Validation of S. Pombe Sequence Assembly by Micro-Array Hybridization

Running title: Mapping Genomes Using Micro-arrays

Joseph West[1], John Healy[2], William Casey[3], Bhubaneswar Mishra[4], and Michael Wigler[2,5]

[1]DesignWrite
189 Wall Street
Princeton, NJ 08540

[2]Cold Spring Harbor Laboratory
1 Bungtown Road, P.O. Box 100
Cold Spring Harbor, NY 11724

[3]Institute for Physical Sciences (IPS)
1365 Beverly Road, Suite 300
McLean, VA 22101

[4]Courant Institute of Mathematical Science
251 Mercer Street, Room 801
New York, NY 10012

[5]Corresponding Author: Michael Wigler
Cold Spring Harbor Laboratory
1 Bungtown Road
Cold Spring Harbor, NY 11724
e-mail: wigler@cshl.org
Office: (516) 367-8376
FAX: (516) 367-8381

Key words: Physical Maps, Microarray Hybridization, Genome Sequence Assembly

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 a distance metric between probes, and then use this metric 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. We suggest conditions under which a "linear ordering" of the genome is not achievable.

Introduction

In theory, a genome can be sequenced and assembled into a linear map without resorting to any outside physical mapping information (Weber, J. et al., 1997; Venter, J.C., et al., 2001). These methods depend upon the recognition of sequence overlaps. In practice, deriving a complete and accurate map this way is not sensible. Any complex genome contains repeats, and these repeats, if longer than sequence reads, result in ambiguous assemblages. If the sequence reads do not cover the entire genome, 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.

In this paper we explore the feasibility of making high-resolution genome maps using micro-array hybridization, and using this data for sequence validation. We published a theoretical treatment of many of the ideas used here (Casey, W., et al., 2001), 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 metric between probes. From this metric 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 a half BAC length of each other, more often than not A and B will both hybridize to the same set of BAC pools. The degree of coincidence of their hybridization signals, over a large series of hybridization experiments, is statistically related to their actual distance in base pairs.

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.

We decided to test these ideas with actual experiments, and based our studies on the yeast S. pombe, because both good BAC libraries and a good sequence assembly were already available. The experiments themselves are expensive, and so pilot experiments with a small model organism is highly desirable. 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.

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

Results

Design of micro-array hybridizations.

DNA from BAC pools were made from a BAC library obtained from Pieter de Jong http://bacpac.chori.org/pombe104.htm. This library consisted of 3072 individual elements, with an average insert size of 160 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, R., et al., 1998).

Probes were designed to be relatively unique and to hybridize to high complexity representations. These representations are under-represented for the genome sequences in Sau3A1 fragments smaller than 200bp or larger than 1200 bp, and hence we designed 70-mer length oligonucleotide probes to reside within 200 to 1200 bp Sau3A1 fragments. We also required our probes to be unique sequence, and used exact mer-matching methods (Healy, J., 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 inter-probe 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, far in excess of the required threshold predicted by theory. Later, we "omitted" data to examine the quality of the resulting physical map assembly. The resulting set of 1224 probes were synthesized (Data set 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, in Cy5 and Cy3 label, 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 did 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.

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 (Data set 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 **Figure 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.

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 deviation of each population, enabling us to assign a probability to each probe 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 outcome of all hybridizations were thus compressed into a set of 1224 "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 (Data set B, see Materials and Methods). Note that the computation of the hit vectors requires no knowledge of the genome order inferred by sequence assembly.

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 d(A, B) as the sum of the absolute value of the differences between identical positions in each of the two vectors. Tabulating these values we obtain the 1224 by 1224 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. This number corresponds to the coverage of the probe in the BAC library. From the hamming distances and number of hits of two probes, we compute an estimate of the distance x between probes A and B using the formula:

(1) $$x = (d(\mathrm{A}, \mathrm{B})/Hits) * BacL * \exp(Hits/(4*NHybs))$$

where *Hits* are the combined number of hits of A and B, *BacL* is the mean BAC length, and *Nhybs* is the length of the hit vector (the number of hybridization experiments used) and *exp(X)* is *e* raised to the power *X*. We tabulate each pairwise estimate into a 1224 by 1224 matrix of distances, the "BDM" (BAC distance metric). The derivation of the formula is as follows:

Assume that the physical distance between two probes A and B is $x < BacL$. In addition to the hamming distance d(A,B), one can also define a coincidence value c(A, B) that measures the number of experiments in which both A and B get "hit". Note that d(A, B) + 2 c(A, B) = *Hits*(A) + *Hits*(B) = *Hits*. Furthermore, the following approximate estimates can be derived, $d(\mathrm{A}, \mathrm{B}) \propto 2x$ and $c(\mathrm{A}, \mathrm{B}) \propto BacL - x$, with the same constant

of proportionality. The intuitive argument is as follows: $pos(A) = pos(B) - x$, where *pos* denotes a linear coordinate position, and $pos(A) <= pos(B)$. Note that only in experiments where BACs from the pools have their left ends either in the interval, $[pos(A) - BacL, pos(B) - BacL]$, of length x, or in the interval, $[pos(A), pos(B)]$, also of length x, do we have a contribution to the function $d(A, B)$. Further, note that only in experiments where BACs from the pools have their left ends in the interval, $[pos(B) - BacL, pos(A)]$, of length $BacL - x$, do we get a contribution to the function $c(A, B)$. Thus, $d(A, B)/Hits = d(A, B)/[d(A, B) + 2\, c(A, B)] = x/BacL$ or

(2) $$x = (d(A,B)/Hits) * BacL.$$

This formula is a good approximation, and is correct if in a given BAC pool, no more than one BAC covers A or B. 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. We can correct the formulas for the expected values:

(3) $$d(A, B) = 2\, p\, s\, q * NHybs \text{ and } c(A, B) = [1 - s\,(1+p)q] * NHybs,$$

where $p + q = 1$, $r + s = 1$, $q = exp(-c\, x / BacL)$ and $s = exp(-c\,(BacL - x) / BacL)$, and where c is the coverage of the pool chosen in each experiment. q and s are simply the probabilities that no left end of any BAC appears in an interval of size x (e.g. $[pos(A), pos(B)]$) and in an interval of size $BacL - x$ (e.g.$[pos(B) - BacL, pos(A)]$), respectively. Thus,

(4) $$d(A, B)/Hits = 2\, p\, s\, q/(2 - 2\, s\, q) = p\, exp(-c)/(1 - exp(-c))$$
$$= (1 - exp(-c\, x / BacL))\, exp(-c)/(1 - exp(-c)).$$

After appropriate simplification we have:

(5) $$x - x^2\, c / (2\, BacL) + o(c^2) = (d(A, B)/Hits) * BacL * exp(c/2).$$

The rest follows from the following local estimation of c as $Hits/(2 * NHybs)$. For small c and $x < BacL$, all but the first term on the left hand side can be ignored; thus, making the right hand expression a good estimate of the inter-probe distance, x. Experimental validation of this formula can be seen from **Figure 2**.

Given the estimates of distances from our hybridization data alone we can begin to derive a physical map, and compare to the map inferred from the sequence assembly.

Assembling the probes into a graph.

Given a matrix of pair-wise 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, Ph.D. thesis, and Casey, Ph.D. thesis). 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. With real data, we found it impossible to derive an unambiguously correct linear ordering, and hence we used a more complex geometric structure into which we embed the distance relationship of our probes.

The ordering algorithm we have chosen involves constructing the minimum spanning tree between our probes. The minimum spanning tree is an acyclic graph that joins all neighboring probes by the shortest possible path. The method 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. It can be proven that this results in the same acyclic graph, no matter the starting point. This is known as Prim's algorithm (Cormen, T.H. et al., 2001). We then extract the longest linear path containing the greatest number of probes.

The result of this method, applied to probes from the S. pombe chromosome 1, is shown in **Figure 3**, in which the output of our algorithm is plotted using GraphViz, a set of graph drawing tools, (at http://www.research.att.com/sw/tools/graphviz/). 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 blow-ups. The three isolated probes that form their own 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 number 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.

Comparison of 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 inequality (i.e., one of the form: if A<B<C then AB +BC < AB + AC, etc.) and the minimum spanning tree is a single contig with all the probes in correct order. However, in real experiments, these conditions are not met throughout, and 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 **Figure 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 1224). We note that probes from the telomeres of different 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 **Figure 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 **Figure 2 panel A** on a full scale of all pair-wise probe distances, and **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 metric. The majority of these are telomeric and centromeric probes, or probes that fall into regions that have 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 pair-wise 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 neighbor 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 **Figure 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 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 (**Figure 3**), as aberrant edge connections (**Figure 4**), discordance between the BAC mapping distance and sequence assembly distance (**Figure 2**), and the pattern of BAC pool hybridization of sequence assembly neighbors (**Figure 5**). These are presumably all related, and to test this, we created a new pair-wise BAC distance matrix by removing the handful of probes that were judged to have distorted BAC hybridization in their neighborhood (by the criteria illustrated in **Figure 5**). We then recomputed the minimal spanning tree, and plotted the resulting edge connections (**Figure 4 panel B**), and plotted again the comparison of the BAC distance metric to the sequence assembly metric (**Figure 2 panels C and D**). Not surprisingly, the most extreme discordances are thereby removed.

The edges computed for the spanning trees (Data set "E") are available from our web site: (XXX), as are our files of pair-wise distances (Data set "D").

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 metric, 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 metric 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 never the less 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 make less than perfectly reliable substrates for hybridization, and oligonuclotide synthesis is expensive. Now, microarrays with very uniform character and with any desired oligonucleotide probe design can be fabricated by laser directed in situ synthesis (NimbleGen Systems, Inc.). Although still not cheap, reproducibility is increased. Relative to the costs of assembly, the costs of physical mapping by array hybridization are minor.

Materials and Methods

Microarrays.

We used the Cartesian PixSys 5500 arrayer to array our probe collection onto commercially prepared silanated glass slides. Each probe was spotted 5 times at random locations on the slide. This was done to control for any geometric or geographic artifacts on the array that was present on the slide itself before printing or that was induced by the processing of the slide during the hybridization or post processing steps.

Probe design.

Our probes are 70 base-pair long oligonucleotides (70-mer) derived from short (200-1200 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, J., et al., 2003). The complete genome sequence of *Schizosaccharomyces pombe* is available for download from the website of The Wellcome Trust Sanger Institute http://www.sanger.ac.uk/Projects/S_pombe. The genomic DNA sequence of *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 1200 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 1224 70-mer fragments so that the midpoint of each fragment was on average 10kb from the midpoint of any of its neighbors to the left and to the right. These 1224 70-mer fragments are what we refer to as our probes (Data set "C").

BAC pools.

The *S. pombe* BAC library has 3072 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. 16 rows per plate x 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 of 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 5ml 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 1000 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.

Representations.

BAC pool representations were prepared as described in Lucito, R., 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, resuspended, and the concentration determined. This material was then used as template in the PCR reaction.

Labeling of representations.

Ten micrograms of representation was denatured by heating to 95°C in the presence of 5 µg random nonamer in a total of 100 µl. After 5 minutes the sample was removed from heat and 20 µl of 5X buffer was added (50mM Tris-HCl [pH 7.5], 25mM $MgCl_2$, 40mM DTT, suspended with 33 µM dNTPs), 10nmol of either Cy3 or Cy5 was

added, and 5 units of Klenow fragment. After incubation of the reaction at 37°C for 2 hours, the reaction 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 μl, at a concentration of 3X SSC and 0.3% SDS, denatured and then hybridized to the array of probes.

Hybridization of representations to microarrays.

Hybridization solution for printed slides consisted of 25% formamide, 5 X SSC, 0.1%SDS. 25ul of hybridization solution was added to the 10ul of labeled sample and mixed. Samples were denatured in a MJ Research Tetrad at 95°C for 5 mins, and then incubated at 37°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.

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 6120 features.

Data pre-processing.

We then calculated the log (Cy5/Cy3) for each of the 6120 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 6120 rows by 128 columns (Data set "A"). Since each probe was printed in quintuplicate, we then calculated the median log ratio over the 5 replicates for each probe and used this value as the value for that probe in that particular hybridization. This condensed our data matrix to 1224 rows (each row representing a single probe), and 128 columns (each column representing a particular hybridization or BAC pool). The final step in the pre-processing 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 1224 by 128, containing values between 0 and 1 (Data set "B"). As described in the text, the computation of physical distances (using equation 1) is accomplished using Data set B.

Data availability.

The Data sets 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 from this site: (XXX).

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 (contract #: F30602-01-2-0556). Additional support was provided by NSF's

Qubic and two ITR programs, the US Department of Energy, and New York State Office of Science, Technology & Academic Research.

Figure Legends

Figure 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.

Figure 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 metric, BDM) computed from equation 1 on the Y-axis, and the sequence assembly distance metric (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 metric 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 **Figure 5**).

Figure 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 are 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.

Figure 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 are 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 **Figure 5**) was recomputed, and the edge connections are shown in **Panel B**.

Figure 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. These parameters, "AllHits", "SingleHit", and "LonelyMiss" are explained in the text.

References

Casey, W., Mishra, B. and Wigler, M. (2001). Placing probes along the genome using pairwise distance data. Algorithms in Bioinformatics, O. Gasscuel and B.M.E. Moret (eds.): First International Workshops, Arhus, Denmark. Springer-Verlag Berlin Heidelberg WABI 2001, LNCS, **2149***:* 52-68.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C. 2001. Introduction to Algorithms. Chapter VI: Graph Algorithms. Page 505. MIT Press under a joint Production-distribution agreement with the McGraw-Hill Book Company.

Healy, J., Thomas, E.E., Schwartz, J.T., and Wigler, M. 2003. Annotating large genomes with exact word matches. Genome Research **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. U S A **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.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., 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., Li, P.W., Mural, R.J., Sutton, G.G., Smith,H.O., Yandell, M.,Evans,C.A., Holt, R.A., 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 Research, **7**: 401-409.

FIGURE 1: Representative Data from a Single Hybridization

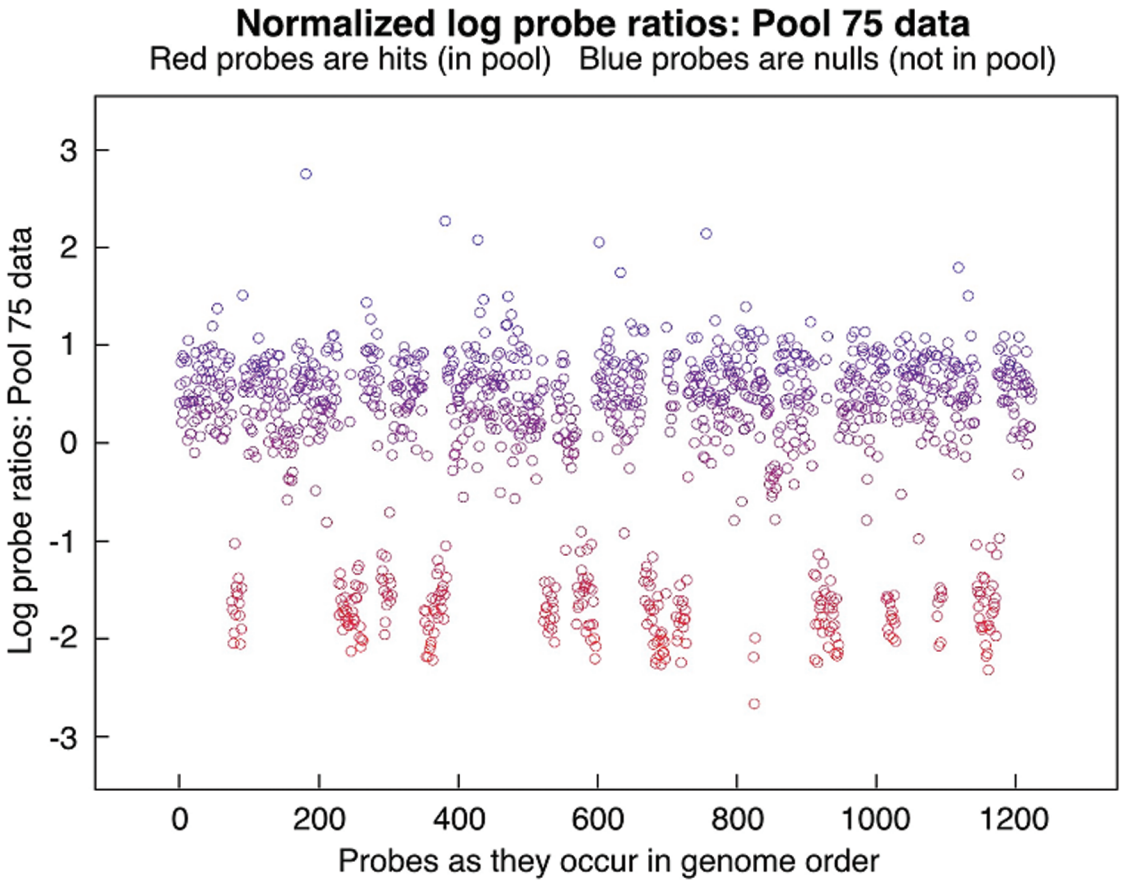


FIGURE 2: Computed Physical Distance Compared to Sequence Assembly

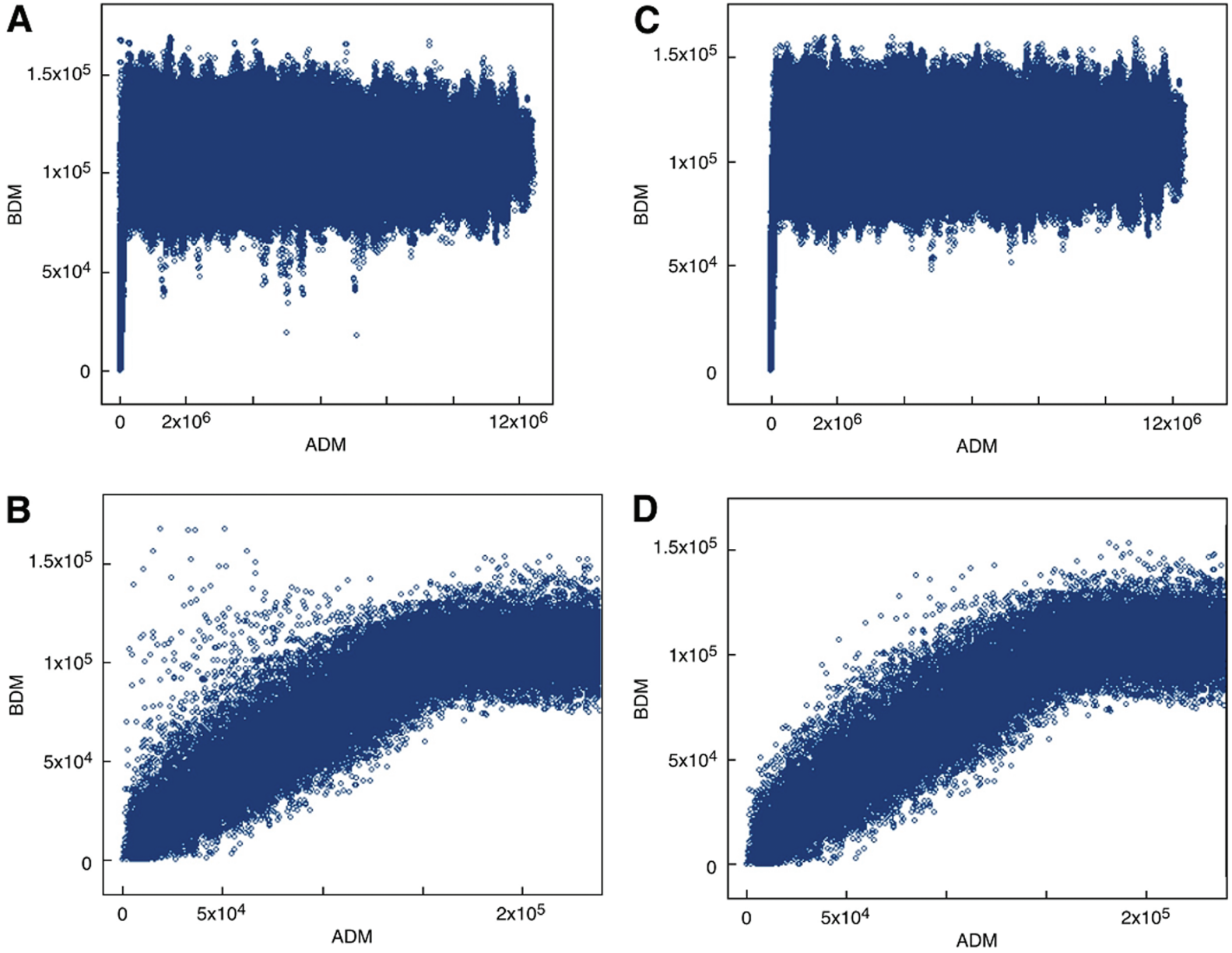


FIGURE 3: Graph of Minimal Spanning Tree, *S.pombe*, Chromosome 1

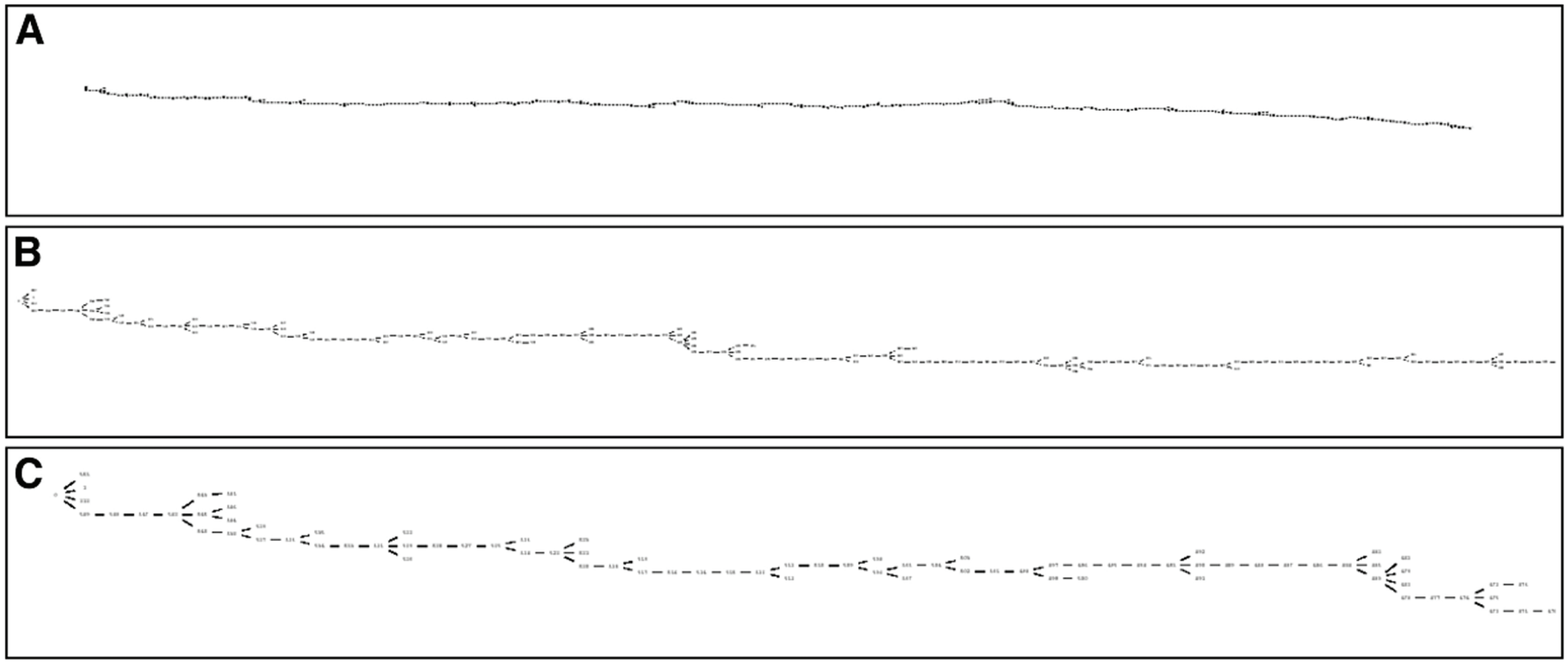
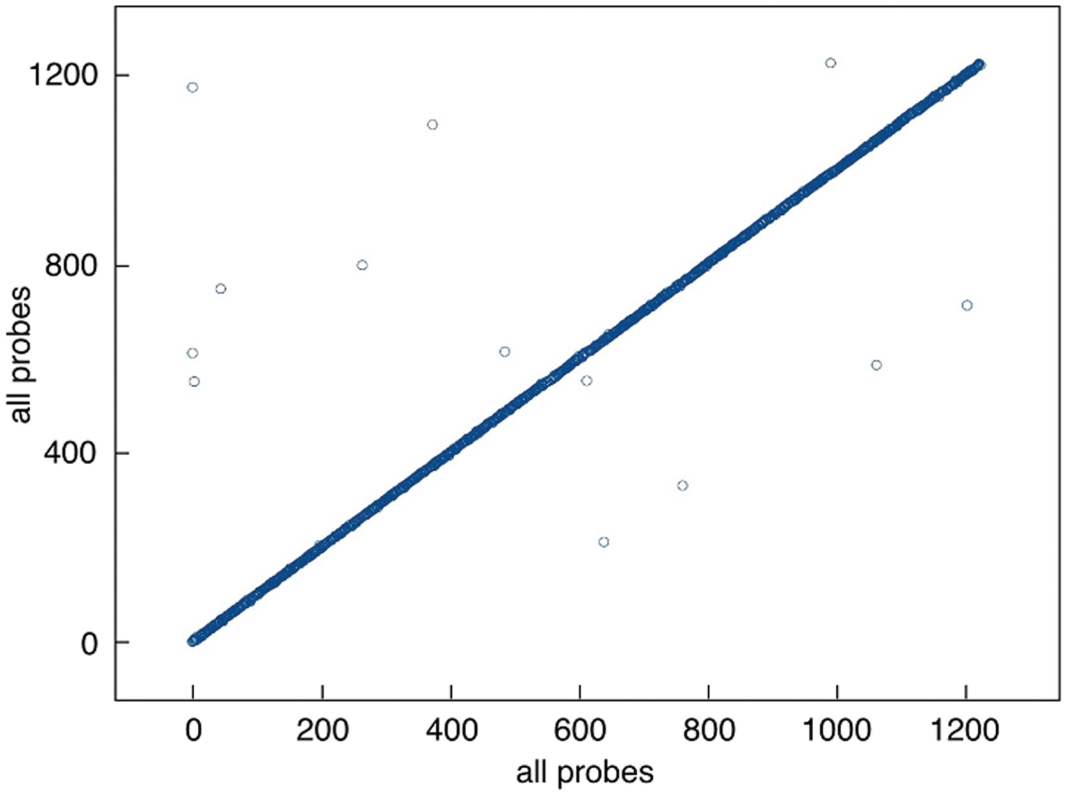
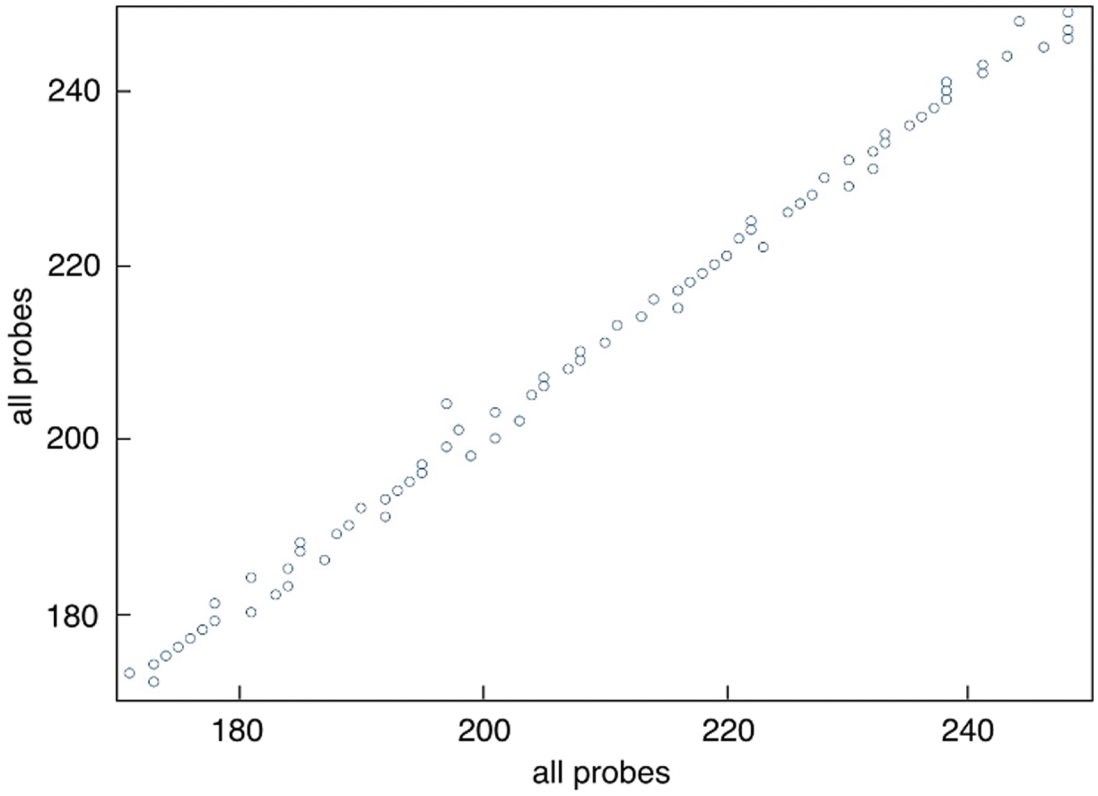


FIGURE 4: Edge Coordinate Pairs from Minimal Spanning Trees



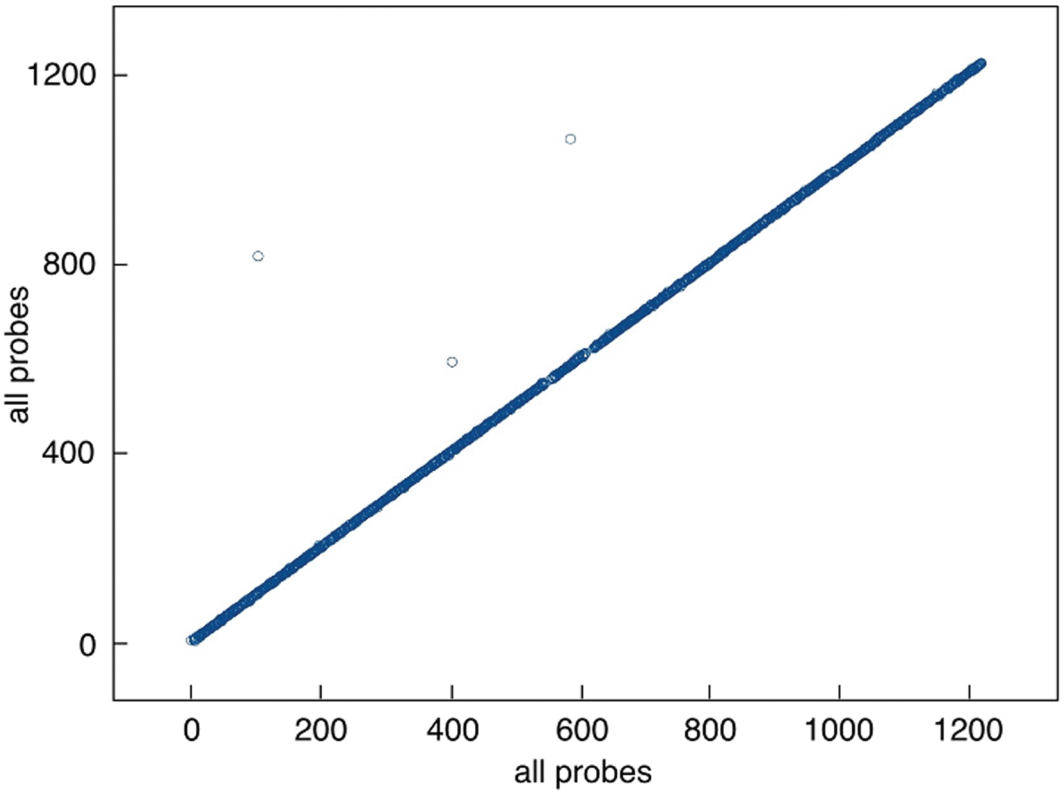
Panel A: Edge Coordinate Pairs, all Probes Whole Genome

FIGURE 4: Edge Coordinate Pairs from Minimal Spanning Trees



Panel B: Edge Coordinate Pairs, "Well-Behaved" Probes Whole Genome

FIGURE 4: Edge Coordinate Pairs from Minimal Spanning Trees



Panel C: Edge Coordinate Pairs, Close-Up from Chromosome 1

FIGURE 5: BAC Hybridization Patterns Across Probes Displayed in Genome Assembly Order