NONPARAMETRIC ESTIMATION OF HETEROGENEOUS RECOMBINATION RATES ACROSS THE MOUSE GENOME

Ying-ming M. Jou, David Ruppert, Cornell University, Ithaca, NY
Gary A. Churchill, The Jackson Laboratory
Gary A. Churchill, The Jackson Laboratory, Bar Harbor, ME 04609

Key Words: Local polynomial kernel estimator, Local polynomial density estimator, Bandwidth, The mouse genome, Genetic map, Physical map, Genetic recombination rates.

Abstract:

This work develops statistical methods for estimating local rates of recombination from genetic mapping data. Heterogeneity of recombination rates has been observed in the genomes of fruitfly species for which detailed physical and genetic maps are available. However, it has proven to be difficult to study recombination rate variation in other species due to the lack of physical mapping data. It has been noted (Lyon, 1970; Nachman and Churchill, 1996) that local recombination rates can be estimated from genetic map data alone. Here we develop a local polynomial kernel regression method to estimate recombination rates across the mouse genome. Simulation results demonstrate that the local polynomial kernel estimator with an appropriately selected bandwidth can recover the true initial recombination rates from the genetic map if markers are uniformly distributed on the physical map. This method was then applied to the Whitehead/MIT mouse genome crossover data. Recombination rates were observed to vary substantially within and between chromosomes. There is evidence for recombinational heterogeneity on most of the 20 mouse chromosomes. The methods developed here provide a useful tool for estimating genetic recombination rates in species for which physical maps are not available.

1. Introduction

Single-celled organisms can reproduce by simple mitotic division (mitosis) which involves chromosome replication, segregation, and division. Such asexual reproduction is simple and direct, but gives rise to offspring that are genetically identical to the parent organism. Sexual reproduction, on the other hand, involves meiosis which produces gametes with only half of the parent organism's chromosome, and fertilization that mixes the genomes from two individuals of opposite sex. An important feature of sexual reproduction is that it produces offspring that differ genetically from one another and from both their parents. Therefore, sexual reproduction has probably been favored by evolution (Maynard Smith, 1978)because it improves the chance of producing at least some offspring that will survive in an unpredictably variable environment. During meiosis, the two replicated homologous chromosomes pair together and crossing-overs may occur which allow the chromosomes to exchange some DNA. With such genetic recombinations, a gamete produced at the end of meiosis receives a totally new assortment of genes on each chromosome, with some from each of paternal and maternal homologues.

The process of genetic exchange can be exploited to locate gene positions on chromosomes. Geneticists have been interested in the pattern and rates of genetic recombinations across the chromosomes. Morgan and his students first reasoned that the greater the distance between two loci (a locus is the site of a gene in the genome), the greater the chance that they will be recombined by crossing-over occurring at a site between them. Using this notion of distance, a map of the relative positions of genes on chromosomes can be produced. If two markers (a marker is a distinct portion of DNA, or a known gene for which the parental origin can be determined) recombined in a proportion r of gametes, they are said to be separated by a genetic map distance of x map units, where $x = -\frac{1}{2}\log(1-2r)$ (Haldane, 1919). If recombination rates were homogeneous across the genome, the genetic map would be in concordance with the physical map. However this is generally not the case.

Many studies have noted heterogeneity in recombination rates across the genomes of *Drosophila melanogaster*(fruit fly) for which both detailed physical and genetic maps exist for the same set of markers. With both maps, Kliman and Hey(1993) estimated recombination rates across the genome of *Drosophila melanogaster* from plots of genetic po-

Funded by NSF grant #DMS 9706863.

sition (in centimorgans, cM) versus physical position (in megabase, Mb) for markers on each chromosome. They fit a least-squares polynomial curve for each chromosome and estimated recombination rate by taking the derivative of the polynomial. Kindahl(1994) estimated recombination rates of *Drosophila melanogaster* by comparing the genetic and physical distance between many pairs of markers over different genomic regions.

Some studies on *Mus musculus* (laboratory mouse) also indicate that recombinations are not uniformly distributted along the physical chromosome. At present, the physical distances between markers on *Mus musculus* chromosomes are incomplete or unknown. As a result, variation in recombination rate across the mouse genome has not been investigated as extensively as in the fruitfly genome. Lyon (1976) proposed that it should be possible to infer variation in recombination rates across chromosomes from a genetic map alone if some conditions are satisfied. Specifically, under the condition that markers are uniformly and randomly located on a physical map, chromosomal regions with low recombination should contain many markers, while regions of high recombination contain few markers. Nachman and Churchill (1996) applied Lyon's theory to estimate recombination rates across the mouse genome. They first estimated the density function of markers for each mouse chromosome by applying a kernel density smoother with a cosine kernel to the histogram of markers along the genetic map, and then estimated the genetic recombination rates by inverting the estimated density function of markers.

Here we will develop a different approach to estimate the local recombination rate across the mouse genome. As in Nachman and Churchill (1996), we will also make the following two assumptions. First, markers are uniformly distributed on the physical maps of chromosomes. Second, for a chromosome of size P in physical map (*i.e.* P Mb in length) and size G in genetic map (*i.e.* G cM), there is a smooth and differentiable monotonely nondecreasing function $\Lambda : [0, P] \to [0, G]$ (P and G may be normalized to be 1) that maps physical locations along the chromosomes onto the genetic locations. The recombination rate is proportional to the derivative, $\lambda = \frac{d}{dt}\Lambda(t)$. In this artical, we will make another assumption that the function Λ has a continuous second derivative. Under these assumptions, Λ can be approximated, up to some constant, by a local quadratic kernel regression whose second coefficient (first derivative) provides an estimate for the local recombination rate (Wand and Jones, 1995).

2. Materials and Methods

2.1 Data

There are several distinct genetic maps for the mouse genome, generated in different ways and maintained in separate databases. The data we used were the intercross genetic map (Dietrich *et al.*, 1992) available on line from the Whitehead/MIT Center for Genome Research. The map consists of 6331 markers of simple sequence length polymorphisms $(CA)_n$ mapped on an Ob×Cast F_2 intercross at an average resolution of 1.1 cM. In those maps, one finds clusters of markers and spacings between clusters (in cM). Two markers will appear in the same cluster when no recombination events have been observed between them. In this case their *estimated* genetic distance is zero. Because the Whitehead/MIT data are based on a small sample of $46 F_2$ mice obtained by intercrossing F_1 progeny, there only are 92 meioses on this map. Thus many of the markers occur in clusters. Note that in those maps only the order of markers and their genetic distances are available, the physical distances are unknown. Now if markers are randomly distributed on the physical map and crossing-over events occur according to a smooth density function, then what we observe is simply an interlacing of two point processes – markers and crossovers – along the chromosomes. We can then count numbers of crossover events between markers. In other words, we have binned data for which the bins are defined by the markers, and the count in each bin is the number of crossovers between the two adjacent markers that defined the bin.

Let X_i be the true but unobservable physical position of marker i ($i = 1, 2, \ldots, M$, where M is the number of markers on a chromosome), and Y_i the cumulative number of recombinations up to marker i (Y_1 is defined to be 0). Under the assumption that markers are uniformly distributed on the physical map, a realization of $(X_1, X_2, \ldots, X_M)$, denoted as $(x_1, x_2, \ldots, x_M)$, would be either the normalized rank order of markers or a sequence of sorted random numbers. The plot of Y_i's versus x_i's gives an empirical estimate of the smooth function Λ up to a constant. The local derivative of the fitted curve obtained by applying local quadratic kernel regression to the data (x_i, Y_i) then provides an estimate of the regional recombination rate. The mathematical theory of local polynomial kernel regression is introduced in the next section.

2.2 Theory of Univariate Local Polynomial Kernel Regression

Suppose our data, $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, satisfy the following relationship

$$Y_i = \Lambda(x_i) + \epsilon_i, \qquad i = 1, 2, \ldots, M \qquad (1)$$

where Y_i's are the response variables (the cumulative recombination counts in our work), x_i's are some fixed and known values of a predictor (the realization of unobserved physical marker positions), Λ is a smooth function with continuous pth derivatives, ϵ_i are random variables with $E(\epsilon_i) = 0$, and M is the number of markers on a chromosome (Wand and Jones 1995). Note that Λ is in proportion to the aforementioned mapping function, and its derivative, λ, is the recombination rate of our interest.

At any arbitrary point x, the estimator for $\Lambda(x)$ is obtained by fitting a polynomial of degree p

$$\beta_0 + \beta_1(\cdot - x) + \beta_2(\cdot - x)^2 + \ldots + \beta_p(\cdot - x)^p \quad (2)$$

to the data using weighted least squares. In this work, we used the weight kernel $K_h(x_i - x)$ which is defined to be $h^{-1}K(\frac{x_i - x}{h})$ where K is the *Epanechnikov* kernel $K(z) = (1 - z^2)_+$. This kernel implies that an observed point (x_i, Y_i) will get zero weight if the distance between x_i and x is greater than h. Therefore, h is usually called the bandwidth. Suppose, at the point x with specified p and h, $\hat{\beta} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)^T$ minimizes the weighted sum of squared errors

$$\sum_{i=1}^{n}\{Y_i - \beta_0 - \beta_1(x_i - x) - \ldots - \beta_p(x_i - x)^p\}^2 K_h(x_i - x). \quad (3)$$

Then, the estimator for $\Lambda(x)$ at the point x is the height of the fitted polynomial, $\hat{\beta}_0$. And the derivative, $\lambda(x)$, is estimated by $\hat{\beta}_1$.

At the point x, let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be the vector of responses,

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & x_1 - x & \ldots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \ldots & (x_n - x)^p \end{bmatrix},$$

and $\boldsymbol{W}_x = diag\{K_h(x_1 - x), \ldots, K_h(x_n - x)\}$. The standard weighted least squares theory leads to the solution

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y},$$

assuming the invertibility of $\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x$. Thus, given a predetermined polynomial degree p and bandwidth h, we have

$$\hat{\Lambda}(x; p, h) = \hat{\beta}_0 = \boldsymbol{e}_1^T \left(\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y}, \quad (4)$$

and

$$\hat{\lambda}(x; p, h) = \hat{\beta}_1 = \boldsymbol{e}_2^T \left(\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y} \quad (5)$$

where $\boldsymbol{e}_i$ is the $(p+1) \times 1$ vector with 1 in the ith entry and zero elsewhere. A smooth local polynomial kernel estimator can be produced by calculating (4) and (5) on a grid of x.

Since only the first two coefficients of the polynomial are of interest, we shall fit a polynomial of degree $p = 2$, as generally suggested by statisticians. To specify the bandwidth h, however, requires a more subtle decision, as discussed in the following section.

2.3 Simulation and Bandwidth Selection

To choose an appropriate bandwidth for our analysis, standard bandwidth selectors such as the cross-validation method are not applicable for our approach due to the fact that cumulative crossover counts are somehow correlated. In this section, simulation is utilized to select an appropriate bandwidth for analyzing real genomic data. This also serves our purposes of investigating whether our approach with an adequate bandwidth can recover the true recombination rate and how the weight kernel bandwidth affects the accuracy of estimation. The simulation algorithm follows. The first step is to generate 500 sorted random numbers on the interval $[0, 1]$, representing the true but unobserved physical positions of markers. The second step is to generate crossover events that follow a nonhomogeneous Poisson process with the true rate

$$\lambda(t) = 1.2 \times e^{-\frac{(x-0.2)^2}{2(0.08)^2}} + 1.5 \times e^{-\frac{(x-0.5)^2}{2(0.08)^2}} + 1.2 \times e^{-\frac{(x-0.8)^2}{2(0.1)^2}}, \qquad (6)$$

and then count numbers of crossovers between markers. Step 3 is to smooth the empirical curve of cumulative recombination numbers versus the normalized rank order of markers by local quadratic kernel regression with a specified bandwidth, and then take the local derivatives. Step 4 repeats steps 2 and 3 to obtain 1000 estimates of the true rate λ and the

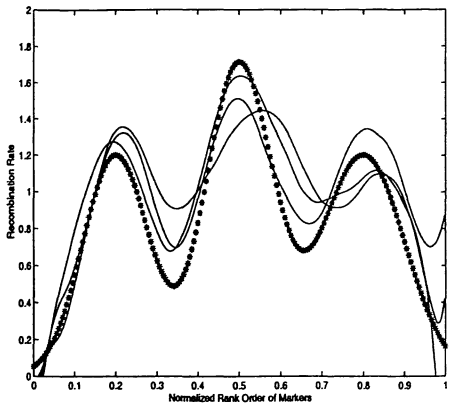

Figure 1: Illustration of Simulation – The thick curve is the true recombination rate λ in (6). Each solid curve represents an estimate of λ. At this specified bandwidth ($h = 1/7$), the simulation is repeated 1000 times to calculate the Monte Carlo MASE and three replicates are shown for illustration.

Monte Carlo mean average squared error (MASE). The simulation is illustrated in Figure 1.

The MASE measures the closeness of an estimator to the true function, and is thus used to evaluate the performance of an estimator. It is calculated with respect to the recombination rate, namely,

$$MASE_h = E\left[\frac{1}{n}\sum_{i=1}^{n}\{\lambda(x_i) - \hat{\lambda}(x_i;h)\}^2\right]. \quad (7)$$

The MASEs are given in Table 1 for two different numbers of markers with different bandwidths within a reasonable range. The bandwidth which provides the smallest MASE is considered appropriate to be employed in our analysis.

Table 1: Simulation results for bandwidth selection

	MASE (1000 reps)	
h	200 markers	500 markers
1/4	0.1293	
1/5	0.1234	0.0914
1/6	**0.1166**	0.0823
1/7	0.1178	**0.0820**
1/8	0.1282	0.0889
1/9		0.0927

It is worth mentioning that when equally-spaced rank orders are used while the true marker positions are not equally-spaced, an observed large crossover count between two markers may reflect two possibilities: a high local recombination rate, or a large physical distance between the two markers. In the latter case, it results in an overestimate of the true regional recombination rate. Likewise, an observed small count may reflect two possibilities: a low local recombination rate, or a short physical distance between the two markers. And in the latter case, it results in an underestimate of the true rate. Such bias, fortunately, can be eliminated by increasing the size of the bandwidth. In general, the local polynomial kernel regression with a small bandwidth can better discover jumps and thus produces a more wiggly curve, but may be too sensitive to the jumps caused by physically unequally-spaced $x_i's$. An estimation with a large bandwidth, on the other hand, is more robust to the nonuniformity of markers and produces a rather smooth curve, but may fail to detect significant jumps caused by a high local recombination rate. For our case, specifically, we should not use a bandwidth that is too small to avoid the bias caused by physical nonuniformity of markers.

As illustrated in Figure 1, the local polynomial kernel estimator with an appropriately selected bandwidth can, although with some variation, capture the underlying pattern of the true recombination rate from the genetic data if the markers are indeed uniformly distributed on the physical map. To analyze recombination rates for the 20 mouse chromosomes, we will use a bandwidth equal to 1/6 or 1/7 for the *Epanechnikov* kernel, depending on the number of markers on the chromosome.

3. Results

The plots of the normalized cumulative crossover counts versus the normalized rank order of markers, including the estimated mapping functions $\hat{\Lambda}$, are given on the left of Figure 2 for mouse chromosomes 4, 6, 7 and X. As in Nachman and Churchill (1996), the dashed identity line ($x = y$) gives the expectation under a uniform physical distribution of markers and no heterogeneity in recombination rate and is shown for reference. The Kolmogorov-Smirnov statistic measures the largest deviation of the cumulative distribution from this line. For each chromosome, this null hypothesis ($x = y$) is rejected by the Kolmogorov-Smirnov test ($P < 0.01$ for each). The estimated recombination rates for these mouse chromosomes are shown on the right of Figure 2. They are given by $\hat{\lambda}$, the pointwise local derivatives of the smoothed mapping functions $\hat{\Lambda}$. The rejection of the above Kolmogorov-Smirnov test for each chromosome implies that these recombination rates

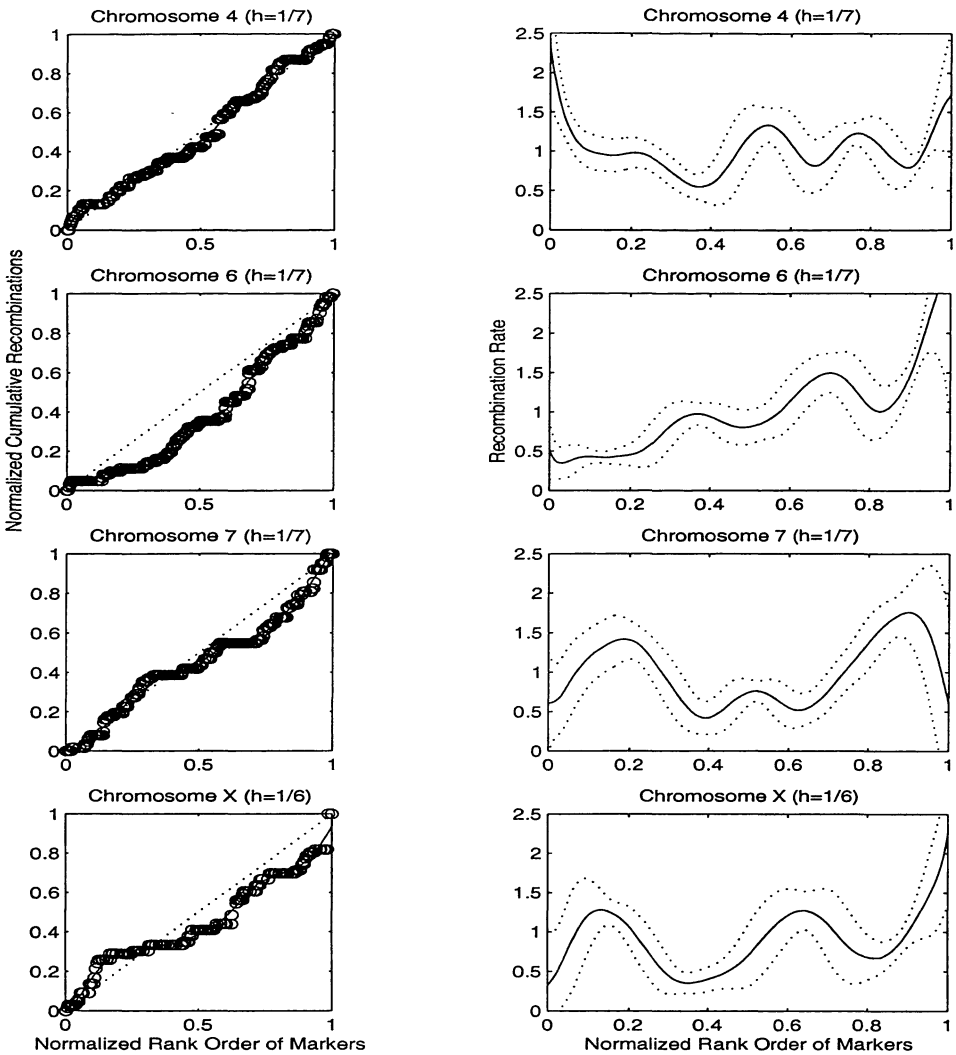

Figure 2: Estimated Recombination Rates

are all significantly different from the flat line $y = 1$.

An interesting comparison reveals high similarity between the patterns of the recombination rates estimated by the local polynomial kernel regression here and those obtained by taking the reciprocals of kernel estimators of marker densities in Nachman and Churchill (1996).

The "confidence" bands bounded by the dashed lines in Figure 2 for the recombination rates reflect the uncertainty and randomness in the physical positions of markers. The are calculated by generating 100 sorted sequences of random numbers between 0 and 1, representing 100 arbitrary possible allocations of markers along a chromosome. For each given sequence of marker positions, a smoother for recombination rate along this chromosome is evaluated at the same set of equally-spaced positions. Thus, there are 100 smoothed curves, representing different estimators of recombination rate. The average and standard deviation of these 100 curves are then calculated pointwise at the equally-spaced positions. A "confidence" band is calculated by adding and subtracting 2 standard deviations from the mean.

Since the mouse chromosomes are all acrocentric, the centromeres are located at one end (left, in our plots) instead of middle, and the telemeres are at the other end. As shown in the graphs, the pattern of recombination rates varies greatly from one chro-

mosome to another, and most chromosomes have the highest rate in regions close to the telemeres. Furthermore, there is evidence for recombinational hotspots on most of the 20 mouse chromosomes.

References

1. Altman, N.S. (1990). Kernel Smoothing of Data with Correlated Errors. *J. Amer. Statist. Assoc.*, **85** 749-759.
2. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. (1994). *Molecular Biology of the Cell, 3rd Ed.* Garland.
3. Dietrich, W.F. *et al.* (1992).A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**: 423-447.
4. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall: New York.
5. Griffiths, A.J.F. *et al.* (1996). *An Introduction to Genetic Analysis, 6th Ed.* Freeman.
6. Haldane, J.B.S. (1919). The Combination of Linkage Values and the Calculation of Distances Between Loci of Linked Factors. *Journal of Genetics* 8: 299-309.
7. Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall: New York.
8. Kindahl, E.C. (1994). Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster.* Ph.D. thesis. Cornell University, Ithaca, NY.
9. Kliman, R.M. and Hey, J. (1993). Reduced natural selection associated with low recombination in *Drosophila melanogaster. Mol. Biol. Evol.* **10**: 1239-1258.
10. Lyon, M.F. (1976). Distribution of crossing-over in mouse chromosomes. *Genet. Res. Camb.* **28**: 291-299.
11. Maynard Smith, J. (1978). *The Evolution of Sex*, Cambridge University Press, Cambridge.
12. Nachman, M.W. and Churchill, G.A. (1996). Heterogeneity in rates of recombination across the mouse genome. *Genetics* **142**: 537-548.
13. Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman and Hall: London.
14. Watson, J.D. *et al,*(1987). *Molecular Biology of the Gene, 4th Ed.* The Benjamin/Cummings.