihood by adding a penalty term proportional to

$$J = \int_a^b f''(x)^2\, dx, \tag{2}$$

the approximate average curvature of $f(x)$. With these adjustments, our problem becomes one of maximizing the penalized loglikelihood

$$L_2 = \sum_{i=1}^{n} \ln y_i - \int_a^b f(x)\, dx - \epsilon J \tag{3}$$

for $\epsilon > 0$ small.

We have intentionally omitted the dependence of J and L_2 in equations (2) and (3) on parametric arguments. In the context of natural cubic splines (Kincaid and Cheney, 1996), the ultimate parameters are the values y_i. This fact will emerge from our subsequent analysis. Besides worrying about the smoothness of $f(x)$, we also need to be concerned that $f(x) \geq 0$. Previous efforts at estimating nonhomogeneous Poisson intensities have used exponential splines or squared splines to finesse the nonnegativity constraint (Gu and Qiu, 1993; Wahba, 1975). Unfortunately, both the exponential and square functions exaggerate fluctuations and almost guarantee a high curvature for the estimated intensity function. We prefer to eliminate negative excursions of $f(x)$ between observed points (or interior knots) by increasing the penalty coefficient ϵ. Note that the presence of the $\ln y_i$ terms in our definition of the objective function L_2 forces positivity at these knots, and if average curvature is sufficiently constrained, then negative excursions become nearly impossible. Between the boundary knots a and b and their nearest neighbors x_1 and x_n, this

argument fails, and we are forced to amend the objective function further by adding partial pseudo-observations at a and b. Thus, our final formulation of the objective function is

$$L = \delta(\ln y_0 + \ln y_{n+1}) + \sum_{i=1}^{n} \ln y_i - \int_a^b f(x)\,dx - \epsilon J \tag{4}$$

for $\delta > 0$ very small and all $y_i > 0$.

Our approach to estimating $f(x)$ is closest to that of Nichols et al. (1999), who use cubic B-splines to estimate the dynamic changes in tracer density in list-mode positron emission tomography. We opt for natural cubic splines since they exactly minimize the roughness penalty (2) in interpolation problems. Our introduction of pseudo-observations at the boundary knots is less arbitrary than it seems. It not only promotes the nonnegativity of $f(x)$, but it also renders L strictly concave and guarantees a unique maximum point.

Penalized Maximum Likelihood Estimation

We now derive an algorithm for maximizing L and estimating the parameters y_i. For notational simplicity, set $m - 1 = n$, $x_0 = a$, $x_m = b$, $h_i = x_i - x_{i-1}$ for $1 \leq i \leq m$, and $f''(x_i) = z_i$ for $0 \leq i \leq m$. The function $f(x)$ is required to be twice continuously differentiable and to reduce to a cubic polynomial $f_i(x)$ on each subinterval $[x_{i-1}, x_i]$. Given its values z_{i-1} and z_i at the knots x_{i-1} and x_i, the linear curve $f_i''(x)$ satisfies

$$f_i''(x) = \frac{z_{i-1}}{h_i}(x_i - x) + \frac{z_i}{h_i}(x - x_{i-1}).$$

If this equation is integrated twice and the interpolation conditions $f_i(x_{i-1}) = y_{i-1}$ and $f_i(x_i) = y_i$ are imposed, then we get the cubic spline

$$f_i(x) = \frac{z_{i-1}}{6h_i}(x_i - x)^3 + \frac{z_i}{6h_i}(x - x_{i-1})^3$$

$$+\left(\frac{y_i}{h_i}-\frac{z_i h_i}{6}\right)(x-x_{i-1})+\left(\frac{y_{i-1}}{h_i}-\frac{z_{i-1}h_i}{6}\right)(x_i-x).$$

Integrating once again produces the expression

$$\int_a^b f(x)\,dx \;=\; \sum_{i=1}^m \int_{x_{i-1}}^{x_i} f_i(x)\,ds \;=\; \frac{1}{2}\sum_{i=1}^m (y_i+y_{i-1})h_i - \frac{1}{24}\sum_{i=1}^m (z_i+z_{i-1})h_i^3. \quad (5)$$

To write the objective function L in matrix notation, we define the basis functions

$$l_i(x) \;=\; \begin{cases} 0 & \text{for } x < x_{i-1} \\ \frac{x-x_{i-1}}{h_i} & \text{for } x_{i-1} \le x < x_i \\ \frac{x_{i+1}-x}{h_{i+1}} & \text{for } x_i \le x < x_{i+1} \\ 0 & \text{for } x_{i+1} \le x \end{cases}$$

for $0 \le i \le m$. Here we take $h_0 = 1$, $h_{m+1} = 1$, $x_{-1} = a-1$, and $x_{m+1} = b+1$. With these definitions, it is straightforward to check that

$$f''(x) \;=\; z_0 l_0(x) + z_1 l_1(x) + \cdots + z_{m-1} l_{m-1}(x) + z_m l_m(x).$$

Given the vectors $\mathbf{y} = [y_0, \ldots, y_m]^t$, $\mathbf{z} = [z_0, \ldots, z_m]^t$, and $\mathbf{l}(x) = [l_0(x), \ldots, l_m(x)]^t$, it follows that $f''(x) = \mathbf{z}^t \mathbf{l}(x)$ and that

$$J(\mathbf{y}) \;=\; \int_a^b f''(x)^2\,dx \;=\; \int_a^b \mathbf{z}^t \mathbf{l}(x)\mathbf{l}(x)^t \mathbf{z}\,dx. \quad (6)$$

In dealing with cubic splines, the standard procedure is to eliminate the z_i by invoking the continuity of $f'(x)$ at the knots. The corresponding $m-1$ matching conditions can be summarized in vector form as $G\mathbf{z} = H\mathbf{y}$ by introducing the two $(m-1)\times(m+1)$ matrices

$$G \;=\; \begin{pmatrix} \frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{h_{m-1}}{6} & \frac{h_{m-1}+h_m}{3} & \frac{h_m}{6} \end{pmatrix}$$

and

$$H \;=\; \begin{pmatrix} \frac{1}{h_1} & -\frac{1}{h_1}-\frac{1}{h_2} & \frac{1}{h_2} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{h_2} & -\frac{1}{h_2}-\frac{1}{h_3} & \frac{1}{h_3} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{h_{m-1}} & -\frac{1}{h_{m-1}}-\frac{1}{h_m} & \frac{1}{h_m} \end{pmatrix}.$$

A natural cubic spline satisfies $z_0 = z_m = 0$, so we drop the first and last columns of G to obtain an $(m-1) \times (m-1)$ symmetric matrix $\bar{G}$ satisfying $\bar{G}\mathbf{p} = H\mathbf{y}$, where $\mathbf{p} = [z_1, \cdots, z_{m-1}]^t$. Because $\bar{G}$ is diagonally dominant, its inverse exists, and we can solve for $\mathbf{p} = \bar{G}^{-1}H\mathbf{y}$. This shows that $\mathbf{y}$ adequately parameterizes the problem.

Substitution of $\mathbf{p} = \bar{G}^{-1}H\mathbf{y}$ in equation (6) gives

$$\begin{aligned} J(\mathbf{y}) &= \int_a^b (\bar{G}^{-1}H\mathbf{y})^t \, \bar{\mathbf{l}}(x)\bar{\mathbf{l}}(x)^t \, (\bar{G}^{-1}H\mathbf{y}) \, dx \\ &= \mathbf{y}^t H^t \bar{G}^{-1} \int_a^b \bar{\mathbf{l}}(x)\bar{\mathbf{l}}(x)^t \, dx \; \bar{G}^{-1}H\mathbf{y}, \end{aligned}$$

where $\bar{\mathbf{l}}(x) = [l_1(x), \cdots, l_{m-1}(x)]^t$. Elementary calculus demonstrates that entry by entry the two symmetric matrices $\int_a^b \bar{\mathbf{l}}(x)\bar{\mathbf{l}}(x)^t \, dx$ and $\bar{G}$ coincide. Since the first is positive semidefinite and the second invertible, both are positive definite. Furthermore, we have the obvious identity

$$J(\mathbf{y}) = \mathbf{y}^t H^t \bar{G}^{-1} \bar{G} \bar{G}^{-1} H\mathbf{y} = \mathbf{y}^t H^t \bar{G}^{-1} H\mathbf{y}. \tag{7}$$

The remainder of the objective function can be simplified by defining the two $m+1$ vectors

$$\begin{aligned} \ln \mathbf{y} &= [\ln y_0, \cdots, \ln y_m]^t \\ \gamma &= [\delta, 1, \cdots, 1, \delta]^t, \end{aligned}$$

the m vector $\mathbf{1} = [1, \ldots, 1]^t$, the $m \times (m+1)$ matrix

$$A = \frac{1}{2} \begin{pmatrix} h_1 & h_1 & 0 & \cdots & 0 & 0 \\ 0 & h_2 & h_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_m & h_m \end{pmatrix},$$

and the $m \times (m-1)$ matrix

$$B = \frac{1}{24} \begin{pmatrix} h_1^3 & 0 & \cdots & 0 & 0 \\ h_2^3 & h_2^3 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & h_{m-1}^3 & h_{m-1}^3 \\ 0 & 0 & \cdots & 0 & h_m^3 \end{pmatrix}.$$

Substituting identities (5) and (7) in definition (4) now yields

$$L(\mathbf{y}) = \gamma^t \ln \mathbf{y} - \mathbf{1}^t(A - B\bar{G}^{-1}H)\mathbf{y} - \epsilon \mathbf{y}^t H^t \bar{G}^{-1} H \mathbf{y}. \quad (8)$$

With this expression in hand, it is easy to calculate the first differential (row vector of first partial derivatives)

$$dL(\mathbf{y}) = [\delta y_0^{-1}, y_1^{-1}, \ldots, y_{m-1}^{-1}, \delta y_m^{-1}] - \mathbf{1}^t(A - B\bar{G}^{-1}H) - 2\epsilon \mathbf{y}^t H^t \bar{G}^{-1} H \quad (9)$$

and second differential (Hessian matrix of second partial derivatives)

$$d^2L(\mathbf{y}) = -Q - 2\epsilon H^t \bar{G}^{-1} H, \quad (10)$$

where Q is the diagonal matrix constructed with entries $[\delta y_0^{-2}, y_1^{-2}, \ldots, y_{m-1}^{-2}, \delta y_m^{-2}]^t$.

The representation (10) of $d^2L(\mathbf{y})$ as the sum of a negative definite matrix and a negative semidefinite matrix shows that $d^2L(\mathbf{y})$ is negative definite. It follows that $L(\mathbf{y})$ is strictly concave and possess at most one maximum point. The existence of a maximum point in the positive orthant $R_+^{m+1} = \{\mathbf{y} : y_i > 0, i = 0, \ldots, m\}$ can be proved by showing that $L(\mathbf{y})$ tends to $-\infty$ whenever any component y_i of $\mathbf{y}$ tends to either 0 or ∞. With this goal in mind, we first note that the quadratic function $\mathbf{1}^t B\bar{G}^{-1}\mathbf{v} - \epsilon \mathbf{v}^t \bar{G}^{-1}\mathbf{v}$ possesses a maximum ω on R^{m-1} because $\bar{G}^{-1}$ is positive definite. Consequently, the inequality

$$\begin{aligned} L(\mathbf{y}) &\leq \gamma^t \ln \mathbf{y} - \mathbf{1}^t A \mathbf{y} + \omega \\ &= \delta(\ln y_0 + \ln y_m) + \sum_{i=1}^{m-1} \ln y_i - \frac{1}{2}\sum_{i=1}^{m}(y_i + y_{i-1})h_i + \omega \end{aligned}$$

holds. To complete our proof, we simply observe that any function of the form $f(s) = c \ln s - ds$ for positive c and d is bounded and tends to $-\infty$ as s tends to either 0 or ∞.

In practice, $L(\mathbf{y})$ can be maximized by a variety of methods. Newton's method iterates according to

$$\mathbf{y}^{j+1} = \mathbf{y}^j - d^2L(\mathbf{y}^j)^{-1} dL(\mathbf{y}^j)^t$$

and converges at a quadratic rate if started in the vicinity of the maximum point. Because of the possibility of overshoot, Newton's method is not globally convergent. In this problem, it can be made so by adding a line search at each iteration (Luenberger, 1984; Lange, 1999). Quasi-Newton methods substitute an approximate Hessian for the true Hessian (Press et al., 1992; Lange 1999). They typically start with the identity matrix and gradually build up a better approximation using function and gradient values only. All Hessian updates are symmetric and positive definite, thus guaranteeing steps in an uphill direction. Just as with Newton's method, some safeguard must be taken to avoid overshooting the maximum point. Conjugate gradient methods dispense with approximate Hessians altogether and are appropriate for high dimensional problems where excessive demands on computer memory become an issue (Luenberger, 1984). Both quasi-Newton and conjugate gradient methods display superlinear convergence.

Special features of this problem make each of these algorithms extremely efficient. For example, the vector $\mathbf{1}^t(A - B\bar{G}^{-1}H)$ and the matrix $H^t\bar{G}^{-1}H$ need only be computed once. Because the matrix $\bar{G}$ is tridiagonal, computation of $\bar{G}^{-1}$ requires only $O(m)$ operations. The biggest drawback to Newton's method is the inversion of the $(m+1) \times (m+1)$ Hessian $d^2L(\mathbf{y}^j)$ at each iteration. Quasi-Newton methods can be phrased as either updates to the approximate Hessian or to the inverse approximate Hessian. The latter is clearly advantageous from the point of view of operation counts. Quasi-Newton updates can be adversely affected by roundoff errors when the spacings h_i between adjacent recombination events become very small or highly nonuniform. These problems can be handled by rescaling the h_i, by restarting the algorithm at the point of convergence, or by storing and updating the Cholesky decomposition of the approximate Hessian rather than the approximate Hessian itself (Dennis and Schnabel, 1983).

Algorithm Tests

We have tested our algorithms on various intensity functions. Given a choice of $f(x)$, it is easy to simulate observations on the number of points N and their locations $X_1, \ldots, X_N$. Comparison of estimated and true intensity functions can then guide the choice of the tuning constants δ and ϵ. Of the two constants, ϵ is the more critical. In practice, our results are insensitive to the choice of δ, and we simply take $\delta = .0001$. The smoothing constant ϵ controls the tradeoff between the conflicting goals of maximizing the loglikelihood and minimizing the roughness penalty. The loglikelihood contribution to the objective function (8) steers the estimate of $f(x)$ toward the data as the number of independent replicates r increases. With little data, the roughness penalty predominates.

Figure 1 summarizes the results from one simulation where we attempted to match the interval length and number of points ($\sim$ 60). The different plots of Figure 1 for the original and reconstructed intensity functions show the effects of different choice of ϵ. It is clear from the figure and similar numerical experiments that the best agreement between the original and estimated intensities is reached for $\epsilon \in [0.0005, 0.001]$. Observe here that the unit of measurement affects the choice of ϵ. To control roundoff error, all computations use a unit of distance of 10Mb. Comparable simulations were carried out corresponding to the observed data on chromosomes 21 and 22 and lead to similar ranges for ϵ.

Application to Real Data

Eight large, three-generation pedigrees from the CEPH consortium provide ideal data our purposes (web site http://www.cephb.fr/cephdb/) (Dausset et al., 1990).

These pedigrees have been genotyped for literally thousands of markers and carefully checked for inconsistencies. Unfortunately, recombination events cannot be precisely pinpointed from these or other currently available human data. As a substitute for precise localization, we find the shortest interval unambiguously containing each recombination. This can be done via application of a haplotyping program such as Simwalk 2.82 (option 1). Once the flanking interval is determined, a recombination point is randomly and uniformly assigned to it. Although this adds an element of stochasticity to the estimation process, in practice the impact is unnoticeable. A total of 40 markers were chosen for analysis from chromosome 21 and 90 from chromosome 22. Chromosome 21 shows 65 maternal recombinations and 54 paternal recombinations, while chromosome 22 shows 68 maternal recombinations and 64 paternal recombinations. The physical order of the genetic markers was obtained from the Marshfield Center for Medical Genetics (Yu et al., 2001).

Figure 2 displays the genetic maps for females (grey circles) and males (black squares) along chromosome 21. Dots indicate the assigned points of recombinations. Notice that dips in estimated intensity correspond to regions with densely spaced points and peaks to regions with sparsely spaced points. Estimation was done by both Newton's method and quasi-Newton methods and takes seconds on a desktop computer. Figure 3 displays the female and male maps for chromosome 22. Figures 4 and 5 graph the ratio of female to male intensity along each chromosome. The chromosome 21 plot is very similar to the corresponding plot appearing in (Broman et al., 1998). The choices of ϵ varied in these computations from 0.001 (maternal graph for chromosome 21), 0.01 (paternal graph for chromosome 21), 0.00035 (maternal graph for chromosome 22), to 0.0005 (paternal graph for chromosome 22).

Discussion

The complete physical maps of chromosomes 21 (Dunham et al., 1999) and 22 (Hattori et al., 2000) have not yet substantially impacted disease linkage studies. However, the precise ordering of markers and accurate estimates of genetic distances between them can only help in the terribly difficult job of pinpointing complex disease loci. Errors in the order of closely spaced markers are far more common than realized; we encountered several in the data given on the CEPH web site (http://www.cephb.fr/cephdb/). The current paper uses orders dictated by the sequence data.

Estimating genetic distances between closely spaced markers requires large numbers of families and high quality recombination data. One can argue that good methods for converting physical distances into genetic distances are apt to give more reliable estimates than direct measurement of recombination fractions. Our results for chromosomes 21 and 22 are consistent with the best previous data (Broman et al., 1998). This suggests that the Poisson process model and penalized likelihood estimation are reasonably accurate. It is probably wise to defer similar analyses of the remaining chromosomes until their drafts are complete. The large differences between females and males has been observed repeatedly for both humans (Broman et al., 1998) and mice (Lindahl, 1991). In fact, the total map lengths in human females and males are 44 Morgans and 28 Morgans, respectively (Broman et al., 1998). Our analysis is confined to families of European descent. Although there is the faint possibility that ethnic differences in recombination will emerge, we know of no evidence suggesting this as a major concern.

Recombination hot and cold spots are the most probable causes of the nonlinearities between genetic and physical maps. Such phenomena have been noted for all eukaryotic organisms studied (Petes, 2001). On the coarse scale of this study,

it is hard to isolate hot spots. Certainly, we would need better localization of recombination events to capture them. Once the biology of recombination is better understood, we may be able to construct a rational theory predicting variations in the intensity of recombination (Kanaar and Hoejimakers, 1998). Until that day, we will have to rely on empirical data and estimation procedures such as those presented here.

References

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L., and Weber, J.L. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.,* 63, 861-9.

Cox, D.R., and Hinkley, D.V. 1974. *Theoretical Statistics,* Chapman and Hall Ltd, London.

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M., and White, R. 1990. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics,* 6, 575-7

Dennis, J.E., and Schnabel, R.B. 1983 *Numerical Methods for Unconstraned Optimization and Nonlinear Equations,* Prentice-Hall, Englewood Cliffs, NJ.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., et al. 1999. The DNA sequence of human chromosome 22. *Nature,* 402, 489-95.

Gu, C., and Qiu, C. 1993. Smoothing spline density estimation: Theory. *Annals Stat.,* 21, 217-234.

Kanaar, R., and Hoeijmakers J.H.J. 1998. From competition to collaboration. *Nature*, 391, 335-337.

Kincaid, D., and Cheney, W. 1996 *Numerical Analysis,* 2nd edition.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., et al.

2000. The DNA sequence of human chromosome 21. *Nature,* 405, 311-9.

Kingman, J.F.C. 1993. *Poisson Processes,* Clarendon Press, Oxford.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 15, 409, 860-921.

Lange, K. 1997. *Mathematical and Statistical Methods for Genetic Analysis,* Springer-Verlag, New York.

Lange, K. 1999. *Numerical Analysis for Statisticians,* Springer-Verlag, New York.

Luenberger, DG. 1984. *Linear and Nonlinear Programming,* 2nd ed. Addison-Wesley, Reading, MA.

Lindahl K. 1991. His and Hers recombinational hotspots. *Trends Genet.,* 7, 273-276.

Nichols, T.E., Qi, J., and Leahy, R. M. 1999. Continuous time dynamic PET imaging using list mode data, 98-111, in *IPMI'99, LNCS,* ed. A. Kuba *et al.*, Springer-Verlag.

Petes, T.D. 2001. Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.,* 2, 360-9.

Press, W., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1992. *Numerical Recipes in FORTRAN. The Art of Scientific Computing,* 2nd edition, Cambridge university Press.

Sobel, E. and Lange, K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.,* 58, 1323-1337.

Snyder, D.L. 1975. *Random Point Processes,* Wiley, New York.

Wahba, G. 1975. Interpolating spline methods for density estimation. I: Equi-spaced knots. *Annals Stat.,* 3, 30-48.

Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A.J., Deloukas, P., Olsen, A., Doggett,

N.A., Ghebranious, N., Broman, K.W., and Weber, J.L. 2001. Comparison of human genetic and sequence-based physical maps. *Nature,* 409, 951-3.

Figure 1: Reconstruction of the Poisson intensity function $f(x) = ae^{x/10}$ (solid line) using 60 random points and various values of ϵ. The horizontal axis correspond to a chromosome length of 35 Mb. The constant a is chosen so the the area under the curve is 60.

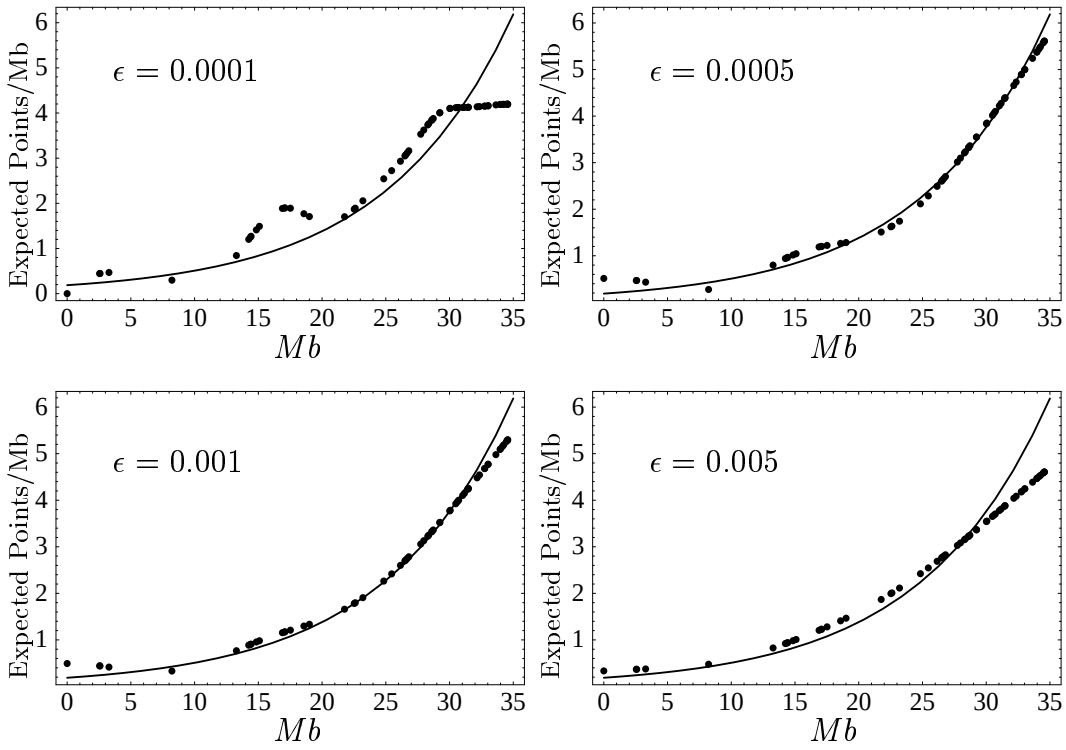

Figure 2: Plot of female (grey circles) and male (black squares) intensities versus genetic distances (in Mb) along chromosome 21.

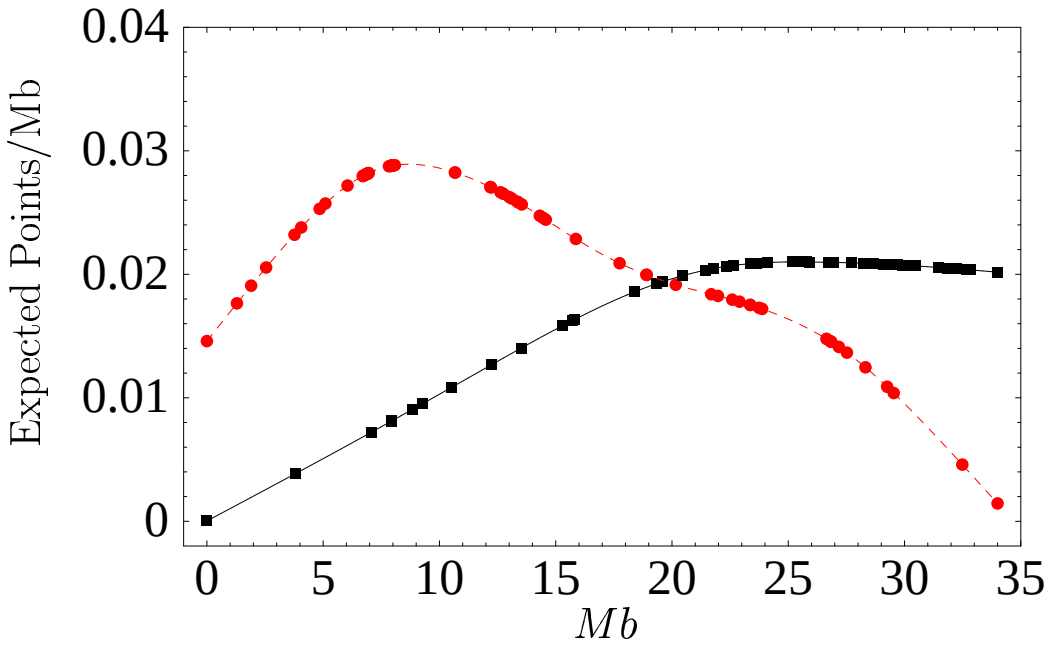

Figure 3: Plot of female (grey circles) and male (black squares) intensities versus genetic distances (in Mb) along chromosome 22.

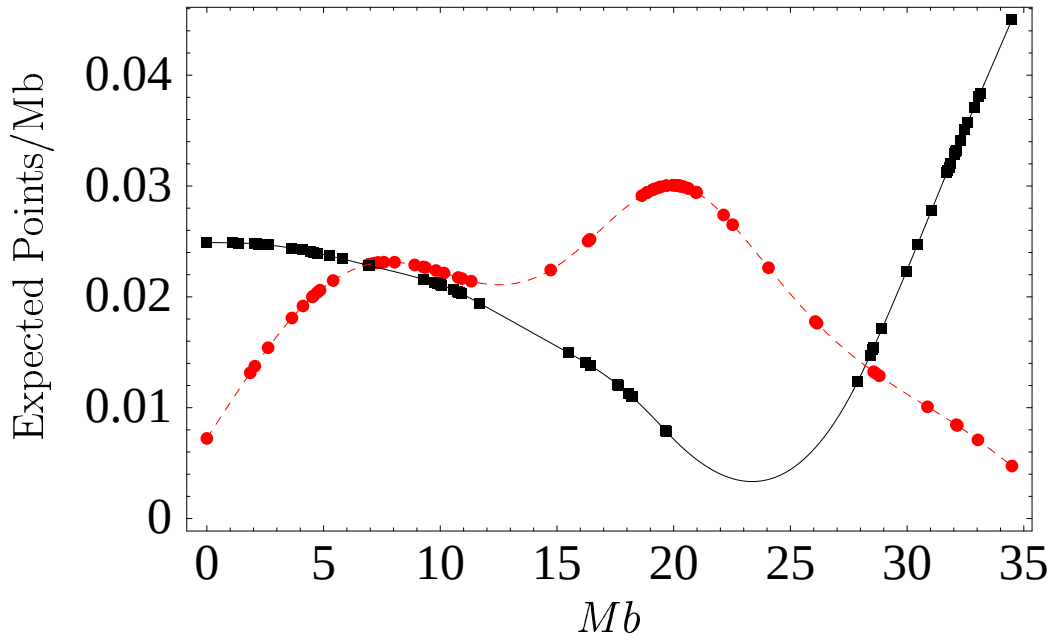

Figure 4: Plot of the female/male intensity ratio versus sex-averaged distance (in Mb) along chromosome 21.

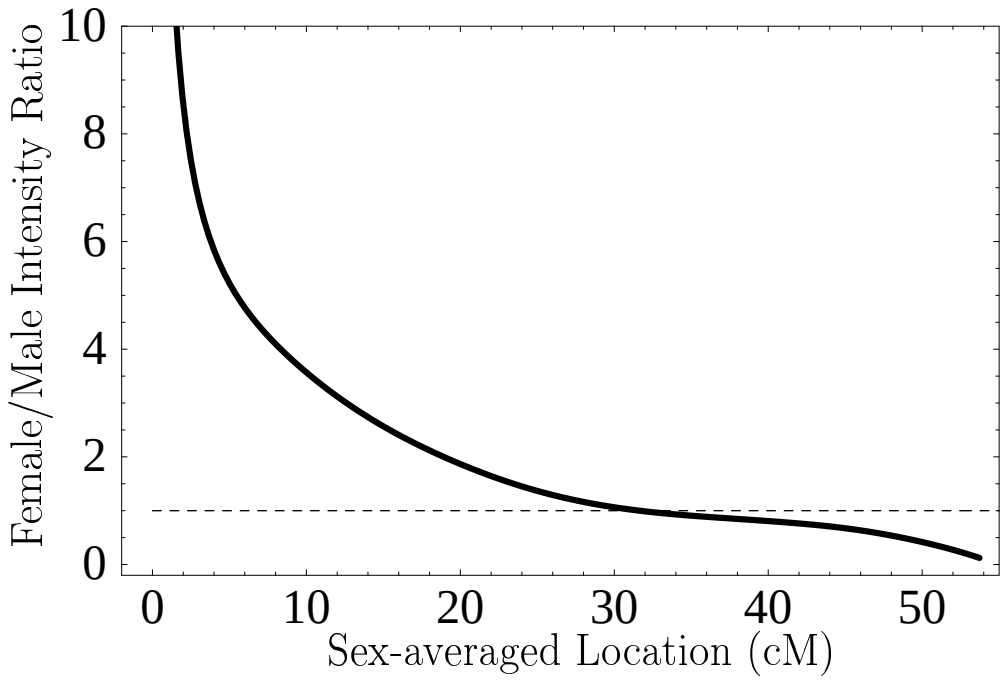
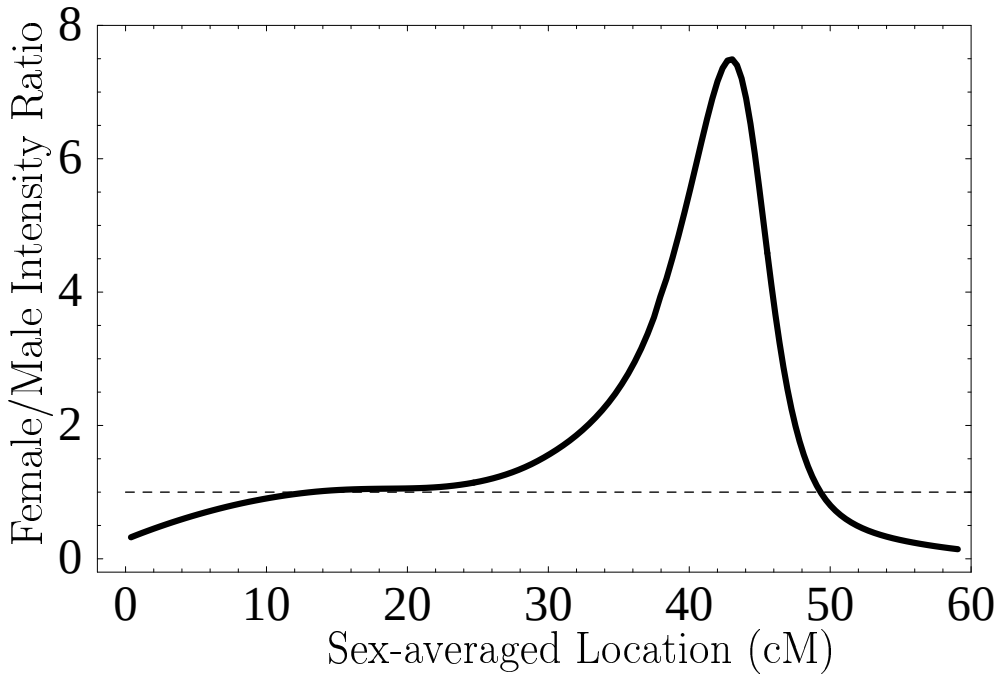
Figure 5: Plot of the female/male intensity ratio versus sex-averaged distance (in Mb) along chromosome 22.