CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Cold Spring Harbor Laboratory Institutional Repository

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 1, 2006
© Mary Ann Liebert, Inc.
Pp. 1–20

Validation of *S. Pombe* Sequence Assembly by Microarray Hybridization

JOSEPH WEST,[1] JOHN HEALY,[1] MICHAEL WIGLER,[1] WILLIAM CASEY,[2] and BUD MISHRA[2,3]

ABSTRACT

We describe a method to make physical maps of genomes using correlative hybridization patterns of probes to random pools of BACs. We derive thereby an estimated distance between probes, and then use this estimated distance to order probes. To test the method, we used BAC libraries from *Schizzosaccharomyces pombe*. We compared our data to the known sequence assembly, in order to assess accuracy. We demonstrate a small number of significant discrepancies between our method and the map derived by sequence assembly. Some of these discrepancies may arise because genome order within a population is not stable; imposing a linear order on a population may not be biologically meaningful.

Key words: physical maps, microarray hybridization, genome sequence assembly.

1. INTRODUCTION

In theory, a genome can be sequenced and assembled into a linear map without resorting to any outside physical mapping information (Weber *et al.*, 1997; Venter *et al.*, 2001). In the absence of any other physical location information, these methods depend upon the recognition of sequence overlaps. In practice, deriving a complete and accurate map this way is not achievable for any complex genome when the sequence reads are shorter than long repeats. By combining a pure shotgun approach with some distance information obtained from "mated-pairs" from end-sequenced clones, sequence-reads have been "contiged" and these contigs, phased, oriented, and ordered along a scaffold. Furthermore, even with a fairly large number of end-sequenced clones of various lengths and sequence reads as well as detailed knowledge of the genome structure, the sequence reads may not sufficiently cover the entire genome. In that case, sequencing cannot bridge the gaps, and a complete map cannot be made. Finally, if the genome is itself variable, containing polymorphic rearrangements within a population or between strains, there is no single true linear structure that will be valid for the organism.

Typically, physical mapping is used to facilitate sequence assembly, offering a large-scale map into which the local sequence assembly fits, bridging gaps and aiding in the organization of the sequencing tasks. And in principle, a high-resolution physical map could also aid in validating a sequence assembly and indicating where errors need correction.

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, P.O. Box 100, Cold Spring Harbor, NY 11724.
[2]Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012.
[3]Department of Cell Biology, NYU School of Medicine, New York, NY 10016.

In this paper, we explore the feasibility of making high-resolution genome maps using microarray hybridization, and using this data for sequence validation. We published a theoretical treatment of many of the ideas used here (Casey *et al.*, 2001; Mishra, 2002), which also contained the results of computer simulations. The basic idea is straightforward. Given that the genome is contained in a vector library of sufficient coverage, we hybridize many independent random pools of the library to arrays of probes, dense in the genome. When each pool from the library has a small depth of coverage, a sufficient informative "binary output" on the probes (hybridizes to the pool or not) allows the establishment of a distance function between probes. Using this distance function, we can infer the relative order and position of the probes in a linear map, within an experimental error. For example, if two probes, a and b, are within less than about a third of a BAC length of each other, more often than not a and b will both hybridize to the same set of BAC pools. More formally, when the distance between a and b is roughly one-third $(1/3 + (2/9 + o(1))c$, where $c =$ coverage in a pool) of a BAC length, the two probabilities of hybridization of "exactly one of a and b" or "both a and b" to a randomly selected small BAC pool become roughly equal, and the latter event dominates the former as a and b get closer. Thus, the degree of coincidence of their hybridization signals over a large series of hybridization experiments is statistically related to their actual distance between any two probes in base pairs. Extending this reasoning, if in many experiments one observes that among three probes a, b, and z, a and z as well as b and z hybridize together more often than do a and b, then it is reasonable to assume that the probe z is "between" a and b.

In our computer simulations and analytical formulation of this process, we modeled a library of BAC clones and tested different densities of probes and different pool sizes. The assembly process obeys "0–1 laws," in which long continuous and relatively error free assembly occurs only when a sharp threshold is exceeded by the available experimental data. We found in these studies that a probe density of about five probes per BAC length, a BAC library of about seven fold in depth, and hybridization with about 80 independently derived pools of BACs, each with about 25% coverage of the genome, produced contiguous maps of probes on the order of several megabases in length. Below, we briefly describe the rationale for the choices of these parameters.

Let $L = 166$ Kb be the length of a BAC. If β is the number of probes per BAC, then $\mu(\alpha) = L/\beta$ is the expected interprobe distance and is chosen based on one's desired map resolution. In this example, a desired resolution of 33 Kb gives us a $\beta = 5$ and is otherwise arbitrary.

If c denotes the BAC-coverage in each pool, in an ideal situation, it takes an optimal value of $c^* = 2\beta/(2\beta - 1)$, where it maximizes a probability p_h that determines the accuracy of interprobe distance estimation (see Lemma 1: $p_h \propto c\exp(-c)(1-\exp(-c/\beta))$. However, cross-hybridization error probability is minimized by making c arbitrarily close to zero. (It takes the form $1 - \exp(-\gamma(1 - \exp(-c)))$, with γ depending on thermodynamics of hybridization.) We chose a value of $c = 1/4 \in (0, 10/9]$. Also, in Lemma 1, we see that the number of experiments N determines the variance $\sigma^2(\alpha) = L^2/(2\beta Nc)$. If we wish the probes to be at least six sigma apart, then $\mu(\alpha)/\sigma(\alpha) = 3$, and $N = (9/2)\beta/c = 90$. Finally, we can compute the necessary BAC-library coverage C, by noting that as C increases, the number of uncovered gaps between our final set of contigs also decreases. We can compute that any chromosomal location belongs to a gap with probability $(1 - C/L)^{L(1-1/\beta)} \approx e^{-C(1-1/\beta)}$, where we may desire this probability to be bounded from above by ϵ_0.

We decided to test these ideas with actual experiments. The experiments themselves are expensive and so pilot experiments with a small model organism is highly desirable. We based our studies on the yeast *S. pombe*,[1] because both good BAC libraries[2] and a good sequence assembly were already available. The genome length for *S. Pombe* is 14 Mb and with an interprobe distance of 33 Kb on average, the total number of probes is 14 Mb/33 Kb = 424. Thus, a satisfactory choice of $\epsilon_0 = 1/424$ leads to a value of $C = \beta/(\beta - 1)\ln(1/\epsilon_0) = 7.562$.

In the experiments described below, we confirmed the computer and analytical predictions. A comparison of our data and inferred probe maps to the *S. pombe* sequence assembly map provides some insights into the difficulties of establishing a canonical and accurate sequence or physical map and suggests ways that the two types of data can be combined to render increased confidence levels of the assembly.

[1]Pombe sequence assembly was obtained from *www.sanger.ac.uk/Projects/S_pombe*.

[2]A full description of how the BAC library was constructed, including the vectors used and other relevant details can be found at *http://bacpac.chori.org/publications.htm*.

The raw and processed data from the entirety of our experiments, as well as inferred pairwise distances, is available on-line for further computational analysis.

The paper is organized as follows. We begin with an extensive discussions of the main results: experimental design, especially of complexity reduction (Subsection 2.1), and processing of raw data (Subsection 2.2). Once the data is converted into a binary form, we show how interprobe distances can be inferred (Subsection 2.3), with detailed statistics of our estimator. We next discuss the complexity of organizing the data in various graph structures (complete graph, tree structure, or linear structure), and discuss the results, when organized in a minimum spanning tree (MST) structure (Subsection 2.4). Equipped with this information, we show how our physical-mapping data can be compared with the sequence data, both analytically and visually (Subsection 2.5). In Section 3, we discuss the implications of our results and their possible applications to sequence assembly, finishing, validation, and correction; we also discuss how our approach is related to other physical mapping approaches (e.g., optical mapping, RH mapping and HAPPY mapping). In an appendix, we provide the details of the materials (microarrays and BAC pools), methods (representation, hybridization, data collection, data processing), and data availability.

2. RESULTS

2.1. Design of microarray hybridizations

DNA from BAC pools were made from a BAC library obtained from Pieter de Jong.[3] This library consisted of 3,072 individual elements, with an average insert size of 166 Kb. A library of this size has an expected depth of coverage of about 40 fold. The library was gridded in random order, and we picked 128 pools of 24 BACs each, covering the entire library. Each pool was expected to cover approximately 30% of the genome. To minimize unevenness of growth, each BAC was grown overnight in a 5 ml culture to saturation and then pooled in groups of 24 to inoculate a one-liter culture, from which highly purified BAC DNA was prepared. To obtain enough DNA for hybridization, these DNAs were amplified by making Sau3A1 high complexity representations (Lucito *et al.*, 1998).

Probes were designed to be relatively unique and to hybridize to high complexity representations. These representations are underrepresented for the genome sequences in Sau3A1 fragments smaller than 200 bp or larger than 1,200 bp, and hence we designed 70-mer length oligonucleotide probes to reside within 200 to 1,200 bp Sau3A1 fragments. We also required our probes to be unique sequences and used exact mer-matching methods (Healy *et al.*, 2003) to minimize the substrings of lengths 12 and 18 bases that matched elsewhere within the remainder of the *S. pombe* genome. Finally, although the physical mapping method works with randomly placed probes, to minimize the problems caused by the exponential distribution of the interprobe distances, we chose probes distributed roughly every 10 Kb in the genome. This resulted in a probe to BAC ratio of about 16 to 1, with approximately 1,224 probes in total, far in excess of the sufficient number (424) predicted by theory. Later, we "omitted" data to simulate the quality of the resulting physical map assembly with fewer probes and to test theory further.

It can be shown analytically that despite using representations that reduce genome complexity, a sufficient probe resolution is almost always possible. Note that when a genome is cleaved by a restriction enzyme with a cutting probability of p_r, the resulting restriction fragments have lengths distributed as $\sim$ Exponential(μ_r), where $\mu_r = [\log(1/(1-p_r))]^{-1} \approx 1/p_r$. Now suppose that a complexity reduction has been obtained by selecting only the restriction fragments of length $w \in [l, u]$. Then, in any genomic region of length L, the number, F, of reduced-complexity restriction fragments has a Poisson distribution with parameter $\lambda_{r,L} = Lp_r[e^{-l/\mu_r} - e^{-u/\mu_r}]$. Using a Chernoff bound, one can bound the probability that the number, F, of such reduced-complexity restriction fragments is less than $(1-\delta)\lambda_{r,L}$:

$$P[F < (1-\delta)\lambda_{r,L}] < e^{-\lambda_{r,L}\delta^2/2}, \quad \text{for } 0 < \delta < 1.$$

A simple computation then shows that, in a reduced-complexity representation with a four-cutter enzyme (e.g., Sau3a1), a single BAC contains 100 or more reduced complexity restriction fragments (of length between 200 and 1,200 bp) with a very high probability (higher than $1 - 1.25 \times 10^{-25}$).

[3]See *http://bacpac.chori.org/pombe104.htm*.

With low complexity representations, such as obtained by six-cutter enzymes, there can be an insufficient number of probes per BAC, and they will not contig well unless there is extremely high BAC coverage.

The set of 1,224 probes, resulting from above analysis, were synthesized (Dataset C; see Materials and Methods) and printed in randomized order, in quintuplicate, along with various controls, on glass slides. Hybridizations were performed as "two-color" experiments, with Cy5 and Cy3 labels, in which DNA from pools were labeled in one color and DNA from the entire BAC library was labeled in the other. To prepare DNA from the entire library, we pooled all the BACs from individual cultures, extracted DNA, and made Sau3A1 representations. We performed a limited number of experiments in color reversal (Shoemaker *et al.*, 2001) to identify probes with color bias. Color bias was not a significant problem, and we thus collected data in which the entire BAC library was labeled with Cy5 and all the BAC pools were labeled with Cy3.

2.2. Processing of raw data

The raw data consisted of 145 hybridizations because some of the 128 BAC pools were analyzed twice. We used only 128 of these hybridizations, because some of the data was judged to be of poor quality.

After normalizing each of the hybridizations (Dataset A; see Materials and Methods), we averaged the five quintuplicate log ratio values for each probe. The results from a typical hybridization are shown in Fig. 1, in which all probes are listed in genome order on the X-axis and their averaged log ratios on the Y-axis. The probe ratios clearly divide into two classes. The majority of probes are "nulls" (blue), meaning they do not hybridize to the BAC pool, while some are clearly "hits" (red), meaning that they do hybridize to the BAC pool. A few ambiguous probes have intermediate log ratios. Note that the hits tend to occur in clusters of adjacent probes, as we would expect since the probes are plotted in genome order, the assembly must be mostly correct, and a BAC would be expected to cover a contiguous set of probes along the genome. Since we know that the median BAC length is 166 Kb, and our probes are spaced every 10 Kb on average, we would expect that a typical BAC should cover approximately 16 contiguous probes. In some cases, there may be overlapping BACs in the same pool, and we would see longer contigs of probe hits as a result. Since our pool size of 24 BACs correspond to $c = 0.285$, we expect to find about $c \cdot \exp[-2c] = 3.9$ (on average) singleton BAC contigs out of a total of $c \cdot \exp[-c] = 5.1$ (on average) BAC contigs. In other words, after hybridization with a pool of randomly chosen 24 BACs, typically, we will see about four clusters of 16 contiguous probes, and one (or infrequently, two) more contig covering more than 16 contiguous probes.

To convert the averaged log ratio data into "probabilistic" form, we used an expectation maximization (EM) algorithm and assumed that that the log ratios from each experiment fell into two normal distributions, the "hits" and the "nulls." The EM finds the best fit of means and standard deviations of each population, enabling us to assign to each probe a probability that it is a hit or a null. Using this algorithm, the majority of probes can be unambiguously assigned to one group or the other. Very few probes have significant memberships in both groups. The outcomes of all hybridizations were thus compressed into a set of 1,224 hit vectors, one for each probe, each vector 128 long, consisting of the probabilistic weights of the probes being hit by a BAC in a pool (Dataset B, see Materials and Methods). Note that the computation of the hit vectors requires no knowledge of the genome order inferred by sequence assembly.

2.3. Computing the physical distance matrix

From the hit vectors, we can compute an estimate of the physical distance between each pair of probes. Given two hit vectors A and B of equal length, we define the Hamming distance $h(A, B)$ as the sum of the absolute value of the differences between identical positions in each of the two vectors,

$$h(A, B) = \sum_{i=1}^{N} |A_i - B_i|.$$

Tabulating these values, we obtain the 1,224 $\times$ 1,224 Hamming distance matrix, HDM. We also compute the number of "hits" of each probe, which is the sum of the weights of its hit vector:

$$\text{hits}(A) = \sum_i \mathbb{I}_{A_i > 1/2}.$$

Note that this number corresponds to the coverage of the probe in the BAC library.

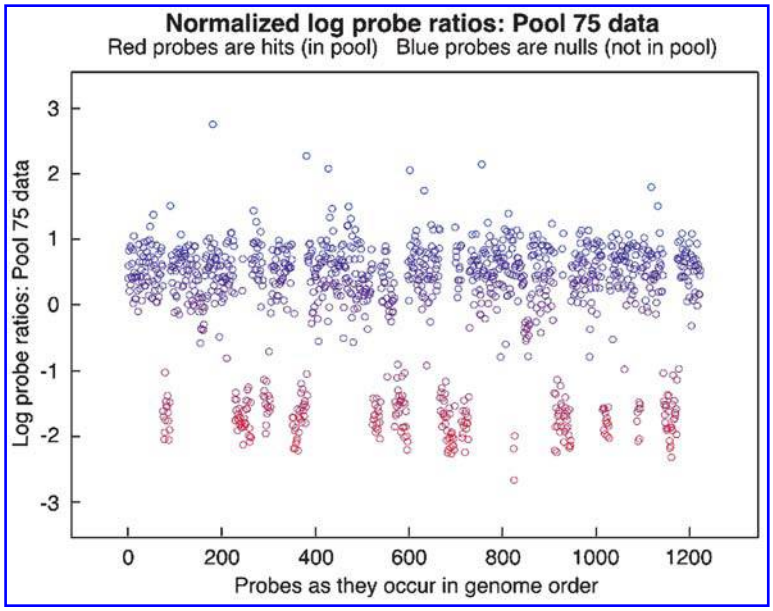


FIG. 1. Representative data from a single hybridization. Figure 1 illustrates the results of a typical hybridization (to BAC pool 75). The log intensity ratios for each probe, in sequence assembly order on the X-axis, are plotted on the Y-axis. See text for details.

From the Hamming distances and number of hits of two probes, we compute an estimate of the distance x between the respective probes a and b using the formula

$$x = \widehat{D}(a, b) \equiv \frac{h(A, B)}{\text{Hits}(A, B)} e^{\left(\frac{\text{Hits}(A,B)}{4N}\right)} L \tag{1}$$

where "Hits" are the combined number of hits of A and B, L is the mean BAC length, and N is the length of the hit vector (the number of hybridization experiments used).

We tabulate each pairwise estimate into a 1,224 $\times$ 1,224 matrix of distances, the "BDM" (BAC distance matrix). The derivation of the formula is as follows: Assume that the physical distance between two probes a and b is $x < L$. In addition to the Hamming distance $h(A, B)$, one also has a coincidence value $c(A, B)$ that measures the number of experiments in which both A and B get hit. Note that

$$\begin{aligned}\text{Hits}(A, B) &= \text{hits}(A) + \text{hits}(B) \\ &\approx h(A, B) + 2\sum_i \mathbb{I}_{(A_i > 1/2) \wedge (B_i > 1/2)} \approx h(A, B) + 2c(A, B).\end{aligned}$$

By a simple intuitive argument, the following approximate estimates can be derived, (when $x < L$):

$$h(A, B) \propto 2x, \quad \text{and}$$

$$c(A, B) \propto L - x,$$

with the same constant of proportionality. The intuitive argument is as follows: position(a) = position$(b) - x$, where "position" denotes a linear coordinate position along the genome and position$(a) \leq$ position(b). Note that only in experiments where BACs from the pools have their left ends either in the interval, [position$(a) - L$, position$(b) - L$], of length x (denoted "LEFT"), or in the interval, [position(a), position(b)] (denoted "RIGHT" interval), also of length x, do we have a contribution to the function $h(A, B)$. Further, note that only in experiments where BACs from the pools have their left ends in the interval, [position$(b) - L$, position(a)], of length $L - x$ (denoted "MIDDLE"), do we get a contribution to the function $c(A, B)$.

Thus,

$$\frac{h(A, B)}{\text{Hits}(A, B)} = \frac{h(A, B)}{h(A, B) + 2c(A, B)} \approx \frac{x}{L},$$

or

$$x = \frac{h(A, B)}{\text{Hits}(A, B)} \cdot L.$$

This formula is a good approximation and is correct if in a given BAC pool, no more than one BAC covers a or b (in the limit $\lim_{c\to 0}$).

However, a BAC may hit a without hitting b, and another may hit b without hitting a. With a better model of Poisson distribution for the terminals of the BACs, we can allow for these multiple hits as follows.

Lemma 2.1. *For two probes a and b, $x = D(a, b)$ distance apart, let A and B denote the N-dimensional vectors of hits and nulls obtained with N hybridizations with BAC pools, each of coverage $0 < c$. Assume that the BAC pools are randomly derived from a sufficiently large BAC library. Let L be the length of a BAC and x^* denote $x^* \equiv \min(x, L)$.*

The estimator for $x*$ is

$$\widehat{D}(a, b) \equiv \frac{h(A, B)}{Hits(A, B)} e^{\left(\frac{Hits(A,B)}{4N}\right)} L \approx x^* + \sqrt{\frac{x^* L}{2Nc}}\mathbb{N}(0, 1),$$

if c is small.

Proof. As before, assume that position(a) = position$(b) - x$, where "position" denotes a linear coordinate position along the genome, and position$(a) \le$ position(b).

Recall the notations

$$\text{"LEFT" interval} \equiv [\text{position}(a) - L, \min(\text{position}(b) - L, \text{position}(a))],$$

$$|\text{"LEFT" interval}| = x^*,$$

$$\text{"MIDDLE" interval} \equiv [\text{position}(b) - L, \text{position}(a)],$$

$$|\text{"MIDDLE" interval}| = L - x^*,$$

$$\text{"RIGHT" interval} \equiv [\max(\text{position}(a), \text{position}(b) - L), \text{position}(b)],$$

$$|\text{"RIGHT" interval}| = x^*.$$

Case 1, $x < L$.

After hybridizing the probes with a BAC pool of coverage c, we need to compute the following probabilities: Let q and s simply be the probabilities that no left end of any BAC appears in an interval of size x (e.g., LEFT or RIGHT intervals) and in an interval of size $L - x$ (e.g., MIDDLE interval), respectively.

$$q = e^{-cx/L} \quad \text{and} \quad s = e^{-c(L-x)/L} = e^{-c}e^{cx/L} = e^{-c}/q$$

Also, write $p \equiv (1 - q)$ and $r \equiv (1 - s)$, respectively, the probabilities that one or more BACs have their left ends in an interval of size x and $L - x$, respectively.

Now, it is straightforward to see that

$$p_b = q^2 s, \;\; p_h = 2psq, \;\; \text{and } p_c = 1 - (p_b + p_h) = 1 - s(1 + p)q$$

are the probabilities that (1) neither a nor b was hit (they are blank), (2) exactly one of a and b was hit (they contributed to Hamming distance), and (3) both a and b were hit (they contributed to coincidence measure, and hence local variations in coverage).

Note that

$$p_h = 2sq(1-q) = 2e^{-c}(1-e^{-cx/L}) \quad \text{and}$$
$$p_c = 1 - sq(2-q) = 1 - e^{-c}(2-e^{-cx/L}).$$

Also

$$p_h + 2p_c = 2 - 2sq = 2(1-e^{-c}) = 2c - c^2 + o(c^3), \text{ and}$$
$$c \approx (p_h + 2p_c)/2.$$

Thus,

$$\frac{p_h}{p_h + 2p_c} = (1-e^{-cx/L})\frac{e^{-c}}{(1-e^{-c})} = (1-e^{-cx/L})\frac{e^{-c/2}}{(e^{c/2}-e^{-c/2})}.$$

Simplifying, we have

$$L\frac{p_h}{p_h + 2p_c}e^{c/2} = L\frac{(1-e^{-cx/L})}{(e^{c/2}-e^{-c/2})}$$
$$= (1 - c^2/24 + o(c^4))x - (c/2L + o(c^3))x^2 + o(c^2(x/L)^2).$$

Thus, we have

$$h(A,B) \sim \text{Binomial}(N, p_h) \quad \text{and} \quad c(A,B) \sim \text{Binomial}(N, p_c).$$

Note that

$$c \approx \frac{h(A,B) + 2c(A,B)}{2N}$$
$$L\frac{p_h}{p_h + 2p_c}e^{c/2} \approx L\frac{h(A,B)}{h(A,B) + 2c(A,B)} \exp\left[\frac{h(A,B) + 2c(A,B)}{4N}\right]$$
$$\sim x + \sigma(x)\mathbb{N}(0,1) + o(c^2x + cx^2/L).$$

We can estimate σ^2 by a normal approximation to binomial distributions:

$$\sigma(x) = \alpha\sqrt{2Npsq}, \text{ where } \alpha \text{ is a scaling factor, } \alpha = \left(\frac{L}{N}\right)\frac{e^{c/2}}{2c}.$$

Thus,

$$\sigma(x) \approx (L/N)\frac{e^{c/2}}{2c}\sqrt{2N(cx/L)e^{-c}} = \sqrt{xL/2Nc}.$$

Case 2, $x \geq L$.

When $x \geq L$, essentially the same arguments work out, except that now $s = 0$ and $q = e^{-c}$. Thus

$$p_h = 2pq = 2e^{-c}(1-e^{-c}) \quad \text{and} \quad p_c = 1 - 2pq - q^2 = 1 - e^{-c}(2-e^{-c}).$$

Also,

$$p_h + 2p_c = 2 - 2q = 2(1-e^{-c}).$$

Thus,

$$\frac{p_h}{p_h + 2p_c} = e^{-c},$$

$$L\frac{p_h}{p_h + 2p_c}e^{c/2} = Le^{-c/2} = (1 - c/2 + c^2/8 + o(c^3))L.$$

After simplifying in the manner similar to above, we have

$$L\frac{h(A, B)}{h(A, B) + 2c(A, B)} \exp\left[\frac{h(A, B) + 2c(A, B)}{4N}\right] \sim L + L\sqrt{1/2Nc}\mathbb{N}(0, 1) \quad \blacksquare$$

Our formula uses a local estimation of c as Hits/$(2N)$ and hence is immune to local variations in the BAC library coverage, a very satisfying solution to the problem of uneven BAC coverage. For small c (e.g., $c = 1/4$) and $x < L$ (e.g., $x \approx L/16$), all but the most dominant term of the estimator formula can be safely ignored; thus, making our expression a good estimate of the interprobe distance, x, especially when it is sufficiently small with respect to the BAC length. Experimental validation of this formula can be seen in Fig. 2. Given the estimates of distances from our hybridization data alone, we can begin to derive a physical map and compare it to the map inferred from the sequence assembly. We will need the following corollary in order to carry out these comparisons.

Corollary 2.2. *Let a, b, $D(a, b)$, $\widehat{D}(a, b)$, L, and $\sigma = \sqrt{L/2Nc}$ be as before. Then we can write down the following conditional distribution functions:*

$$f(\widehat{D}|D) = \mathbb{I}_{D<L}\phi_{\sigma\sqrt{D}}(\widehat{D} - D) + \mathbb{I}_{D\geq L}\phi_{\sigma\sqrt{L}}(\widehat{D} - L),$$

$$f(D|\widehat{D}) \approx \mathbb{I}_{\widehat{D}<L}\phi_{\sigma\sqrt{\widehat{D}}}(\widehat{D} - D) + \mathbb{I}_{\widehat{D}\geq L}\mathbb{I}_{D\in[L,G]}\frac{1}{G - L}$$

where $\phi_\tau(y) = (1/\sqrt{2\pi\tau})\exp(-y^2/2\tau^2)$.

Proof. The first part is simply a restatement of the earlier lemma. The second part follows from Bayes' Rule:

$$f(D|\widehat{D}) = \frac{f(\widehat{D}|D)f(D)}{f(\widehat{D})} \tag{2}$$

$$\approx \frac{\frac{1}{G}\left(\mathbb{I}_{D<L}\phi_{\sigma\sqrt{D}}(\widehat{D} - D) + \mathbb{I}_{D\geq L}\phi_{\sigma\sqrt{L}}(\widehat{D} - L)\right)}{\left(\left(\frac{1}{G}\right)\mathbb{I}_{\widehat{D}<L} + \left(1 - \frac{L}{G}\right)\delta_{\widehat{D}=L}\right)}. \tag{3}$$

We have estimated $f(\widehat{D})$ using the following approximation when σ^2 is relatively small:

$$f(\widehat{D}) = \int_0^G f(\widehat{D}|D)f(D)\, dD.$$

The rest follows from appropriate algebraic simplifications. $\blacksquare$

Using this corollary, we now have the following way of measuring the goodness of an assembled sequence. Imagine that the locations of the unique probes $a_1, \ldots, a_n$ along the genomic sequence are

$$y_1 < y_2 < \cdots < y_{k-1} < y_k < y_{k+1} < \cdots < y_n.$$

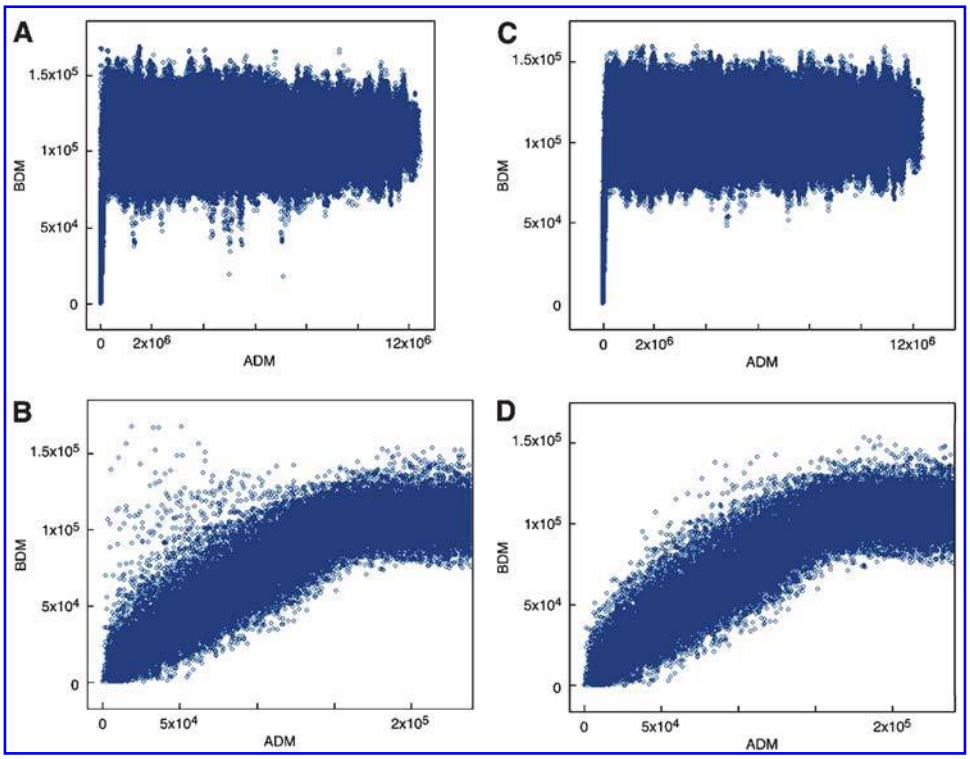


FIG. 2. Computed physical distance compared to sequence assembly distance. In all panels, the distances (in base pairs) between pairs of probes are plotted, with the physical distance (BAC distance measure, BDM) computed from Equation (1) on the Y-axis and the sequence assembly distance measure (ADM) plotted on the X-axis. The panels show different scales and slightly different sets of probe pairs. In **Panels A** and **C**, we use the full scale on the X-axis, and in **Panels B** and **D**, a smaller section on the X-axis where linearity from the physical distance is most apparent. Panels A and B are for all probe pairs, while Panels B and D are for all probe pairs less those edited out because of poor BAC coverage or aberrant pattern of BAC hybridization (see text and Fig. 5).

Then a reasonable measure of goodness of this assembly can be given as

$$\left\langle -\ln f(\widehat{D}(a_i, a_j)|D(a_i, a_j) = |y_i - y_j|)\right\rangle_{1\le i,j\le n}.$$

Thus, it can be measured by a global distance function:

$$d^2 = \left(\frac{1}{|\{i, j : 0 < |y_i - y_j| < L\}|}\sum_{i,j:0<|y_i-y_j|<L}\frac{(\widehat{D}(a_i, a_j) - |y_i - y_j|)^2}{|y_i - y_j|}\right)$$

$$= \left\langle\frac{(\widehat{D}(a_i, a_j) - |y_i - y_j|)^2}{|y_i - y_j|}\right\rangle_{i,j:0<|y_i-y_j|<L}.$$

Other asymmetric and local but more informative distances, which are better at localizing sequencing and assembly errors, can be given as follows:

$$d^2_{left,i} = \left\langle\frac{(\widehat{D}(a_i, a_j) - |y_i - y_j|)^2}{|y_i - y_j|}\right\rangle_{j:y_i-L<y_j<y_i} \quad \text{and}$$

$$d^2_{right,i} = \left\langle\frac{(\widehat{D}(a_i, a_j) - |y_i - y_j|)^2}{|y_i - y_j|}\right\rangle_{j:y_i<y_j<y_i+L}.$$

A symmetric situation arises when we have the measured distances $\widehat{D}(a_i, a_j)$ and we wish to organize the probes by embedding them on a real-line, which induces a linear order and a consistent set of pairwise distances:

$$\tilde{x}_1 < \tilde{x}_2 < \cdots < \tilde{x}_{k-1} < \tilde{x}_k < \tilde{x}_{k+1} < \cdots < \tilde{x}_n,$$

such that we minimize the following negative log likelihood function (under a mild independence assumption):

$$\left\langle -\ln f(\tilde{D}(a_i, a_j) = |\tilde{x}_i - \tilde{x}_j| \, |\widehat{D}(a_i, a_j)) \right\rangle_{1 \le i, j \le n}.$$

Thus, our problem reduces to the following optimization problem:

$$\text{minimize} \sum_{1 \le i, j \le n} W_{ij}(|\tilde{x}_i - \tilde{x}_j| - D_{ij})^2,$$

where

$$W_{ij} = \begin{cases} \dfrac{1}{2\sigma^2 D_{ij}}, & \text{if } D_{ij} < L; \\ \epsilon, & \text{otherwise,} \end{cases}$$

with $\epsilon = O(G^{-2})$ and $D_{ij} = \widehat{D}(a_i, a_j)$.

We will say more about this problem in the next subsection.

2.4. Assembling the probes into a graph

Given a matrix of pairwise distances between points (i.e., probes) on a line, there are several algorithms that can be used to derive a linear ordering of the points, or a map. If in fact the points lie on a line, if the distance matrix has no errors, and if there is no missing data, then there is always a single correct mapping. However, these assumptions do not necessarily hold in the present case, and even in "errorless" computer simulations we do not derive unambiguous orderings of our probes (West, 2003; Casey, 2002). Additionally, the experimental data is "noisy," and, as we shall see, even the assumption that our probes have a true linear ordering may not be correct. In fact, with real data we could not derive an unambiguously correct linear ordering, and hence we have explored other geometric structures into which we embed the distance relationship of our probes.

Note that, in theory, using an optimization criteria (e.g., negative log likelihood function, mentioned earlier), one could attempt to embed the probes on a line in such a manner that the desired criteria are satisfied. However, the structures of any reasonable optimality criteria are somewhat unruly and are rather closely related to difficult (NP-hard) combinatorial problems, as illustrated by the following problem.

Input: An $n \times n$ positive real-valued matrix D with $O(n)$ of the entries taking values in $[1-1/n, 1+1/n]$ (mean $\mu_g = 1$ and variance $\sigma_g^2 = 1/n^2$) and the remaining $O(n^2)$ entries taking values in $[1, 2]$ (mean $\mu_b = 3/2$ and variance $\sigma_b^2 = 1/2n$). Constant k is an arbitrary positive constant, and $n > 2 + 2\sqrt{k+1}$ is sufficiently large. A threshold Θ is given.

Cost: Given a permutation $\pi \in S_n$, assume a mapping of $\{1, \ldots, n\}$, $x_{\pi(i)} = i$, and a cost $C_D(\pi)$:

$$\sum_{i,j:|i-j|=1} W^{(1)}(|x_i - x_j| - D[i, j])^2 + \sum_{i,j:|i-j|\neq 1} W^{(2)}(\mu_b - D[i, j])^2,$$

with $W^{(1)} = 1/2\sigma_g^2 = n^2/2$ and $W^{(2)} = \epsilon = k/n^2$.

Output: Find a permutation $\pi \in S_n$ such that $C_D(\pi) < \Theta$.

This problem can be shown to be NP-complete. Since it is polynomially verifiable if a particular permutation π satisfies $C_D(\pi) < \Theta$, the problem is in NP. To see that the problem is NP-hard, consider a

cubic graph $G = (V, E)$ with vertices $1, \ldots, n$ and $|E| = \sum_{v \in V} \deg(v)/2 = 3n/2$, for which we wish to determine if the graph is Hamiltonian. Construct an instance of our problem with a matrix D as follows: $D[i, j] = 1$, if $\langle i, j \rangle \in E(G)$. Of the remaining $\binom{n}{2} - n + 1$ entries, make exactly $n/2 + 1$ entries $= 2$ and the other $= 3/2$. Let $\Theta = (n + k)/2$. If $\pi(1), \pi(2), \ldots, \pi(n)$ is a Hamiltonian path in G, then for D, it generates a cost bounded from above by

$$< W^{(1)} \times 0 + W^{(2)} \times 1/4 \left[\binom{n}{2} - n + 1 \right] < \frac{k}{2}[1 + 2/n^2] < (n + k)/2 = \Theta.$$

If, on the other hand, $\pi(1), \pi(2), \ldots, \pi(n)$ is *not* a Hamiltonian path in G, then for D, it generates a cost bounded from below by

$$> W^{(1)} \times (1 - 3/2)^2 > n^2/8 > (n + k)/2 = \Theta.$$

In spite of these computational complexity issues, there are several heuristics that can be shown to perform reasonably well.

For instance, in a greedy algorithm one can assume that the first $(i - 1)$ probes (where $i > 3$) $a_{\pi(1)}, \ldots, a_{\pi(i-1)}$ have already been embedded at

$$\tilde{x}_1 < \tilde{x}_2 < \cdots < \tilde{x}_{i-2} < \tilde{x}_{i-1},$$

and the next probe can be found by finding an a_j such that $\widehat{D}(a_{\pi(i-1)}, a_j)$ is the shortest among all probes satisfying the following constraint:

$$\widehat{D}(a_{\pi(i-2)}, a_j) > \widehat{D}(a_{\pi(i-2)}, a_{\pi(i-1)}).$$

The probe a_j found this way is labeled $a_{\pi(i)}$, and $\tilde{x}_i$ is computed as follows:

$$\tilde{x}_i = \frac{W_1[\tilde{x}_{i-1} + \widehat{D}(a_{\pi(i-1)}, a_{\pi(i)})] + W_2[\tilde{x}_{i-2} + \widehat{D}(a_{\pi(i-2)}, a_{\pi(i)})]}{W_1 + W_2},$$

where $W_p = (\widehat{D}(a_{\pi(i-p)}, a_{\pi(i)}))^{-1}$, $p = 1, 2$. The algorithm can be further generalized so that we may consult more than $p > 2$ many probes $a_{\pi(i-p)}, a_{\pi(i-p+1)}, \ldots, a_{\pi(i-2)}$, and $a_{\pi(i-1)}$ to choose the next probe, and also to assign it a location (West, 2003). An even stronger generalization can be achieved by creating many "contigs" in parallel and combining them in a union-find-like structure, by always selecting probe-pair distances from the shortest-to-longest order. The details of a variant of this algorithm is given by Casey *et al.* (2003) (also see Appendix B).

The other algorithmic choices can be based on (1) finding an optimal traveling salesman tour on the underlying graph, (2) finding a minimum spanning tree and linearizing it with some local heuristics, (3) successive-edge pruning from the underlying graph by eliminating edges that are unlikely to connect two adjacent probes, or (4) perturbing distance metrics to "linearize" the graph (Casey, 2002, Chap. 2).

For the current paper, we kept the matter simple by selecting one of the simplest ordering algorithms in order that the underlying data (e.g., experimental data and available sequence data) will be the easiest to interpret and debug. This algorithm involves constructing the minimum spanning tree of the weighted graph, induced by the estimated distances among probes. Our implementation uses Prim's algorithm (Cormen *et al.*, 2001) directly to the complete graph on probes with edge-weights chosen by the experimental data. Our implementation starts at a random probe, adjoins the nearest probe to the growing tree, and then halts when there is no probe left that, assuming a Gaussian distribution of probe distances among unrelated probes, would be expected to be a true neighbor.

The result of this method, applied to probes from the *S. pombe* chromosome 1, is shown in Fig. 3, in which the output of our algorithm is plotted using GraphViz, a set of graph drawing tools.[4] There is one long "contig" that is nearly linear, but not quite, having short branches (panel A). The branching structure is seen more clearly in panels B and C, successive blowups. The three isolated probes, which form their own

[4]Available at *www.research.att.com/sw/tools/graphviz/*.

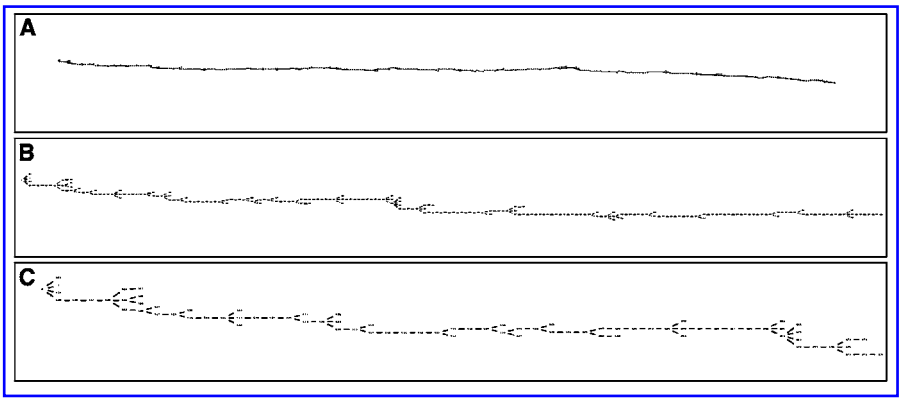


FIG. 3. Graph of minimal spanning tree, *S. pombe*, chromosome 1. A graphical representation of the minimal spanning tree generated from the BDM for all probes of chromosome 1 is shown full scale in **Panel A**, a blow up of the first quarter of the chromosome in **Panel B**, and the first eighth in **Panel C**. The beginning of the graph shows four branches, one for each telomere, one for the centromere, and the main long branch.

contigs of one (see the start of the graph, in panel C) and are not computed to be neighbors of any other probes, correspond to the centromeric and telomeric probes. They are either sparsely covered by BACs, having very low numbers of "hits," or behave anomalously in hybridization and have very high numbers of "hits". We obtain similar results with each of the other two *S. pombe* chromosomes. However, when our program is run on the entirety of *S. pombe* probes together, we obtain a single tree that, while still mainly linear, contains significantly long branches. The individual chromosomes are not recognized as separate contigs, and in contrast to the computation performed on the individual chromosomes, the telomeres and centromeres are joined to statistically significant neighbors. Some of the anomalies we observe may result from actual variation in the genomic structure of the *S. pombe* genome, and some from repetitive structure that is not apparent in the published sequence. We explore these aspects further in the next section.

2.5. Comparison of the hybridization map to the sequence map

Note that if the estimated distance between every consecutive pair of probes is small and has small relative error, then locally the distances satisfy a "triangle-like" inequality, i.e., one of the form

$$a < b < c \quad \Leftrightarrow \quad ab + bc = \min(ab + bc, ab + ac, ac + bc).$$

In this case, the minimum spanning tree is a single contig with all the probes in correct order.[5] However, in real experiments, these conditions are not met throughout: for instance, if we choose three probes a, b, and c that are separated from each other by distances longer than a BAC length, then all pairwise measured distances will be rather similar with values sampled from the distribution $L+(L/\sqrt{2NC})\mathbb{N}(0, 1)$. Consequently, the resulting minimum spanning tree is found to be mainly linear, with short branches. Within the longest linear path, the order of the probes closely matches the sequence assembly, and the branches contain nearby probes.

To see an overview of the minimal spanning tree and how it compares to the sequence assembly, we plot in Fig. 4 panel A all the "joins" of the minimal spanning trees for the entire *S. pombe* genome. In this display, for every edge of the spanning tree we plot "x" and "y," where x and y are the indices of the joined probes in the sequence assembly order (from 1 to 1,224). We note that probes from the telomeres of different

[5]This statement has a rather trivial proof. Consider a sequence of probes $a_1 < a_2 < \cdots < a_n$ with the measured distances satisfying triangle-like inequalities. Assume further that its minimum spanning tree is not a single contig. Thus, there exists an i such that the MST contains all edges along the path $a_1, \ldots, a_{i-1}$, but misses $\langle a_{i-1}, a_i \rangle$. Let $S = \{a_1, \ldots, a_{i-1}\}$, $A =$ the set of edges along the path. Let $(S, V - S)$ be a cut of G that respects A. By our triangle-like inequalities, $\langle a_{i-1}, a_i \rangle$ is the only light edge crossing $(S, V - S)$. Since it is also the only safe edge, we have a contradiction. See Cormen *et al.* (2001).

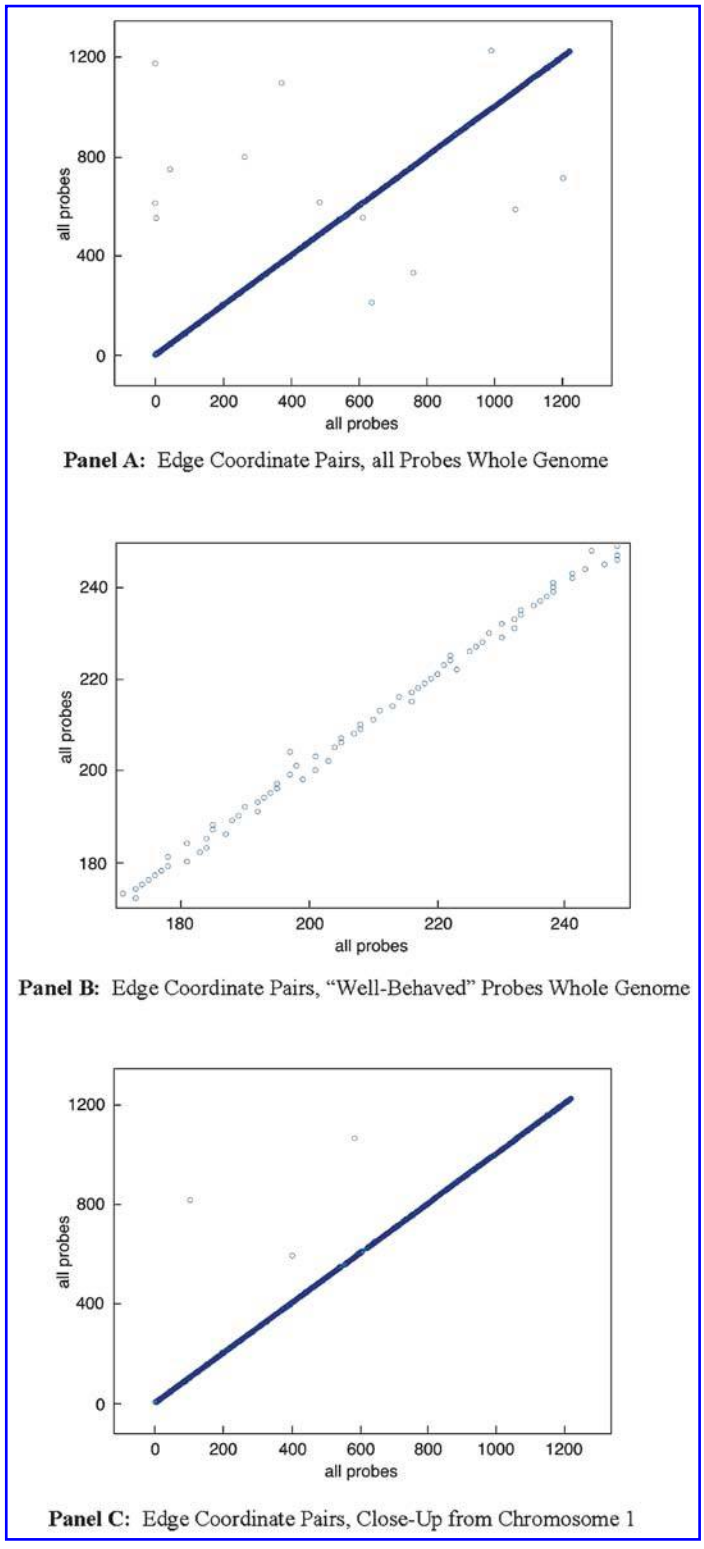


FIG. 4. Edge coordinate pairs from minimal spanning trees. The (sequence assembly) order of the probes that are joined by edges from the minimal spanning tree of the entire *S. pombe* genome is plotted in **Panel A** (see text). A higher resolution from a portion of chromosome 1 is shown in **Panel C**. The spanning tree of the "well behaved" probes (removing probes that show aberrant behavior in the BAC hybridization patterns, see text and Fig. 5) was recomputed, and the edge connections are shown in **Panel B**.

chromosomes are joined as neighbors and some centromeric probes are joined to essentially random probes within another chromosome. Of course, these associations disrupt a linear ordering of the genome.

At the resolution of panel A, the fine detail of the orderings is not apparent, so we show in Fig. 4 panel C a blow up of a randomly chosen region of chromosome 1. It is clear that at the fine level, the precise physical ordering of the probes is not coherent with the sequence ordering, but this is predicted from theory and results from statistical sampling noise and the paucity of BACs in the library with boundaries that fall between nearby probes.

A gross overview of the relationship between the physical map distance and the sequence assembly distance between probes can be viewed by plotting the two distances between all pairs of probes against each other: the sequence assembly distance on the X-axis and the physical distance (Equation (1)) on the Y-axis. This is shown in Fig. 2 panel A on a full scale of all pairwise probe distances and on panel B for the probe pairs that are closer together from the view of the sequence assembly. The overall shape of these plots closely resembles our theoretical predictions. Panel B shows the intrinsic limit of our method, namely, that distances between probes that are more than a BAC's length apart simply cannot be measured by this method. It is apparent on the full scale that a few probe pairs predicted in the sequence assembly to be distant appear close according to our BAC distance measure (BDM). The majority of these are telomeric and centromeric probes, or probes that fall into regions that have a very low number of BAC hits (regions of poor coverage in our library), and these are not a surprise. However, it is apparent that a few probes predicted by the sequence map to be close are mapped as distant by our method. This class is somewhat more disconcerting, but could in theory be caused by sequences complementary to our probes that are duplicated at two distant sites in the genome that was used for the library construction, but that were not duplicated in the genome that was used in the sequence assembly. Other discrepancies could be due to errors in either method.

There is perhaps a more informative way to examine the same question. We can display data from the BAC hybridization with probes in their sequence assembly order, and "view" where the BAC hybridization data and the sequence assembly deviate most radically from expectation. Then we can specifically query the physical pairwise BAC distance matrix to gather more information. From the BAC hybridization data, we compute three statistics for each probe in its genome assembly order: the number of experiments in which the probe and its left and right neighbors all hybridize to a BAC pool ("AllHits," blue open circles); the number of experiments in which the probe hybridizes to a BAC pool but its left and right neighbor do not ("SingleHit", open red triangles); and the number of experiments in which the left and right neighbor of a probe hybridize to a BAC pool, but the probe itself does not ("LonelyMiss," open green squares). In a noiseless experiment, except for those rare times when a BAC pool contains BACs just to the left and just to the right of a probe, SingleHits and LonelyMisses should be zero. For most probes, these values are low, but not zero. For a few probes there is a great variation from expectation.

In Fig. 5 we illustrate the plots of these statistics for a window from probe 560 to 730, all on chromosome 2. Three exceptional cases are seen, for probes 611, 639, and 712. Probe 611 has a high value of SingleHit, the other statistics being zero. In fact, this is a region predicted to derive from the centromere of chromosome 2, and its neighborhood must have very poor coverage by BACs. Like probe 212 from the centromere of chromosome 1, probe 611 displays a promiscuous hybridization pattern, and like probe 611, the neighborhood of probe 212 has poor coverage by BACs. The second probe, 639, has a high value for SingleHit, equal to its value for AllHits. When we ask which probes 639 maps closest to, it correctly maps to its closest assembly neighbor probes, although we calculate it as more distant from them than expected (data not shown). However, we also calculate probe 639 to be close to probe 212, the promiscuous probe from the centromere of chromosome 1. This fortuitous pattern of hybridization thus increases its apparent distance to its neighbors, as ascertained by our physical mapping methods.

The third probe is the most interesting of the three. It has high statistics for SingleHit and LonelyMiss, with a low statistic for AllHits. In fact, we map it to be very close to the neighborhood of probe 1203 on chromosome 3, which is otherwise close to its sequence assembly neighbors.

Clearly, unexpected behavior is seen in the map assembly, as branches in the minimal spanning tree (Fig. 3), aberrant edge connections (Fig. 4), discordance between the BAC mapping distance and sequence assembly distance (Fig. 2), and the pattern of BAC pool hybridization of sequence assembly neighbors (Fig. 5). These are presumably all related, and to test this, we created a new pairwise BAC distance matrix by removing the handful of probes that were judged to have distorted BAC hybridization in their

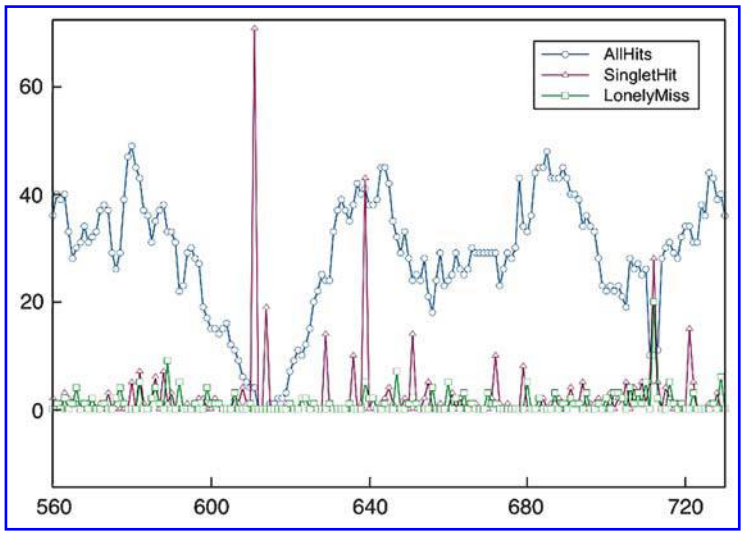


FIG. 5. BAC hybridization patterns across probes displayed in genome assembly order. The BAC hybridization parameters of probes 560 through 730, a region around the centromere of chromosome 2, are displayed (along the X-axis). Along the Y-axis, three statistical parameters for each probe have been plotted. These parameters, "AllHits," "SingleHit," and "LonelyMiss" are explained in the text.

neighborhood (by the criteria illustrated in Fig. 5). We then recomputed the minimal spanning tree, plotted the resulting edge connections (Fig. 4 panel B), and plotted again the comparison of the BDM (BAC distance measure) to the sequence assembly distance measures (Fig. 2 panels C and D). Not surprisingly, the most extreme discordances are thereby removed.

3. DISCUSSION

We have demonstrated empirically that with appropriate experimental conditions, microarray hybridization can be used to establish a physical distance between probes, and that this distance can be used to assemble physical maps and validate sequence assemblies of genomes. The critical conditions include libraries of genomic inserts of deep coverage, probes that are both reasonably unique in the genome and reasonably dense with respect to the length of the library insert, and a sufficient number of hybridizations. Our particular conditions were suggested by a theoretical model, and the empirical outcome in turn largely supported the theoretical modeling. Even the computer simulations of our method predict noise in the inferred distance, largely due to Poisson fluctuations in coverage. Theory predicts we cannot expect the method to give an accurate fine grain ordering because the probes are too dense relative to the BAC coverage, even with an unlimited number of hybridizations. There is more noise in the real data than we find in our "noiseless" simulations, causing both fine and coarse grain distortions in inferred distance. This additional noise can come from many sources: infidelity of the genomic inserts in the library, such as chimerism, deletions, and duplications; uneven amplification of DNA resources, both in library DNA preparation and in high complexity representations; poor or spurious hybridization patterns of the microarray probes; cryptic duplications of probe sequences in the genome; networking between library inserts during the hybridization stage; and even possibly variation in the genomic DNA from a single strain used for library production.

Despite all these possible sources of error, the method works well, as judged by its match to the *S. pombe* sequence assembly. Although we fail to assemble a linear map, the probes can be ordered into a minimal spanning tree which is largely linear (few long branches). The order between the nodes of this tree largely matches the order of the probes in the linear genome, especially if certain probes, such as probes from the telomeres, centromeres, poorly hybridizing probes, or probes with low BAC coverage, are removed.

There are areas where the inferred distance appears distorted, relative to the genome sequence assembly. These areas include all the probes that map to the telomeres and centromeres. The discrepancy of the

distance measure in these areas perhaps reflects poor BAC coverage, but there may be other factors at play. For example, we find probes from the centromeres appear to map to specific regions that are not centromeric or telomeric, despite the fact that our probes, designed from the public sequence assembly, are predicted to be unique. The public assembly may be in error, or these regions may be prone to rearrangement, or there may be differences in the strain used to build the library and the strain used to build the sequence assembly. Also, probes from different telomeres that are predicted to be unique nevertheless show proximity by our method, and this may be due to networking between repeated regions that are adjacent to our probes, or it may reflect high frequency recombination between telomeric sequences.

Even excluding telomeric and centromeric probes, there still remain a few areas of our map which do not match the assembly. In one set of cases, a small number of probes appear to map to two regions: one region that was predicted and one very distant unexpected region. In another case, a probe mapped to an altogether different region than was predicted. Some of these discrepancies can be explained as errors in the sequence assembly or differences between strains, such as duplicated or rearranged regions.

In any case, a high throughput method for physical mapping based on array hybridization is feasible and can serve as an independent method for validating a sequence assembly or as an aid to that assembly (when the sequence and the library of inserts are made from the same strain). When we initiated these studies, we used microarrays printed using pin technology from individually synthesized oligonucleotides. Physical printing using pins makes less than perfectly reliable substrates for hybridization, and oligonucleotide synthesis is expensive. Now, microarrays with very uniform character and with any desired oligonucleotide probe design can be fabricated via *in situ* synthesis. Although still not cheap, reproducibility is increased. Relative to the costs of assembly, the costs of physical mapping by array hybridization are minor.

A different approach to physical mapping and sequence validation (Mishra, 2003) of genomes comes from optical mapping, where random pieces of genomic DNA are stretched on a glass surface, cleaved by a restriction enzyme, photo-labeled with fluorochromes, and imaged by an optical camera. The distance between two restriction sites is measured by the integrated intensity of the corresponding imaged restriction fragment. This approach thus leads to a direct and accurate physical mapping technique, but relies on complex chemistries and algorithms. In contrast, the approach we present here is technologically simpler and can be adapted by any laboratory with access to microarray technology.

The costs, accuracy, and accessibility of our method compare very favorably to other physical mapping techniques. Our method is useful for validating a sequence assembly. But physical maps to correctly order a set of reagents, in the absence of a sequence assembly, are often very useful.

A radiation hybrid (RH) mapping technique (Boehnke *et al.*, 1991) uses high doses of X-rays to randomly fragment the genome to be mapped (genome of the donor cell) and fuses the fragments in to the chromosomes of a second species (genome of the recipient cell). The distance between two gene markers on the donor genome is computed from the estimated frequency of how often fragmentation events occurred at locations between the two markers, which ultimately placed the markers in two distant positions in the recipient cell's chromosome. In the case of HAPPY mapping (Dear, 1997), DNA carrying STS markers is extracted from cells (whose genome is to be mapped) and broken randomly to give a pool of fragments. Each pool is diluted and then screened by PCR to find the collection of markers in each pool. In a manner similar to our techniques and RH mapping, the distance between any pair of markers is then estimated by observing how often two markers co-occur in the same pool. As both of these techniques depend on estimating distances from the probability distribution of events modulated by relative positions of any pair of markers, our analysis and algorithm will apply equally well to all such approaches.

APPENDIX A. MATERIALS AND METHODS

A.1. Microarrays

We used the Cartesian PixSys 5500 arrayer to array our probe collection onto commercially prepared silanated glass slides. Each probe was spotted five times at random locations on the slide. This was done to control for any geometric or geographic artifacts on the array that were present on the slide itself before printing or that were induced by the processing of the slide during the hybridization or postprocessing steps.

A.2. Probe design

Our probes are 70 base-pair long oligonucleotides (70-mer) derived from short (200–1,200 base pairs) Sau3A1 restriction endonuclease fragments that were predicted to exist from analysis of the reference sequence of the *S. pombe* genome. Additionally, we used algorithms to maximize the uniqueness of the probe sequences (Healy *et al.*, 2003). The complete genome sequence of *Schizosaccharomyces pombe* is available for download from the website of The Wellcome Trust Sanger Institute.[6] The genomic DNA sequence of the *S. pombe* genome consisting of three chromosomes, each 5.5 million base pairs (Mbp), 4.4 Mbp, and 2.4 Mbp, respectively, were concatenated in silico to yield one large DNA molecule 12.3 Mbp in length. We then identified every subsequence of the genome that was flanked by a Sau3A1 restriction enzyme site and that was between 200 and 1,200 base pairs in length. Each of these identified subsequences was then tested for its constituent overlapping 12-mer and 18-mer frequencies against the entire *S. pombe* sequence. Only those subsequences with unique overlapping 18-mer frequencies were considered further. From the surviving subsequences with unique overlapping 18-mer frequency, we then selected a contiguous 70-mer fragment which had the minimal arithmetic mean of its constituent 12-mer frequency and with a GC content that was as close as possible to the overall average GC content of the *S. pombe* genome. Each of the selected 70-mer fragments was then tested for uniqueness in the *S. pombe* genome by conducting a low homology BLAST search. Finally, we selected 1,224 70-mer fragments so that the midpoint of each fragment was on average 10 Kb from the midpoint of any of its neighbors to the left and to the right. These 1,224 70-mer fragments are what we refer to as our probes (Dataset C).

A.3. BAC pools

The *S. pombe* BAC library has 3,072 BACs arrayed in eight 384-well micro-titer plates. The median clone size was determined to be 166,000 base pairs. Since the clones are unordered and each plate's dimensions are 24 wells by 16 wells, we simply chose a row of 24 clones to be a BAC pool. Sixteen rows per plate $\times 8$ plates $= 128$ pools of 24 clones each. Each pool is thus a random subset of 24 intervals of the *S. pombe* genome with the median length of each interval approximately 166,000 base pairs, and each pool of 24 clones thus represents approximately one-third of the *S. pombe* genome. Each clone of a pool was inoculated into an individual 5 ml culture media and grown to saturation overnight. The 24 saturated 5 ml cultures were then combined, and this 120 ml pooled culture was used to inoculate a larger 1,000 ml volume of broth. This was grown to saturation, and the bacteria collected by centrifugation. The pellets were drained and stored at $-70°$C until ready for further processing. BAC DNA was recovered from the frozen pellets by processing with the Qiagen Large Construct Kit protocol.

A.4. Representations

BAC pool representations were prepared as described by Lucito *et al.* (2000). Briefly, BAC pool DNA was digested to completion with Sau3A1, and cohesive adapters were ligated to the digested ends. PCR primers complementary to the ligated adapters were then used for amplification. Representations were cleaned by phenol:chloroform extraction, precipitated, and resuspended, and the concentration was determined. This material was then used as template in the PCR reaction.

A.5. Labeling of representations

Ten micrograms of representation was denatured by heating to 50°C in the presence of 5 mg random nonamer in a total of 100 ml. After five minutes, the sample was removed from heat and 20 ml of $5\times$ buffer was added (50 mM tris-HCl [pH 7.5], 25 mM MgCl2, 40 mM DTT, suspended with 33 mM dNTPs), 10 nmol of either Cy3 or Cy5 was added, and 5 units of Klenow fragment was added. After incubation of the reaction at 70°C for two hours, the reactions were combined and the incorporated probe was separated from the free unbound nucleotide by centrifugation through a Microcon YM-30 column. The labeled sample was then brought up to 15 ml at a concentration of $3\times$ SSC and 0.3% SDS, denatured, and then hybridized to the array of probes.

[6]Available at *www.sanger.ac.uk/Projects/S_pombe.*

A.6. Hybridization of representations to microarrays

Hybridization solution for printed slides consisted of 25% formamide, 5× SSC, 0.1% SDS. Then 25 μl of hybridization solution was added to the 10 μl of labeled sample and mixed. Samples were denatured in a MJ Research Tetrad at 50°C for 5 minutes and then incubated at 70°C for 30 minutes. Samples were spun down and pipetted onto slides prepared with lifter slip and incubated in a hybridization oven at 60°C for 14 to 16 hours. After hybridization, slides were washed, dried, and then scanned.

A.7. Scanning and data collection

An Axon GenePix 4000B scanner was used with a pixel size setting of 10 microns. GenePix Pro 4.0 software was utilized for quantitation of intensity for the arrays. Array data was imported into S-PLUS 6.1 for further analysis. Measured intensities without background subtraction were used to calculate ratios. For each pool (each hybridization corresponds to a separate pool of 24 BACs), we collected the median Cy3 and Cy5 channel intensities for each feature on the array. The Cy3 channel corresponded to the BAC pool DNA, and the Cy5 channel corresponded to the total genomic representation of the BAC library. Excluding controls, we collected intensity data on 6,120 features.

A.8. Data preprocessing

We then calculated the log (Cy5/Cy3) for each of the 6,120 features on the array. We did this for every pool that was hybridized (a total of 128 hybridizations). This resulted in a data matrix that was 6,120 rows by 128 columns (Dataset A). Since each probe was printed in quintuplicate, we then calculated the median log ratio over the five replicates for each probe and used this value as the value for that probe in that particular hybridization. This condensed our data matrix to 1,224 rows (each row representing a single probe) and 128 columns (each column representing a particular hybridization or BAC pool). The final step in the preprocessing of the data involved normalizing each column in the matrix so that the log ratios for each hybridization had a mean of zero and a standard deviation of 1. These values were then processed using an EM algorithm (see text), yielding a matrix 1,224 by 128, containing values between 0 and 1 (Dataset B). As described in the text, the computation of physical distances (using Equation (1)) is accomplished using Dataset B.

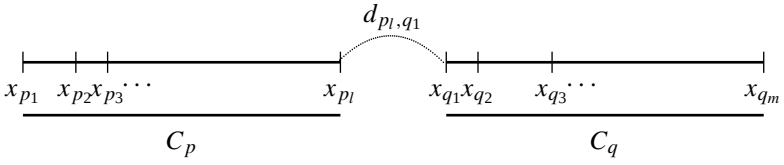
A.9. Data availability

The Datasets A (raw intensity ratios) , B (EM processed average log ratios, as probabilities), and C (all probe sequences), as tab delimited text files, are available for downloading as follows:

A: http://cs.nyu.edu/mishra/PUBLICATIONS/05.LinearMapData/DataSetA
B: http://cs.nyu.edu/mishra/PUBLICATIONS/05.LinearMapData/DataSetB
C: http://cs.nyu.edu/mishra/PUBLICATIONS/05.LinearMapData/DataSetC

APPENDIX B. SIMPLE ALGORITHM

The simplest algorithm to place probes proceeds as follows: Initially, every probe occurs in just one singleton contig, and the relative position of a probe $\tilde{x}_i$ in contig C_i is at the position 0. At any moment, two contigs $C_p = [\tilde{x}_{p_1}, \tilde{x}_{p_2}, \ldots, \tilde{x}_{p_l}]$ and $C_q = [\tilde{x}_{q_1}, \tilde{x}_{q_2}, \ldots, \tilde{x}_{q_m}]$ may be considered for a "join" operation: the result is either a failure to join the contigs C_p and C_q or a new contig C_r containing the probes from the constituent contigs. Without loss of generality, assume that $|C_p| \geq |C_q|$ and that the probe corresponding to the right end of the first contig (x_{p_l}) is closest to the left end of the other contig (x_{q_1}). That is, the estimated distance d_{p_l,q_1} is smaller than all other estimated distances: d_{p_1,q_1}, d_{p_1,q_m}, and d_{p_l,q_m}.



Let $0 < \theta \leq 1$ be a parameter that can be selected suitably (see Casey *et al.* [2001]), and $L' = L\theta \leq L$. If $d_{p_l,q_1} \geq L'$, then the join operation fails. Otherwise, the join operation succeeds with the probes of C_p placed to the left of the probes of C_q, with all the relative positions of the probes of each contig left undisturbed. We will estimate the distance between the probes in C_p and the probe x_{q_1} by minimizing the function:

$$\text{minimize} \sum_{i \in \{p_1,\ldots,p_l\}: d_{i,q_1} < L'} \frac{(\tilde{x}_{q_1} - \tilde{x}_i - d_{i,q_1})^2}{d_{i,q_1}},$$

where $\tilde{x}_i$'s ($i \in \{p_1, \ldots, p_l\}$) are fixed by the locations assigned in the contig C_p. Thus, taking a derivative of the expression above with respect to $\tilde{x}_{q_1}$ and equating it to zero, we see that the optimal location for x_{q_1} in C_r is

$$d^* = \max \left[\tilde{x}_{p_l}, \frac{\sum_{i \in \{p_1,\ldots,p_l\}: d_{i,q_1} < L'} \left(\tilde{x}_i + d_{i,q_1}\right)/d_{i,q_1}}{\sum_{i \in \{p_1,\ldots,p_l\}: d_{i,q_1} < L'} 1/d_{i,q_1}} \right].$$

Once the location of x_{q_1} is determined in C_r at d^*, the locations of all other probes of C_q in the new contig C_r are computed by shifting them by the value d^*. Thus,

$$C_r = [\tilde{x}_{r_1}, \ldots, \tilde{x}_{r_l}, \tilde{x}_{r_{l+1}}, \ldots, \tilde{x}_{r_{l+m}}],$$

where $r_i = p_i$ and $\tilde{x}_{r_i} = \tilde{x}_{p_i}$, for $1 \leq i \leq l$; $r_{l+i} = q_i$ and $\tilde{x}_{r_{l+i}} = d^* + \tilde{x}_{q_i}$, for $1 \leq i \leq m$. Note that when the join succeeds, the distance between the pair of consecutive probes $\tilde{x}_{r_l}$ and $\tilde{x}_{r_{l+1}}$ is

$$0 \leq \tilde{x}_{r_{l+1}} - \tilde{x}_{r_l} \leq L',$$

and the distances between all other consecutive pairs are exactly the same as what they were in the original constituent contigs. Thus, in any contig, the distance between every pair of consecutive probes takes a value between 0 and L'. Note that one may further simplify the distance computation by simply considering the k nearest neighbors of $\tilde{x}_{q_1}$ from the contig C_p: namely, $\tilde{x}_{p_{l-k+1}}, \ldots, \tilde{x}_{p_l}$.

$$d_k^* = \max \left[\tilde{x}_{p_l}, \frac{\sum_{i \in \{p_{l-k+1},\ldots,p_l\}: d_{i,q_1} < L'} \left(\tilde{x}_i + d_{i,q_1}\right)/d_{i,q_1}}{\sum_{i \in \{p_{l-k+1},\ldots,p_l\}: d_{i,q_1} < L'} 1/d_{i,q_1}} \right]$$

At any point, we can also improve the distances in a contig by running an "adjust" operation on a contig C_p with respect to a probe $\tilde{x}_{p_j}$, where

$$C_p = [\tilde{x}_{p_1}, \ldots, \tilde{x}_{p_{j-1}}, \tilde{x}_{p_j}, \tilde{x}_{p_{j+1}}, \ldots, \tilde{x}_{p_l}].$$

We achieve this by minimizing the following cost function:

$$\text{minimize} \sum_{i \in \{p_1,\ldots,p_l\} \setminus \{p_j\}: d_{i,p_j} < L'} \frac{(|\tilde{x}_{p_j} - \tilde{x}_i| - d_{i,p_j})^2}{2d_{i,p_j}},$$

where $\tilde{x}_i$'s ($i \in \{p_1, \ldots, p_l\} \setminus \{p_j\}$) are fixed by the locations assigned in the contig C_p.

Let

$$I_1 = \{i_1 \in \{p_1, \ldots, p_{j-1}\} : d_{i_1,p_j} < L'\},$$

$$I_2 = \{i_2 \in \{p_{j+1}, \ldots, p_l\} : d_{i_2,p_j} < L'\},$$

$$x^* = \frac{\sum_{i_1 \in I_1} \left(\tilde{x}_{i_1} + d_{i_1,p_j}\right)/d_{i_1,p_j} + \sum_{i_2 \in I_2} \left(\tilde{x}_{i_2} - d_{i_2,p_j}\right)/d_{i_2,p_j}}{\sum_{i_1 \in I_1} 1/d_{i_1,q_1} + \sum_{i_2 \in I_2} 1/d_{i_2,p_j}}.$$

At this point, if $x^* \neq \tilde{x}_{p_j}$, then the new position of the probe $\tilde{x}_{p_j}$ in the contig C_p is x^*. Note that the "adjust" operation always improves the quadratic cost function of the contig locally and since it is positive valued and bounded away from zero, the iterative improvement operations terminate.

ACKNOWLEDGMENTS

This work was supported by grants to M.W. from the National Institutes of Health R21HG02606; NYU/DARPA F5239. M.W. is an American Cancer Society Research Professor. B.M. is supported by grants from DARPA's BioCOMP project and AFRL contract F30602-01-2-0556. Additional support was provided to B.M. by NSF's Qubic and two ITR programs, and New York State Office of Science, Technology and Academic Research.

REFERENCES

Boehnke, M., Lange, K., and Cox, D.R. 1991. Statistical methods for multipoint radiation hybrid mapping. *Am. J. Human Genet.* 49, 1174–1188.

Casey, W. 2002. *Graph Embeddings with Application in Genomic Experiments*. Ph.D Thesis, NYU.

Casey, W., Mishra, B., and Wigler, M. 2001. Placing probes along the genome using pairwise distance data. *Algorithms in Bioinformatics, WABI 2001, LNCS* 2149, 52–68.

Cormen, T.H., Leiserson, C.E., and Rivest, R.L., and Stein, C. 2001. *Introduction to Algorithms*, MIT Press, Cambridge, MA.

Dear, P.H. 1997. HAPPY mapping. *Genome Mapping: A Practical Approach*. (Ed. P.H. Dear). Oxford University Press, London, UK. 95–124.

Healy, J., Thomas, E.E., Schwartz, J.T., and Wigler, M. 2003. Annotating large genomes with exact word matches. *Genome Res.* 13, 2306–2315.

Lucito, R., Nakimura, M., West, J.A., Han, Y., Chin, K., Jensen, K., McCombie, R., Gray, J.W., and Wigler, M. 1998. Genetic analysis using genomic representations. *Proc. Natl. Acad. Sci. USA* 95, 4487–4492.

Lucito, R., West, J., Reiner, A., Alexander, J., Esposito, D., Mishra, B., Powers, S., Norton, L., and Wigler, M. 2000. Genetic alterations in cancer detected by hybridization to micro-arrays of genomic representations. *Genome Res.* 10, 1726–1736.

Mishra, B. 2002. Comparing genomes. *Comput. Sci. Eng.* Jan/Feb, 42–49.

Mishra, B. 2003. Optical mapping. *Encyclopedia of the Human Genome* 4, 448–453.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., *et al.* 2001. Experimental annotation of the human genome using microarray technology. *Nature* 409, 922–925.

Venter, J.C., Adams, M.D., Myers, E.W., *et al.* 2001. The sequence of the human genome. *Science* 291, 1304–1351.

Weber, J., and Myers, E. 1997. Human whole genome shotgun sequencing. *Genome Res.* 7, 401–409.

West, J.A. 2003. *Micro-Array Based Genomic Mapping*. Ph.D Thesis, Cold Spring Harbor Laboratory.

Address correspondence to:
Bud Mishra
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY, 10012

E-mail: mishra@nyu.edu

This article has been cited by:

1. J.-H. Choi, S. Kim, H. Tang, J. Andrews, D. G. Gilbert, J. K. Colbourne. 2008. A machine-learning approach to combined evidence validation of genome assemblies. *Bioinformatics* **24**:6, 744-750. [CrossRef]