Genome Mapping by Random Anchoring: A Discrete Theoretical Analysis

M. Q. Zhang[1] and T. G. Marr[1]

As a part of the international human genome project, large-scale genomic maps of human and other model organisms are being generated. More recently, mapping using various anchoring (as opposed to the traditional "fingerprinting") strategies have been proposed based largely on mathematical models. In all of the theoretical work dealing with anchoring, an anchor has been idealized as a point on a continuous, infinite-length genome. In general, it is not desirable to make these assumptions, since in practice they may be violated under a variety of actual biological situations. Here we analyze a discrete model that can be used to predict the expected progress made when mapping by random anchoring. By virtue of keeping all three length scales (genome length, clone length, and probe length) finite, our results for the random anchoring strategy are derived in full generality, which contain previous results as special cases and hence can have broad application for planning mapping experiments or assessing the accuracy of the continuum models. Finally, we pose a challenging nonrandom anchoring model corresponding to a more efficient mapping scheme.

1. INTRODUCTION

A complete set of ordered DNA clones spanning the genome of an organism provides a very powerful template for further genomic analysis.[2]

---

[1] Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, New York 11724, U.S.A.

[2] A genome is a long chain of DNA molecule which contains all the genetic information of an organism. To analyze a genome, biologists often randomly cut the DNA into many small pieces by restriction enzymes and insert each piece into the DNA of a simple organism (called a vector), then maintain and duplicate (clone) these vectors inside bacteria host cells. A set of clones carrying random pieces of a large genome is called a library. By mapping, one orders the clone pieces in a library and finds the distances among them according to their original genomic positions. A clone may be thought of as a duplicate of a DNA fragment; those with relative distances determined according to their genome positions are called ordered clones.

It would reveal the relationship between physical and recombination distances throughout the genome and, once aligned with the genetic map[3] of the organism, would greatly facilitate map-based cloning of genes for which only a mutant phenotype and linkage map position are known. Complete or nearly complete physical maps[4] of clones have already been constructed for genomes of *Escherichia coli*,[2] *Saccharomyces cerevisiae*,[3] and *Schizosaccharomyces pombe*.[11] Numerous efforts are underway to construct physical maps of considerably larger genomes, such as the human and mouse genomes.

There are two major schemes used to order the clones. One is known as "fingerprinting," which consists in detecting overlapping clones using restriction fragment length patterns (called "fingerprints"); the other is known as "anchoring," which consist in detecting shared unique sequences (called "anchors").

Theoretical considerations of planning mapping projects based on fingerprinting were published initially in this area by Lander and Waterman.[4] Analogous analyses for projects based on anchoring have also appeared recently.[5–7,1] In all these works dealing with anchoring, an anchor was idealized as a point on a continuous genome. This approximation is good when the size of the probes (a probe recognizes a specific anchor sequence; it is for our purpose synonymous to an anchor) is much smaller than the size of the clones. Recently, an approximate discrete analysis has also been published[8] where the authors were not able to derive closed-form formulas for short genome length. Theoretically, it is desirable not to make these assumptions (zero-length anchors and a continuous, infinite-length genome) if we can still solve the problem. More importantly, in practice, these assumptions may sometimes break down. Although these assumptions are applicable in many situations, we believe it is still useful to provide the exact results for those who may only want to map a particular segment of a genome (where the continuous approximation may not be applicable) or for those who use probes with length comparable to that of clones.

From an experimental standpoint, mapping the genome will take place in two stages: first, by building islands of linked clones with all available probes, and second, by linking these islands into a contiguous map by bridging the gaps between them. The theoretical analysis presented in this paper can answer some of the statistical questions in the first stage of

---

[3] A genetic (linkage) map is a map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together.

[4] A physical map is a set of overlapping clones with distances measured by base pairs.

contig (island) building by random anchoring. Since the random anchoring scheme is inefficient experimentally, we shall also discuss a more efficient nonrandom model at the end.[9]

2. EXPECTED NUMBER OF ISLANDS

It is assumed that a library of clones and collection of probes are generated independently, randomly, and uniformly from the genome of size G. Let c and s be the densities of the clones and the probes, respectively. Thus, $c = N_c/G$ and $s = N_s/G$, where N_c and N_s are the total number of clones and probes. If the lengths of clones and probes are denoted by L and M, then the parameters a (the redundancy of coverage) and b (the average number of probes contained in a random clone) are given by $a = cL$ and $b = sM$.

First, we calculate the expected number of islands. For simplicity of notation, we define

$$\bar{c} \equiv 1 - c \qquad \text{and} \qquad \bar{s} \equiv 1 - s$$

We assume $L \geqslant M$. For a clone to be anchored, there has to be at least one probe with its entire length embedded within the clone.

The number of islands is of course equal to the number of the left ends of islands. The event that there is a left end of an island starting (from the left to the right) at some arbitrary position 0 is characterized by the following: there is a clone starting at 0, the first probe should start at i for $i = 0, 1,..., L - M$, and there should be no clones starting to the left of 0 and anchored by the probe (see Fig. 1). Hence, the probability p an island starts at 0 is given by

$$p = \sum_{i=0}^{L-M} \bar{c}^{L-M-i} c \bar{s}^i s \tag{1}$$

Simple summation of the geometric series yields

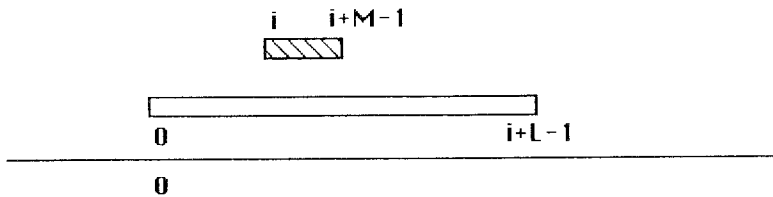$$p = cs \frac{\bar{s}^{L-M+1} - \bar{c}^{L-M+1}}{c - s} \tag{2}$$

Fig. 1.

The expected number N_{island} of island is given by Gp. It is clear from these results that the expected number of islands depends only on the relative length $L - M$ between the clones and the probes [see Eq. (2)].

Having gotten the general result, we would also like to inspect the continuous limit defined by

$$c, s \to 0 \qquad \text{with} \quad cL, sL \to a, b \tag{3}$$

We have

$$N_{\text{island}} \to N_c b \, \frac{e^{-b(1-t)} - e^{a(1-t)}}{a-b} \tag{4}$$

where

$$t \equiv M/L$$

In the continuous limit, the result of Eq. (4) differs from the zero-probe-length case only by a trivial scale factor $1 - t$, which means if we plot N_{island} (in units of G/L) against $(1-t)b$ for different values of $(1-t)a$, we would get the same figure as in the $t = 0$ case. This is shown in Fig. 2. Therefore, reducing the ratio t is equivalent to increasing the coverage a and b.
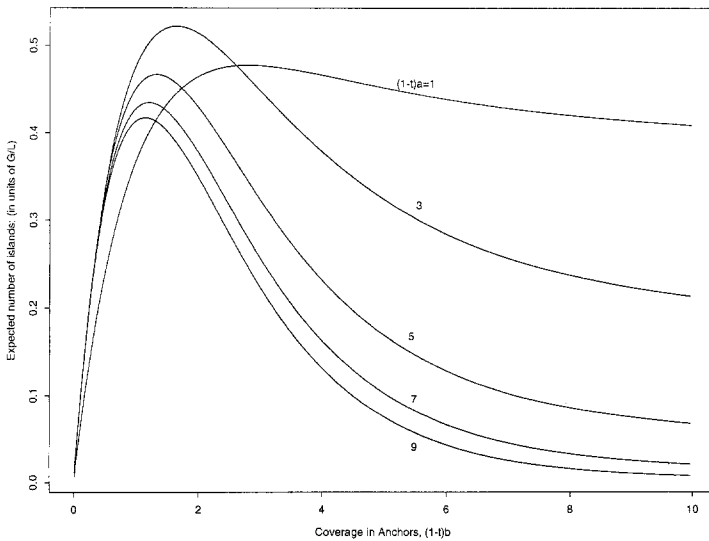

Fig. 2. The expected number of islands (in units of G/L) vs. coverage in anchors $(1 - t)t \equiv$ $(1 - M/L)(N_p L/G)$ for various coverages in clones $(1 - t)a \equiv (1 - M/L)(N_c L/G)$. Increasing the ratio t of the probe length over the clone length is equivalent to decreasing both the probe numbers and the clone numbers.
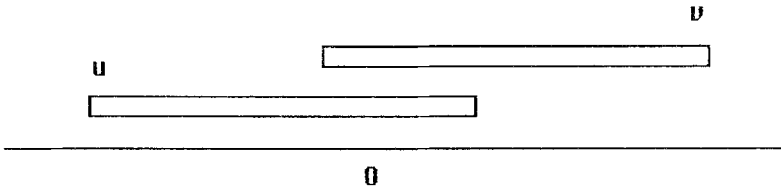
Fig. 3.

3. EXPECTED COVERAGE OF THE GENOME BY ISLANDS

Second, we calculate the expected fraction of genome length covered by islands, which should serve as a measure of the progress of a mapping project. If we define r_i for $i = 0, 1, ..., m$ (if $L > 2M$, $m = 4$)[5] to be the probability that an arbitrary base pair position is covered by i islands, it is obvious, by definition, that

$$r_0 + r_1 + R_2 = 1 \qquad \text{where} \qquad R_2 \equiv \sum_{i \geqslant 2} r_i \tag{5}$$

To compute μ, the expected fraction of genome length covered by the islands, it is easiest to compute r_0, the expected fraction not covered by any islands first.

Take an arbitrary position 0; then the probability that 0 is not covered by any clones is $\bar{c}^L$ and the probability that 0 is covered by one unanchored clone is $Lc\bar{c}^{L-1}\bar{s}^{L-M+1}$. To calculate the probability q that 0 is covered by more than one unanchored clone, we assume (see Fig. 3) the left end of the leftmost anchored clone covering 0 is at u (which means $1 - L \leqslant u \leqslant -1$) and the right end of the rightmost anchored clone covering 0 is at v (which means $u + L \leqslant v \leqslant L - 1$); therefore q is given by

$$q = \sum_{u=1-L}^{-1} \sum_{v=u+L}^{L-1} c^2 \bar{s}^{v-u-M+2} \bar{c}^{2L-2+u-v}$$

$$= c^2 \bar{s}^{L-M+2} \frac{(L-1)(s-c)\bar{c}^{L-1} + \bar{s}(\bar{s}^{L-1} - \bar{c}^{L-1})}{(c-s)^2}$$

Summing up these three possibilities, we get

$$r_0 = \bar{c}^L + Lc\bar{c}^{L-1}\bar{s}^{L-M+1} + q \tag{6}$$

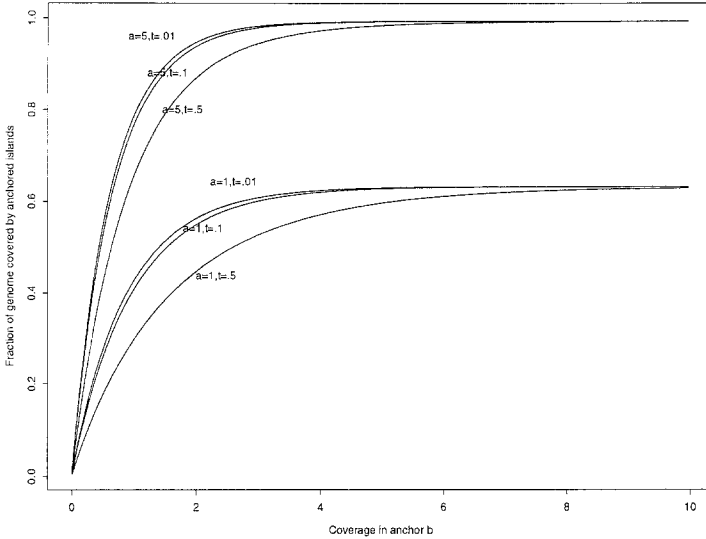[5] We thank Prof. Q. Yu for pointing this out to us.

Fig. 4. The expected fraction of genome coverage by islands vs. coverage by islands vs. coverage in anchors $b \equiv N_p L/G$ for various clone coverage $a \equiv N_c L/G$ and length ratios $t \equiv 1 - M/L$. For a fixed clone coverage a, increasing the length ratio t reduces the fraction of genome coverage, with its largest effect occurring about $b = 2$.

which, in the continuous limit, becomes

$$r_0 \rightarrow e^{-a} + ae^{-(a + (1-t)b)} - \frac{a^2(a-b+1)}{(a-b)^2} e^{-(a+(1-t)b)} + \frac{a^2}{(a-b)^2} e^{-(2-t)b} \quad (7)$$

The fraction $(r_1 + R_2)$ of genome covered by islands is, of course, $1 - r_0$. This is plotted in Fig. 4 for the continuous limit; we see that reducing the ratio t would increase the effective coverage of the genome by islands. Previous results maybe obtained by setting $t = 0$.

4. EXPECTED FRACTION OF GENOME COVERED BY MORE THAN ONE ISLAND

In this section, we first calculate the probability R_2 (defined above) that an arbitrary position 0 is covered by more than one island. Due to the discrete nature of the problem, it turns out that this calculation is very tedious. But the idea is fairly simple and each computation is never more than a straightforward evaluation of some geometric series. We shall illustrate the idea and omit intermediate steps.

In order for the position 0 to be covered by at least two islands (see Fig. 5) (a) there has to be a rightmost clone of the left island, starting at
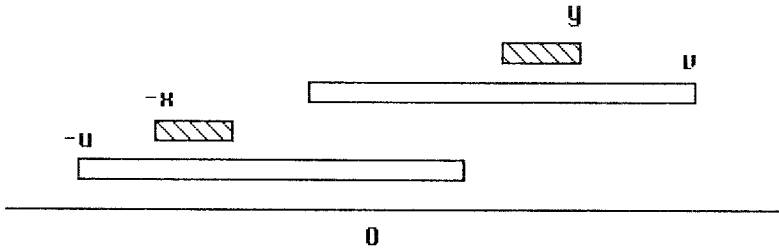
Fig. 5.

$-u$ and covering 0; (b) there has to be a leftmost clone of the right island, ending at v and covering 0; (c) there has to be a rightmost probe of the left island, starting at $-x$ and anchoring the rightmost clone of the left island, but not anchoring the right island; (d) there has to be a leftmost probe of the right island, ending at y and anchoring the leftmost probe of the right island, but not anchoring the left island. The positions of the two probes can vary according to $1 \leqslant x, y \leqslant L - 1$, which has to be subjected to the condition $x + y \geqslant M$ because they cannot overlap themselves. Once the probe positions are fixed, the rightmost clone of the left island can only vary according to $m1 \equiv \max(x, L - y) \leqslant u \leqslant m2 \equiv \min(L - 1, x + L - M)$, which specifies that it contains the left probe but not the right and it covers the proposition 0; similarly, the leftmost clone of the right island can only vary according to

$$m3 \equiv \max(y, L - x) \leqslant v \leqslant m4 \equiv \min(L - 1, y + L - M)$$

by symmetry. Therefore, the probability that the position 0 is covered by two islands can be expressed as the following sum:

$$R_2 = c^2 s^2 \sum_{\substack{x, y = 1 \\ M \leqslant x + y}}^{L-1} \sum_{u = m1}^{m2} \sum_{v = m3}^{m4} \bar{s}^{x + y - M} \bar{c}^{v + u - L - \max(0, x + y - L)} \tag{8}$$

The factors in front of the summation say that there are two clones and two probes. The first factor in the summand says that there can be no other probes in between the two probes, because those two probes are the outmost probes of the two islands, by our assumption. The second factor in the summand says that there can be no other clones in between the two clones if they cannot be fit in between the two probes, otherwise they would either become the outmost clones of the islands or link the two islands, hence contradicting the assumption. The singularity domains in the x, y summation plane are indicated in Fig. 6, they depend on the lengths
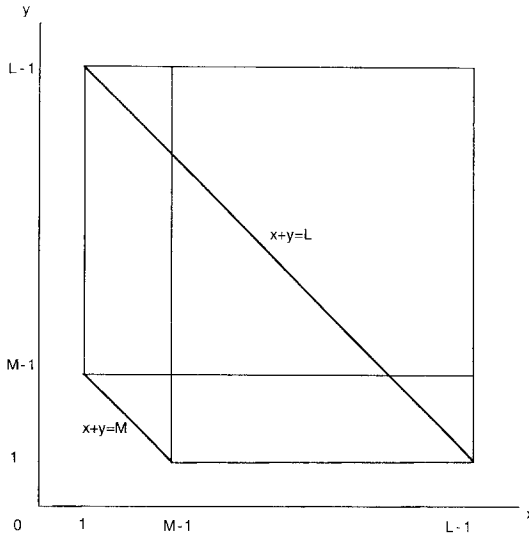
Fig. 6.   The singularities in the (x, y) domain (i.e., the summand has to change across each singularity line).

L and M. To avoid the pathological complications, we assume $L > M > 2$; then, after carrying out the u, v sums, we are left with

$$R_2 = \bar{s}^2 \sum_{\substack{x, y = 1 \\ M \leqslant x + y < L}}^{L-1} \frac{\bar{s}^{x+y-M}}{\bar{c}^M} (\bar{c}^{L-y} - \bar{c}^{m2+1})(\bar{c}^{L-x} - \bar{c}^{m4+1})$$

$$+ \frac{s^2}{\bar{s}^M} \sum_{\substack{x, y = 1 \\ L \leqslant x + y}}^{L-1} \left(\frac{\bar{s}}{\bar{c}} \right)^{x+y} (\bar{c}^x - \bar{c}^{m2+1})(\bar{c}^y - \bar{c}^{m4+1})$$

$$= S_1 + S_2 + S_3 + S_4 + S_5$$

This sum has been decomposed into five terms as indicated above; their definitions and the results are given in the Appendix.

In the continuous limit, we find (remembering that $0 < t = M/L < 1/2$)

$$R_2 \to b \left(\frac{2a - b + t(3ab - 2a^2 - b^2)}{(a-b)^2} + \frac{2a + b + bt(a+b)}{(a+b)^2} \right) e^{-(1-t)a}$$

$$+ \frac{a}{(a+b)^2} e^{-(a+tb)}$$

$$+ a\frac{ab - b^2 - a}{(a-b)^2}e^{-(1-t)b} + \frac{a^2}{(a-b)^2}e^{-(2-t)b} + 2e^{-(1-t)(a+b)}$$

$$- 2\frac{a^2}{(a-b)^2}e^{-((1-t)a+b)}$$

$$+ \frac{ab^2 - a^2 - a^2b + 4ab - 2b^2 + 2abt(a-b)}{(a-b)^2}e^{-(a+(1-t)b)} \tag{9}$$

With R_2 calculated, the expected fraction of genome covered by more than one island is GR_2 [the expected fraction of genome covered by exactly one island is equal to $G(1 - r_0 - R_2)$]. They agree with the previous results when $t = 0$.

5. EXPECTED SIZE OF AN ISLAND

To calculate the mean size of an island, we use the approach of Ewens *et al.*,[7] which begins with the following two calculations:

5.1. Mean Distance between Probes on the Same Island d_1

Suppose that the right end of a probe is at 0 and the right end of the rightmost clone anchored by the probe is at y; conditioning on this event, the probability that the right end of a right nearest neighbor probe is at x is proportional to $\bar{s}^x\bar{c}^{-y}$. Therefore, the mean distance d_1 between probes on the same island may be calculated as

$$d_1 = \frac{\sum_{0 \leqslant x \leqslant y \leqslant L-M} x\bar{s}^x\bar{c}^{-y}}{\sum_{0 \leqslant x \leqslant \leqslant L-M} \bar{s}^x\bar{c}^{-y}} = \frac{A}{B} \tag{10}$$

where

$$A = \frac{(\bar{s}/\bar{c})[1 - (\bar{s}/\bar{c})^{L-M+1}]}{[1 - (\bar{s}/\bar{c})]^2} - \frac{(L-M+1)(\bar{s}/\bar{c})^{L-M+1}}{1 - (\bar{s}/\bar{c})}$$

$$- \bar{c}^{M-L-1}\left[\frac{\bar{s}}{s^2}(1 - \bar{s}^{L-M+1}) - (L-M+1)\frac{\bar{s}^{L-M+1}}{s}\right]$$

$$B = \frac{1 - (\bar{s}/\bar{c})^{L-M+1}}{1 - (\bar{s}/\bar{c})} - \frac{\bar{c}^{M-L-1}}{s}(1 - \bar{s}^{L-M+1})$$

5.2. Mean Distance from the Rightmost Probe on an Island to the Right-Hand End of the Island d_2

This is calculated in a similar fashion and results in

$$d_2 = \frac{\sum_{0 \leqslant y \leqslant L - M} y c \bar{c}^{L - M - y} s \bar{s}^y}{\sum_{0 \leqslant y \leqslant L - M} c \bar{c}^{L - M - y} s \bar{s}^y}$$

$$= \frac{1}{(\bar{c}/\bar{s}) - 1} - \frac{L - M + 1}{(\bar{c}/\bar{s})^{L - M + 1} - 1} \tag{11}$$

Any island will have some number i ($i \geqslant 1$) of probes on it, and thus $i - 1$ "interprobe" distances. It will also have two other distances, to the left (right) of the leftmost (rightmost) probe. The mean island size l, using d_1 and d_2 above, is thus

$$l = d_1 \mathbf{E}(i - 1) + 2d_2 + M - 1 \tag{12}$$

Now the mean of i is the mean number of probes on islands $[N_s(1 - \bar{c}^{L - M1})]$ divided by the mean number of islands, N_{islands}, as before. So we may write Eq. (12) as

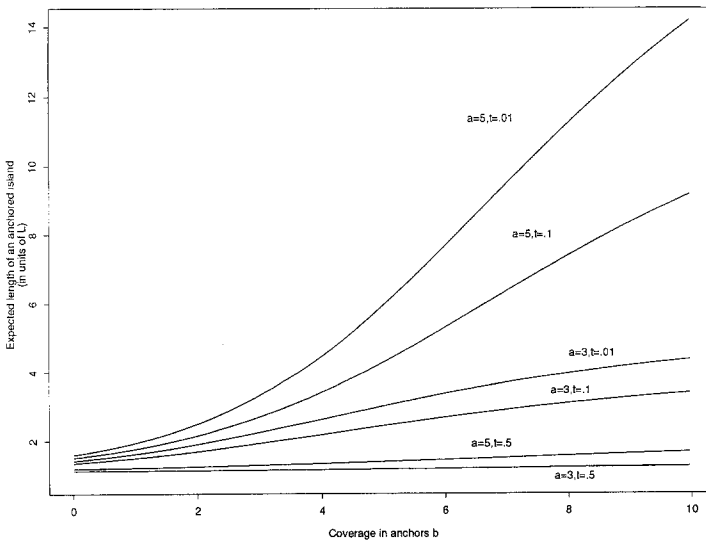$$l = d_1 \left(\frac{N_s(1 - \bar{c}^{L - M + 1})}{N_{\text{islands}}} - 1 \right) + 2d_2 + M - 1$$

Fig. 7. The expected length of an island (in units of L) vs. coverage in anchors $b \equiv N_p L/G$ for various coverages in clones $a \equiv N_c L/G$ and the lengths $t \equiv 1 - M/L$. For a fixed clone coverage a, increasing the length ratio t reduces the average island length.

The mean island size in the continuous limit is plotted in Fig. 7, it is obvious that reducing the ratio t results in increasing the length of islands. Previous results correspond to the case of $t = 0$.

6. COMMENTS

The mapping strategy discussed in this paper is a random one. We have assumed fixed clone and anchor sizes which correspond to the average sizes in real applications. In principle, one may assume Gaussian distribution of these sizes, and obtain analogous results as was done in the continuum case.[6] In order to improve the efficiency, recently a new non-random strategy has been proposed, tested in computer simulations, and implemented in real experiments.[10,11] In this nonrandom strategy, instead of selecting anchors randomly from the genome, anchors are selected from both ends of an unlinked clone in a sequence way. Ideally it works as follows: (1) generate a library of random clones—DNA pieces of size L which can cover the genome five times on average; (2) take an arbitrary clone, make two proves of size M from its two ends, and link some other clones in the library by detecting whether they share the anchor sequences; (3) repeat the procedure 2 only with unlinked clones until all the clones are linked. In this way, islands can be built up very rapidly. Here the distribution of clones and probes are highly correlated, although an approximate model[9] has been solved by using the hard-rod statistics[12] (the random strategy corresponds to a mixture of idea gases) which agreed well with simulation and experiments in the applicable region; it is still a great mathematical challenge to formulate the exact model rigorously.

APPENDIX

$$S_1 \equiv s^2 \sum_{\substack{x, y = 1 \\ M \leqslant x+y < L}}^{L-1} \frac{\bar{c}^L}{\bar{s}^M} \left(\frac{\bar{s}}{\bar{c}}\right)^{x+y}$$

$$= -s^2 \bar{c} \frac{(M-1)\bar{c}^{L-M} - (L-1)\bar{s}^{L-M}}{c-s} + s^2 \bar{c}\bar{s} \frac{\bar{c}^{L-M} - \bar{s}^{L-M}}{(c-s)^2}$$

$$S_2 \equiv -2s^2 \sum_{\substack{x, y = 1 \\ M \leqslant x+y}}^{L-1} \bar{s}^{x+y-M} \bar{c}^{m2}$$

$$= -2s(M-1)\bar{c}^{L-M+1} + 2(\bar{c}\bar{s})^{L-M+1}(1 - \bar{s}^{M-1})$$

$$+ 2s\bar{c}\bar{s}(1 - \bar{s}^{L-1}) \frac{\bar{c}^{L-M} - \bar{s}^{L-M}}{c-s}$$

$$S_3 \equiv s^2 \sum_{\substack{x,\,y\,=\,1 \\ M \leq x+y < L}}^{L-1} \bar{s}^{x+y-M} \bar{c}^{L+\min(0,x-M+1)+\min(0,y-M+1)}$$

$$= s^2 \bar{c}^{L+2} \bar{s}^M \left(\frac{(M-1)(\bar{c}\bar{s})^{-M}}{1-\bar{c}\bar{s}} + \frac{1-(\bar{c}\bar{s})^{1-M}}{(1-\bar{c}\bar{s})^2} \right)$$

$$+ s\bar{c}^L \bar{s}^M [1-(L-2M+1)\bar{s}^{L-2M}]$$

$$+ \bar{c}^L \bar{s}^{M+1}(1-\bar{s}^{L-2M}) + 2s\bar{c}^{L-M+1} \frac{\bar{c}\bar{s}-(\bar{c}\bar{s})^M}{1-\bar{c}\bar{s}}$$

$$- 2s\bar{s}^{L-M} \frac{\bar{c}^{L-M+2}-\bar{c}^{L+1}}{c}$$

$$S_4 \equiv s^2 \sum_{\substack{x,\,y\,=\,1 \\ L \leq x+y}}^{L-1} \bar{s}^{x+y-M}$$

$$= s(L-1)\,\bar{s}^{L-M} - \bar{s}^{L-M+1}(1-\bar{s}^{L-1})$$

$$S_5 \equiv s^2 \sum_{\substack{x,\,y\,=\,1 \\ L \leq x+y}}^{L-1} \bar{s}^{x+y-M} \bar{c}^{\min(L-x,L-M+1)+\min(L-y,L-M+1)}$$

$$= s^2 \bar{c}^{L+1} \bar{s}^{L-M} \left(-\frac{L-1}{c-s} + \bar{s}^L \frac{\bar{c}^{1-L}\bar{s}^{1-L}}{(c-s)^2} \right) - 2\bar{c}^{L-M+3}\bar{s}^{L-M} \frac{s^2(1-\bar{c}^{M-1})}{c(c-s)}$$

$$+ 2s\bar{c}^{L-M+2}\bar{s}^{L-M+1} \frac{1-\bar{s}^{M-1}}{c-s} + 2s^2\bar{c}^{L+1}\bar{s}^{L-M} \frac{M-1}{c-s}$$

$$+ 2s^2\bar{c}^{L-M+2}\bar{s}^{L-M+1} \frac{\bar{c}^{M-1}-\bar{s}^{M-1}}{(c-s)^2}$$

ACKNOWLEDGMENTS

REFERENCES

1. R. Arratia, E. S. Lander, S. Tavaré, and M. S. Waterman, *Genomics* **11**:806 (1991).
2. Y. Kohara, K. Akiyama, and K. Isono, *Cell* **50**:495 (1987).
3. M. V. Olson, J. E. Dutchik, M. Y. Graham, G. M. Brodeur, C. Helms, M. Frank, M. Maccollin, R. Scheinman, and T. Frank, *Proc. Natl. Acad. Sci. USA* **83**:7826 (1986).
4. E. S. Lander and W. S. Waterman, *Genomics* **2**:231 (1988).

5. E. Barillot, J. Dausset, and D. Cohen, *Proc. Natl. Acad. Sci. USA* **88**:3917 (1991).
6. D. Torney, *J. Mol. Biol.* **217**:259 (1991).
7. W. J. Ewens, C. J. Bell, P. J. Donnelly, P. Dunn, E. Matallana, and J. R. Ecker, *Genomics* **11**:799 (1991).
8. T. G. Marr, X. Yan, and Q. Yu, *Mamm. Genome* **3**:644 (1992).
9. M. Q. Zhang and T. G. Marr, *Proc. Natl. Acad. Sci. USA* **90**:600 (1993).
10. M. J. Palazzolo, S. A. Sawyer, C. H. Martin, D. A. Smoller, and D. L. Hartl, *Proc. Natl. Acad. Sci. USA* **88**:8034 (1991).
11. T. Mizukami, W. I. Chang, I. Garkavtsev, N. Kaplan, D. Lombardi, T. Matsumoto, O. Niwa, A. Kounosu, M. Yanagida, T. G. Marr, and D. Beach, *Cell*, in press (1993).
12. M. Q. Zhang, *J. Stat. Phys.* **63**:1191 (1991).