JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 3, 2007

Pp. 255–266
DOI: 10.1089/cmb.2006.0006

A Quantile Method for Sizing Optical Maps

HAIFENG LI,[1] ANTON VALOUEV,[2] DAVID C. SCHWARTZ,[3]
MICHAEL S. WATERMAN,[1] and LEI M. LI[1,2]

ABSTRACT

Optical mapping is an integrated system for the analysis of single DNA molecules. It constructs restriction maps (noted as "optical map") from individual DNA molecules presented on surfaces after they are imaged by fluorescence microscopy. Because restriction digestion and fluorochrome staining are performed after molecules are mounted, resulting restriction fragments retain their order. Maps of fragment sizes and order are constructed by image processing techniques employing integrated fluorescence intensity measurements. Such analysis, in place of molecular length measurements, obviates need for uniformly elongated molecules, but requires samples containing small fluorescent reference molecules for accurate sizing. Although robust in practice, elimination of internal reference molecules would reduce errors and extend single molecule analysis to other platforms. In this paper, we introduce a new approach that does not use reference molecules for direct estimation of restriction fragment sizes, by the exploitation of the quantiles associated with their expected distribution. We show that this approach is comparable to the current reference-based method as evaluated by map alignment techniques in terms of the rate of placement of optical maps to published sequence.

Key words: alignment, computational molecular biology, genomics, optical map, statistics.

1. INTRODUCTION

OPTICAL MAPPING is a proven, high-throughput, single-molecule system that constructs physical maps spanning entire genomes. Because restriction maps describe large-scale structure (0.5 kb—entire chromosomes), they intrinsically reveal a broad range of genomic alterations reflecting polymorphisms and aberrations. Also, such maps aid genome sequencing projects by facilitating assembly, independent validation and finishing efforts (Zhou et al., 2007). Although optical mapping analyzes clones and polymerase chain reaction (PCR) amplicons, significant automation efforts have centered on the construction of whole genome maps using large, randomly sheared genomic DNA molecules (Armbrust et al., 2004; Reslewic

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, California.
[2]Department of Mathematics, University of Southern California, Los Angeles, California.
[3]Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics, University of Wisconsin–Madison, University of Wisconsin Biotechnology Center, Madison, Wisconsin.

et al., 2005). Here, a microfluidic device deposits genomic DNA molecules as a series of centimeter long stripes onto critically charged glass surfaces. Since large molecules exist in solution as random coils, this procedure unravels them, producing a stretched form, which is electrostatically adhered to the surface, thus ensuring optimum presentation prior to restriction digestion. After staining with a fluorochrome dye, these cleavage sites are imaged by fluorescence microscopy as micron-sized gaps formed due to the local release of stored tension within stretched versus relaxed molecules. Restriction fragments retain their order and an automated image-acquisition and machine-vision system provides capacious single molecule data sets for analysis. Optical map construction uses integrated fluorescence intensity measurements for estimation of restriction fragment sizes. This procedure considers internal fluorescence standards consisting of co-mounted small DNA molecules (reference molecules) of known sequence composition for establishing a standard measure of integrated fluorescence intensity per Kb of DNA; a simple relationship then sizes restriction fragments (Lin et al., 1999) as follows:

$$\text{Estimated fragment size (kb)} = \frac{\text{integrated intensity for fragment}}{\text{standard intensity per Kb}}. \quad (1)$$

This sizing method suffers errors in the estimated standard intensity because of the estimation in Equation (1). Such errors arise because optical mapping estimates restriction fragment sizes by essentially assessing the ratio of dye molecules bound to a target and reference, and these measurements must deal with issues regarding dye uptake by DNA molecules bound to surfaces (Dimalanta et al., 2004).

We overcome these limitations by development of a new approach for the estimation of restriction fragments associated with optical maps. Currently, our approach assumes that a reference genome is available. We treat the sizing problem as a "blind inversion" problem. That is, we regard the process of measuring the fluorescence signals from single molecules as a system that considers inputs and outputs as just fragment sizes and intensities. Although both the effective system and input DNA fragment sizes are unknown to us, we can still estimate sizes given their intensities through a quantile-quantile correspondence. The procedure is dictated by the rationale of Blind Inversion Needs Distribution (BIND) (Li, 2003). A similar quantile method has been applied to the normalization of microarrays (Bolstad et al., 2003). Since our method uses only fluorescence intensity determinations of analyte molecules and a sequence from a reference genome, this approach eliminates the need for assaying internal reference molecules, and renders moot any errors stemming from such measurements. Furthermore, reduction of experimental variables gives increased simplicity, and this can foster development of new single molecule platforms aimed at genomic analysis. Here, preliminary analysis shows that the performance of our approach is comparable to that of the current reference molecule method as judged by rate of map placements to a reference map.

The remainder of the paper is organized as follows. Section 2 describes the proposed method with discussion regarding the treatment of typical errors associated with optical maps, including missing or spurious (false) restriction sites and missing fragments. In Section 3, we describe experimental results on some optical mapping data sets. Section 4 concludes the paper with some directions of further research.

2. METHODS

A statistical model for optical mapping

Our method utilizes the following statistical models dealing with measurement errors associated with optical mapping described in Valouev et al. (2006), and we briefly summarize them in what follows.

- **M1: A Poisson model for enzyme cuts.** Sizes (in Kb), Y, of genomic restriction fragments have an exponential density with mean $EY = \lambda$ that depends on the endonuclease used for the digestion. Consequently, the number of restriction sites in s Kb of DNA is a Poisson process with intensity s/λ (Churchill et al., 1989; Waterman, 1995). Typical enzymes produce fragments with average lengths of 12–40 Kb.
- **M2: Missing cuts.** After the DNA is enzymatically cleaved, some restriction sites are left uncut by the endonuclease (partial digestion). Observations of restriction sites on the optical maps are assumed to

be Bernoulli trials with the digestion rate p (usually $p \approx 0.8$). Hence, observed restriction sites are explained by a thinned Poisson process with intensity p/λ (Grimmett and Stirzaker, 1982).

- **M3: False cuts.** False cuts result from random DNA breaks or non-specific activity of endonuclease. We assume that random breaks show no sequence dependent bias and are uniformly distributed across the entire reference genome. According to the model, the number of false cuts per s Kb of DNA is a Poisson process with intensity ζs. On average, we observe about five false cuts per 1 Mb of DNA.
- **M4: Measurement error.** During the staining of DNA, the fluorescent dye randomly binds to sites on DNA molecules, despite saturation conditions. For a true fragment size Y, its observed fluorescence intensity X approximately follows a normal density $N(\mu Y, \mu\sigma^2 Y)$ for sufficiently large fragments (e.g., >4 Kb). This result is analytically supported by the central limit theorem (Valouev et al., 2006).

In addition, we have the following considerations.

- **Missing fragments.** Because electrostatic forces are used for retaining DNA molecules, smaller restriction fragments preferentially detach from the surface, and are consequently under-represented in data sets.
- **Chimeric maps.** Molecules that physically overlap on glass surfaces cannot be spatially resolved for revealing true connectivity of their respective portions. Consequently, some maps are chimeric, meaning that they represent multiple unrelated genomic regions.

Distribution of observed fragment intensities and an exponential approximation

Under the model assumptions, it was shown in Valouev et al. (2006) that measured sizes X of optical map fragments have an exponential density with the mean

$$\theta = \mu \left[\frac{1}{\sigma} \sqrt{\frac{2}{\tau} + \frac{1}{\sigma^2}} - \frac{1}{\sigma^2} \right]^{-1} \tag{2}$$

for $X \geq \Delta$, where Δ is a positive number and $\tau = \left(\zeta + \frac{p}{\lambda}\right)^{-1}$. We note that the exponential distribution can be viewed as an approximation to that of the real fragment intensities. In Figure 1, we plot the histograms of intensities from a human mapping project (see Section 3) using the enzyme SwaI with and without chimeric reads. Note that a large fraction of short fragments are missing as can be seen at the lower ends of the histograms. Also, histograms have longer tails than an exponential distribution. In the Q-Q plot, data are from samples in the quantile interval (0.15, 0.95) and the reference exponential distribution has

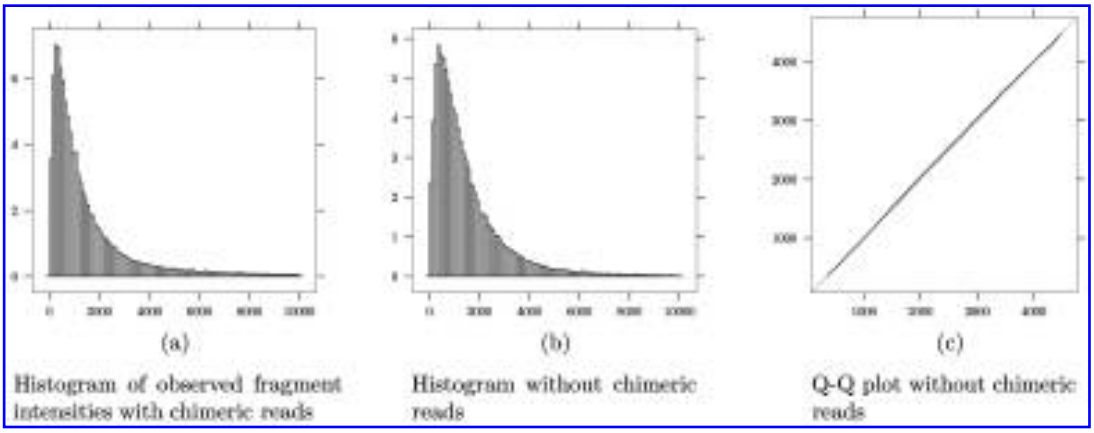


FIG. 1. Histograms and a Q-Q plot from SwaI-human optical maps with and without chimeric reads. In figures, intensity unit is 1000. Histograms have longer tails than an exponential distribution. In the Q-Q plot, data are from samples in the quantile interval (0.15, 0.95). Corresponding exponential distribution has mean 1.27×10^6. (**a**) Histogram of observed fragment intensities with chimeric reads. (**b**) Histogram without chimeric reads. (**c**) Q-Q plot without chimeric reads.

mean 1.27×10^6. It is clear that the distribution of the real fragment intensities is well approximated by an exponential distribution except for the lower and upper tails.

Distribution of fragment lengths

Now let us consider the hypothetical situation in which one fluorescence unit corresponds to exactly one base pair. If we apply the exponential approximation (2), then we have $\mu = 1$. Namely, measured fragment lengths of optical maps follow an exponential density with the mean $\theta = [\frac{1}{\sigma}\sqrt{\frac{2}{\tau} + \frac{1}{\sigma^2}} - \frac{1}{\sigma^2}]^{-1}$. Note that the three parameters—the digestion rate p, the false break rate ζ, and the standard deviation of measurement σ—are confounded in the expression of the exponential scale parameter. To reflect the detailed features of the human genome, we can simulate missing cuts and false cuts on the human genome according to M2 and M3, and common rate values. That is, restriction sites are skipped by 20% independently of each other, and on average, five false cuts per 1 Mb are induced according to a Poisson process. For the moment, we ignore the measurement error in the simulation. The distribution of the fragment lengths is shown in Figure 2a. Although a longer tail exists on the right, the middle part of the histogram looks fairly close to an exponential density. This is confirmed by the Q-Q plot of data versus exponential in the quantile range (0.15, 0.95). Furthermore, we observe excessive short fragments under 1 Kb in the log-scale Q-Q plot. The deviations from the exponential distribution on the lower and upper ends indicate the non-randomness of DNA sequences. Although it is well known that genome sequences are non-random, Churchill et al. (1989) found that the exponential density fits best for modeling the distribution of restriction fragment sizes compared to several other densities they examined in the case of *E. coli*. This assumption remains largely valid for other genomes from our experience. However, the assumption may be violated at the lower and upper ends as shown in Figure 2a and 2b. We take into account these issues in our sizing scheme by a censoring strategy.

Correspondence of the two distributions

Suppose we have n observed intensities $x_1, \ldots, x_n$. Their corresponding fragment lengths are denoted by $y_1, \ldots, y_n$, which are unknown. The goal is to estimate fragment length y_i of given intensity x_i. We formulate their relationship by

$$x = h(y), \tag{3}$$

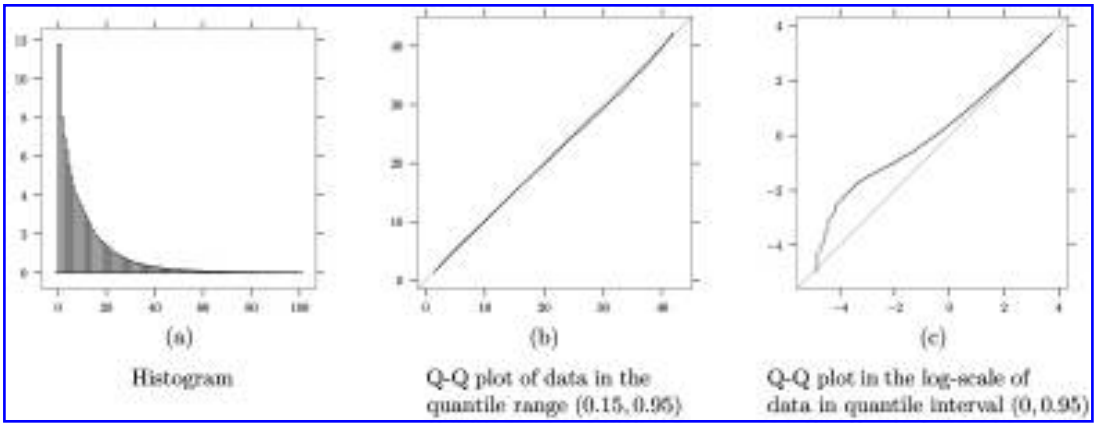


FIG. 2. Histogram and Q-Q plots of fragment sizes (kb) of human restriction map for enzyme SwaI. The histogram has a longer right tail (not shown in the figure) than the exponential distribution. Note that some long fragments could be more than 400 kb. In the Q-Q plots, the horizontal axis is the sorted fragment sizes of human restriction map while vertical axis is the quantiles of a reference exponential distribution with mean 11.49. Clearly, the major part of data follow an exponential distribution well as verified in (**a**) and (**b**). However, (**c**) indicates that there exist excessive short fragments (under 1 kb). (**a**) Histogram. (**b**) Q-Q plot of data in the quantile range (0.15, 0.95). (**c**) Q-Q plot in the log-scale of data in quantile interval (0, 0.95).

where $h(\cdot)$ represents the system of optical mapping instruments. In this formulation, both input y and the system function h are unknown. This is a typical "blind inversion" problem that appears in many scientific measurement problems such as DNA sequencing (Li, 2003). If it is true that larger fragment sizes have larger intensities, then we assume the following:

A1. h is monotone. That is, if $y_i \leq y_j$, then $x_i \leq x_j$.

Sort, respectively, the intensities and fragment sizes in the ascending orders, namely, $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$, and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Under the monotone transformation, the size of $y_{(i)}$ should map to $x_{(i)}$. But the latter are unfortunately not observed. However, if the distribution of the input is known, we can approximate $\{x_{(i)}\}$ by the i/n-th quantiles of x. Ideally, the larger the n is, the more accurate the approximation is. Explicitly, the rationale for the sizing method is as follows:

A2. The distribution of the reconstructed fragment lengths should match the expected distribution of fragment lengths.

If we have the complete data set of all fragment intensities, small or large, the problem is solved since we can obtain the expected distribution of fragment lengths as illustrated earlier. In this case, we even do not need to assume any parametric model other than the random mechanisms of missing and false cuts. However, the size estimation is difficult because many short fragments are missing in the observed intensities as shown in Figure 1b and 1c. Besides, there are only a few very large fragments and the match of quantiles on the very right end is difficult. To be able to apply our method, we use the following strategy to deal with the complications.

A3. For both the distribution of the observed fragment intensities and the expected distribution of fragment lengths, the middle part of the distribution approximately follows a truncated (from both ends) exponential distribution.

Figures 1 and 2 show supporting evidence to this claim. Moreover, we describe each of the two distributions respectively by a mixture model of three components. The principal component is an exponential distribution; two other components respectively have their mass located mostly at the lower end and the upper end. Note that the mixing fractions of these two components are small and we avoid assuming any specific forms. Besides, a fraction of small fragments is missing in the distribution of intensities. To deal with the difficulties on the two ends, we modify the scheme as follows: first we estimate the fragment sizes in the middle part by matching quantiles; second, we estimate the fragment sizes on the two ends by the ratio of the two exponential scales, respectively, estimated from intensities and fragment lengths of the simulated reference genome.

MLE for the exponential scale and two cutoff quantiles from the truncated data

As explained earlier, the major and middle part of both intensities and fragment sizes can be approximated by exponential distributions. To reduce the influence of the two end components on the parameter estimation of the principal exponential distribution, we remove data from the two ends. The above sizing schemes requires estimates of the parameters in the exponential distribution. They include the scale parameter and the two quantiles corresponding to two given cutoff values.

We propose two algorithms to compute the parameter estimates. One assumes that the middle part of data follows a doubly truncated exponential distribution and uses standard maximum likelihood estimation (MLE) method to estimate its mean. The other approach is an EM algorithm (Dempster et al., 1977; Redner and Walker, 1984), which treats the exponential components on the two ends as missing data. Our simulation (not shown) indicates that both algorithms perform well and give consistent estimates.

Let X be a random variable following an exponential distribution with mean λ. Suppose we have n observations $x_1, \ldots, x_n$. Given two cutoff values $a < b$, we remove observations such that $x_i < a$ or $x_i > b$ and treat them as missing data. Suppose n' observations are left and are denoted by $x'_1, \ldots, x'_{n'}$.

Now the remaining samples $x'_1, \ldots, x'_{n'}$ are drawn from a doubly truncated exponential distribution (Cohen, 1991) with density

$$f(x;\lambda) = \begin{cases} 0 & x < a \text{ or } x > b \\ \dfrac{\frac{1}{\lambda}e^{-x/\lambda}}{F(b) - F(a)} & a \le x \le b, \end{cases} \tag{4}$$

where $F(x) = 1 - e^{-x/\lambda}$ is the distribution function of exponential distribution. Therefore, the log likelihood function is

$$\mathcal{L} = -\frac{1}{\lambda}\sum_{i=1}^{n'} x'_i - n' \log[\lambda[F(b) - F(a)]]. \tag{5}$$

Newton-Raphson algorithm

By setting the derivative $\partial\mathcal{L}/\partial\lambda$ equal to zero, we have

$$\lambda = \bar{x} - a + \frac{(b-a)[1 - F(b)]}{F(b) - F(a)}, \tag{6}$$

where $\bar{x} = \frac{1}{n'}\sum_{i=1}^{n'} x'_i$. Because no closed form solution exists, we compute a numerical solution using the Newton-Raphson algorithm.

EM algorithm

Starting with an initial value of λ, iterate over the following steps until λ_k converges.

E-Step. For the current parameter value λ_k, calculate $q_a = \Pr(X < a|\lambda_k)$ and $q_b = \Pr(X < b|\lambda_k)$. Since n' observations are in the range $[a, b]$, on average $n'q_a/(q_b - q_a)$ and $n'(1-q_b)/(q_b - q_a)$ samples are missing in the lower and upper ends, respectively. We impute the value of the lower end samples by $x_a = E(X|X < a; \lambda_k) = (\lambda_k - (a + \lambda_k)e^{-a/\lambda_k})/(1 - e^{-a/\lambda_k})$ and the upper end ones by $x_b = E(X|X > b; \lambda_k) = b + \lambda_k$.

M-Step. Combine observations and imputed missing data, and compute the standard MLE on the complete data. Namely,

$$\lambda_{k+1} = \frac{\sum_{i=1}^{n'} x'_i + x_a n' q_a/(q_b - q_a) + x_b n'(1 - q_b)/(q_b - q_a)}{n'/(q_b - q_a)}. \tag{7}$$

Detailed sizing scheme

Given an optical map data of n intensities, we sort them in the ascending order. We choose two cutoff values a and b corresponding respectively to two quantiles of the raw intensities, r_a and r_b such that $0 < r_a < r_b < 1$. Suppose n_1 intensities are smaller than a, n_2 intensities are larger than b, and n' intensities are in $[a, b]$, where $n_1 + n' + n_2 = n$. We next apply either the proposed EM or Newton-Raphson algorithm to the n' intensities in $[a, b]$ to estimate the mean value λ of underlying exponential distribution. Let $q_a = F(a|\lambda)$ and $q_b = F(b|\lambda)$. Similar to the imputation in the EM algorithm, there should be, on average, $n_a = q_a n'/(q_b - q_a)$ intensities less than a and $n_b = (1 - q_b)n'/(q_b - q_a)$ intensities greater than b. Note that the imputation is for the principal exponential component. Denote the sorted intensities in the range $[a, b]$ by $a \le y_{(n_1+1)} \le y_{(n_1+2)} \le \cdots < y_{(n_1+i)} \le \cdots y_{(n_1+n')} \le b$. As a result, the quantile of $y_{(n_1+i)}$ in the principal exponential distribution is given by

$$q_i = \frac{i + n_a}{n' + n_a + n_b}. \tag{8}$$

ALGORITHM A QUANTILE METHOD FOR SIZING OPTICAL MAPPING FRAGMENTS

Input: Intensities $y_1, \ldots, y_n$ of an optical map, the reference genome sequences and a given restriction enzyme, digestion rate p, random breakage rate ζ, and parameters r_a and r_b such that $0 < r_a < r_b < 1$.

Output: Fragment sizes.

Steps:

1. Generate reference restriction map of given DNA sequence and enzyme. Simulate missing cuts and false cuts with digestion rate p and random breakage rate ζ on the whole sequence, respectively. Repeat the simulation sufficient times and denote simulated fragment sizes $x_1, \ldots, x_m$.
2. Sort intensities. Let a and b be intensities at quantiles r_a and r_b, respectively. Suppose n_1 intensities are smaller than a and n' intensities are in $[a, b]$. Apply the proposed EM or MLE algorithm to intensities such that $a \le y_i \le b$. Denote the estimated mean of exponential distribution by λ.
3. Let $q_a = F(a|\lambda)$, $q_b = F(b|\lambda)$, $n_a = q_a n'/(q_b - q_a)$, and $n_b = (1 - q_b)n'/(q_b - q_a)$. For intensities $a \le y_{(n_1+1)} \le y_{(n_1+2)} \le \cdots < y_{(n_1+i)} \le \cdots y_{(n_1+n')} \le b$, calculate the corresponding quantiles of $y_{(n_1+i)}$ by $q_i = \frac{i+n_a}{n'+n_a+n_b}$.
4. Repeat Steps 2 and 3 on simulated fragment sizes. Denote the estimated mean of the underlying exponential distribution by λ_r.
5. For intensities $a \le y_i \le b$, map to the fragment sizes at the same quantiles.
6. For intensities $y_i < a$ or $y_i > b$, calculate the corresponding size as $y_i \lambda_r / \lambda$ and shift them so that the estimated sizes are continuous at the cutoffs.

FIG. 3. The algorithm of the quantile sizing method.

Similarly, we apply the above procedure to fragment sizes of the reference map to obtain the adjusted quantiles excluding excessive short and long fragments. Denote by λ_r the estimated mean of fragment sizes computed from the EM or Newton-Raphson algorithm.

Finally, we can map intensities $\{y_i\}$ in $[a, b]$ to fragment sizes $\{x_j\}$ at the same quantile. For intensities less than a or greater than b, we cannot estimate their accurate quantiles because they do not follow the same distribution. Instead, we estimate the standard intensity μ by λ/λ_r and determine the fragment size x of an intensity y by $x = y/\mu$ if $y < a$ or $y > b$. Note that the fragment sizes estimated in this way may differ from those obtained through quantile mapping at points a and b. Thus, we make the following continuity correction: shift the sizes on each end by a constant so that the estimated fragment sizes are continuous at cutoff points a and b. We summarize the entire sizing procedure in Figure 3.

Practical issues

Clearly, q_a and q_b are critical for the precision of size estimation. The interval (q_a, q_b) should be appropriately selected so that we use as many fragments as possible in the quantile method whereas excessive short and long fragments are excluded. Histograms and Q-Q plots are helpful for selecting q_a and q_b. We may also optimize them according to the placement performance (Valouev et al., 2006). The digestion rate p and random breakage rate ζ can be estimated from abundant previous optical map data by fitting them to the reference map.

3. METHOD VALIDATION

Experiments

We implemented our method by analysis of optical mapping results from four separate "mounts" that were part of an ongoing human genome mapping project. The DNA was digested using the restriction enzyme SwaI. Mounts 1 and 2 are of high quality, whereas 3 and 4 are of typical quality.

The quality of these measurements is assessed by the number of "flagged" restriction fragments assigned by the image processing software. Briefly, the software "flags" an imaged fragment when associated pixels show abnormally high fluorescence intensities (Runnheim, private communication). This phenomenon occurs—as one example—when molecules overlap, giving rise to chimeric maps. Because such occurrences produce ambiguous results, we remove these fragments from the quantile sizing procedure.

Typically, the optical mapping microfluidic device deposits 48 independent stripes (channels) of DNA molecules onto a surface. However, measurements from only 47 and 27 channels were acquired for Surfaces 1 and 4. The details of the four data sets are described in Table 1. Surface 1 has 77,986 fragments from 4370 molecules, of which 22.9% fragments are flagged. Surface 2 has 75,991 fragments from 4565 molecules, of which 22.2% fragments are flagged. Surface 3 has 73,843 fragments from 5014 molecules, of which 28.6% fragments are flagged. Surface 4 has only 48,596 fragments from 3015 molecules since only 27 channels are available. Among them, 26.3% fragments are flagged. Clearly, Surfaces 1 and 2 are of better quality than Surfaces 3 and 4 in terms of percentage of flagged fragments.

Given that this data set consists of randomly sheared molecules from a human genome, we do not have independent validation of our size measurements. As such, we take an alignment strategy to indirectly assess the sizing scheme. That is, we align optical maps sized by both the reference-DNA and the quantile methods to the reference genome, and then check the number of significant and consistent placements. The genomic identity of a single molecule optical map is found through its alignment to a reference genome. When fragment sizes are precisely estimated, the alignment procedure correctly places a map; this is noted by a significant score. However, when maps show a small number of fragments, often they are ambiguously placed at many genomic locations. Accordingly, we only consider maps containing at least 20 non-flagged restriction fragments.

The alignment and scoring system

We use the dynamic programming algorithm in Waterman et al. (1984) and Valouev et al. (2006) to compute the optimal restriction map alignments. Define a reference map $R = (r_0, \ldots, r_m)$ by a set of site positions relative to the start of the map, i.e., $0 = r_0 < r_1 < \cdots < r_m$. Similarly, let $Q = (q_0, \ldots, q_n)$ be an optical map such that $0 = q_0 < q_1 < \cdots < q_m$. The alignment between R and Q is given by a sequence of ordered pairs of matching site indices $(i_0, j_0), (i_1, j_1), \ldots, (i_d, j_d)$, such that $i_0 < i_1 < \cdots < i_d$, $j_0 < j_1 < \cdots < j_d$, i_t and j_t correspond to sites r_{i_t} and q_{j_t} of maps R and Q. Let $S(i, j)$ be the optimal score of the alignment between R and Q that the rightmost pair of sites i and j are aligned. The algorithm calculates the score $S(i, j)$ using the recursion

$$S(i, j) = \max_{(0 \vee i-\delta_1) \le g < i, (0 \vee j-\delta_2) \le h < j} [S(g, h) + X(q_i - q_g, r_j - r_h, i - g, j - h)], \tag{9}$$

TABLE 1. DESCRIPTION OF THE TEST DATA SETS

Surface	*Channels*	*Fragments*	*Flagged fragments*	*Molecules*	*Molecules with ≥20 non-flagged fragments*
1	47	77,986	22.9%	4370	852
2	48	75,991	22.2%	4565	769
3	48	73,843	28.6%	5014	437
4	27	48,596	26.3%	3015	357

where δ_1 and δ_2 are parameters that specify the sizes of maximum matching regions respectively for the reference and optical maps, and the function $X(q_i - q_g, r_j - r_h, i - g, j - h)$ calculates the score for the region $(g, h) \rightarrow (i, j)$.

The alignment algorithm adopts the scoring function proposed Valouev et al. (2006). The scores are log-likelihood ratios under two hypotheses: a given optical map is indeed from the region to which it is aligned or is independent of the region. The odds are based on the statistical model of optical map measurements and other complications such as missing cuts and false cuts (Valouev et al., 2006). Currently we treat the flagged fragments as missing when fitting optical maps to reference map. The scores for flagged fragments are simply set to be zero at this point.

Results

The fragment intensities of the four data sets are converted into fragment sizes by both the reference DNA and quantile methods. The obtained optical maps are then fitted to the reference genome. We want to compare the two sizing methods. In the quantile sizing method, the values of the two end quantiles r_a and r_b are around 0.30 and 0.95, respectively. In the scoring function of the alignment algorithm, the parameter λ, the mean of reference exponential distribution, is set to be 12.72 Kb, the average fragment length of the reference map; the maximum matching region size δ is 5; the standard deviation σ of size error for long fragments ($\geq$4 kb) is 0.46; the standard deviation η of size error for short fragments (<4 kb) is 0.53. The details of these parameters can be found in Valouev et al. (2006). The values of the digestion rate and the random breakage rate are typical for many optical maps. The parameters r_a, r_b, σ, and η are obtained by optimizing the fitting of long molecules with no flagged fragments to the reference genome.

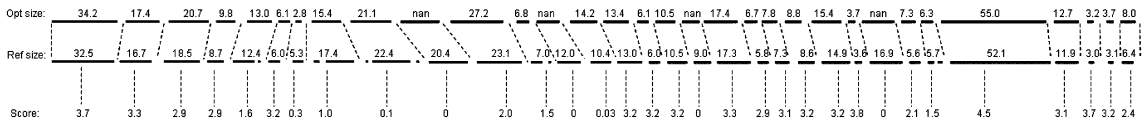
After obtaining the optimal fitting scores of molecules to the reference genome, we need to determine the significant hits. Note that molecules consisting of more fragments tend to have larger scores. We address this by considering the normalized score S/d, where S and d are, respectively, the fitting score of a molecule and the number of matching blocks. A matching block is defined to be a minimal set of fragments flanked by matching sites in the corresponding alignment to the genome. The significant placements are then determined based on the normalized scores. In Table 2, we list the numbers of significant placements for the cutoff value of 2.0 and 2.3. In order to assess the false positive rate, we did a simulation study. Namely, we simulated random maps and real maps from the reference genome and fit them back to the reference genome by the same set of parameters. The result shows a 5% false positive rate at the cutoff value of 2.3. However, in this simulation, the sizing errors follow the model of Valouev et al. (2006) and do not include the issue of flagged fragments. Thus, this evaluation only serves as a guide. We note that more multiple hits are observed in the quantile method.

In Table 2, we show not only the numbers of molecules placed by each method but also the numbers of molecules placed by both methods. In Surfaces 2, 3, and 4, large fractions of the significant placements are common to both methods. The consistency for Surface 1 is not as good as others. We checked closely into the placements of the "good" molecules that do not have flagged fragments by examining their alignments. Since false cuts are rare compared to missing cuts, we take 1 for the value of δ_1 and 5 for δ_2. Again most

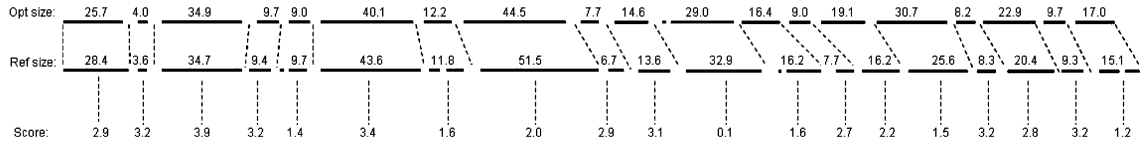
TABLE 2. A COMPARISON OF THE TWO SIZING METHODS

	2.0 cutoff value			*2.3 cutoff value*		
Sizing method	*Reference molecule*	*Quantile*	*Both*	*Reference molecule*	*Quantile*	*Both*
Surface 1	282	173	133	159	86	67
Surface 2	299	260	223	191	165	136
Surface 3	32	22	11	14	7	7
Surface 4	27	36	17	9	10	7

For each surface, we fit optical maps sized by both methods. A placement is called by two cutoff values 2.0 and 2.3. The numbers of placements by the reference molecule method, the quantile method, and by both methods are shown.



(a)



(b)

FIG. 4. Two alignment examples. (**a**) One alignment hit by both sizing methods. "Flagged" fragments are represented by "nan." (**b**) One alignment hit by only the quantile method.

of the significant hits are common to both methods. We checked those "good" molecules placed by the reference DNA method but not the quantile method, and found they contain more false cuts, indicating the possibility of spurious alignments. There are three molecules placed by the quantile method but not by the reference DNA method.

The average fragment size of the reference genome cut by SwaI is 12.72 Kb. After applying the 20% missing cuts and five false cuts per Mb, the average size of fragment lengths becomes 14.76 Kb. As we have explained, the principal exponential component of fragment lengths plays an important role in our approach. If we take 30% and 95% as the truncation thresholds, the estimates of the principal exponential components of the two kinds of fragment lengths are respectively 11.82, 13.85. The estimates of the exponential scale parameter for the sizes determined by the quantile method are 13.08, 14.80, 14.91, and 14.92 Kb for Surfaces 1, 2, 3, and 4. In comparison, the corresponding estimates for the sizes determined by the reference molecule method are 15.64, 18.63, 18.13, and 17.57 Kb. The difference may be accounted by the sampling bias due to the missing fragments and the flagged fragments. Remember that we do not include the flagged fragments in our quantile procedure. The average sizes of all fragments are larger than the estimates of the exponential scale parameter for both methods and in all chips. This is not a surprise because we only use an exponential density to approximate the principal component.

In Figure 4, we show two examples of alignment. The first one is hit by both sizing methods. "Flagged" fragments are represented by "nan." In the second alignment, only the score obtained by the quantile method is significant.

4. DISCUSSION

Our quantile sizing method is a robust procedure. In the statistical literature, there are several perspectives of robustness (Hampel et al., 1986; Huber, 1981). First, the influence curve or sensitivity curve of the quantiles less than a given $\alpha(< 1)$ is bounded. That is, the influence of an outlier of any kind is finite. As α approaches 1, the sensitivity gets higher as the density function decreases to zero. This is one reason why we exclude the very large fragments in the quantile matching. Second, the breakdown value of the α-th quantile is $\min(\alpha, 1-\alpha)$. This says that only quantiles on the two extremes are sensitive to perturbations. In the proposed approach, we re-scale fragments on the two ends by extrapolating size information in the middle part. Our numerical results support the view that our method is robust.

Even though the major and middle part of the observed fluorescence intensities is well approximated by an exponential distribution, we do not use exponential quantiles for size estimation. Rather, we use the quantiles obtained by mimicking the missing and false breaks on the reference genome. Importantly, the exponential model is only used to determine the quantiles of the two cutoffs.

Here, the quantile method has only been applied for the analysis of molecular fluorescence intensities derived from each surface. Thus we have ignored any potential spatial effect. Recall that an optical mapping microfluidic device contains multiple channels. A natural extension is to apply the method to each channel. But in this data set, a single channel contains about 100 molecules and may not be sufficient to accurately estimate the sizes. Alternatively, we can divide the surface into subareas and apply the quantile method to each subarea.

According to our preliminary tests, the quantile method is quite consistent with the current reference DNA method. Since the two perspectives are entirely independent, the new method can help validate the results obtained from the reference DNA method. In addition, it could be very important in situations where placement of reference DNA is difficult.

We describe a quantile method of sizing in the context of fitting optical maps to a reference genome. The same idea with necessary modifications may be applicable to the optical map assembly problem. In this problem, all pairwise overlaps are computed between optical maps (Valouev et al., 2006), and the reference genome is unknown. A possible scheme for size estimation is sketched as follows. First we compare the distribution of fragment intensities with an exponential distribution using Q-Q plots. Second, we map intensities to a reference exponential distribution instead of a known reference genome. The mean λ of the reference distribution can be guessed from a highly related genome or simply a random genome. We use the exponential distribution of mean $\tau = (\zeta + p/\lambda)^{-1}$ to simulate fragment sizes. Third, we fit an exponential distribution to the observed intensities excluding the lower and upper ends and apply our method to estimate the sizes of optical maps. We will explore this scheme in our future research.

ACKNOWLEDGMENTS

We gratefully thank Steve Goldstein and Rod Runnheim for preparing the data used by this research. This work was supported by Center of Excellence in Genome Science at University of Southern California, the NIH P50 HG002790 CEGS grant, and the NCI R33 Ca111933 grant. LML is partially supported by the NIH R01 GM75308-01 grant.

REFERENCES

Armbrust. E.V., Berges, J.A., Bowler, C., et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306, 79–86.

Bolstad, B., Irizarry, R.A., Astrand, M., et al. 2003. A comparison of normalization methods for high-density oligonucleotide array data based on bias and variance. *Bioinformatics* 19, 185–193.

Churchill, G.A., Daniels, D.L., and Waterman, M.S. 1989. The distribution of restriction enzyme sites *Escherichia coli*. *Nucleic Acids Res.* 18, 589–597.

Cohen, A.C. 1991. *Truncated and Censored Samples: Theory and Applications*. CRC Press, Boca Raton, FL.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39, 1–38.

Dimalanta, E.T., Lim, A., Runnheim, R., et al. 2004. A microfluidic system for large DNA molecule arrays. *Anal. Chem.* 76, 5293–5301.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., et al. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Huber, P.J. 1981. *Robust Statistics*. Wiley, New York.

Li, L.M. 2003. Blind inversion needs distribution (BIND): the general notion and case studies. *IMS Lect. Note Series* 40, 273–293.

Grimmett, G., and Stirzaker, D. 1982. *Probability and Random Processes*. Oxford University Press, Oxford, UK.

Lin, J., Qi, R., Aston, C., et al. 1999. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* 285, 1558–1562.

Redner, R.A., and Walker, H.F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* 26, 195–239.

Reslewic, S., Zhou, S., Place, M., et al. 2005. Whole-genome shotgun optical mapping of *Rhodospirillum rubrum*. *Appl. Environ. Microbiol.* 71, 5511–5522.

Valouev, A., Li, L.M., Liu, Y.-C., et al. 2006. Alignment of optical maps. *J. Comput. Biol.* 13, 442–462.

Waterman, M.S., Smith, T.F., and Katcher, H. 1984. Algorithms for restriction map comparisons. *Nucleic Acids Res.* 12, 237–242.

Waterman, M.S. 1995. *Introduction to Computational Biology*. Chapman and Hall/CRC, Boca Raton, FL.

Zhou, S., Herschleb, J., and Schwartz, D.C. 2007. Optical mappling: A single molecule system for genome analysis. *In* Mitchelson, K.R., ed., *New Methods for DNA Sequencing*. Elsevier, New York, pg. 265–300.

Address reprint requests to:
Dr. Lei M. Li
Molecular and Computational Biology Program
Department of Biological Sciences
University of Southern California
Los Angeles, CA 90089-2910

E-mail: lilei@usc.edu