GENOMICS 11, 806–827 (1991)

Genomic Mapping by Anchoring Random Clones: A Mathematical Analysis

RICHARD ARRATIA,* ERIC S. LANDER,†,‡ SIMON TAVARÉ,*,§ AND MICHAEL S. WATERMAN*,§

*Department of Mathematics and §Department of Biological Sciences, University of Southern California, Los Angeles, California 90089; †Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142; and ‡Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received March 8, 1991; June 7, 1991

A complete physical map of the DNA of an organism, consisting of overlapping clones spanning the genome, is an extremely useful tool for genomic analysis. Various methods for the construction of such physical maps are available. One approach is to assemble the physical map by "fingerprinting" a large number of random clones and inferring overlap between clones with sufficiently similar fingerprints. E. S. Lander and M. S. Waterman (1988, *Genomics* 2:231–239) have recently provided a mathematical analysis of such physical mapping schemes, useful for planning such a project. Another approach is to assemble the physical map by "anchoring" a large number of random clones—that is, by taking random short regions called anchors and identifying the clones containing each anchor. Here, we provide a mathematical analysis of such a physical mapping scheme.

1. INTRODUCTION

A complete physical map of the DNA of an organism, consisting of overlapping clones spanning the genome, is an extremely useful tool for genomic analysis. Complete or nearly complete physical maps have already been constructed for the genomes of *Escherichia coli* (Kohara *et al.*, 1987), *Saccharomyces cerevisiae* (Olson *et al.*, 1986), and *Caenorhabditis elegans* (Coulson *et al.*, 1986) and numerous efforts are underway to construct physical maps of considerably larger genomes, such as the human and mouse genomes.

Most physical mapping projects to date have used the approach of linking random clones by fingerprinting. Each clone is individually analyzed to obtain a "fingerprint" reflecting partial information about its sequence. Depending on the precise nature of the clones and of the genome studied, various fingerprinting schemes are possible (including lengths of all restriction fragment lengths, lengths of restriction fragments containing particular repeat sequences, or complete restriction maps). Clones with sufficiently similar fingerprints are likely to overlap. Overlapping clones are then assembled into "islands" or "contigs," which cover larger regions of the genome.

To plan a physical mapping project, it is important to know how the distribution of islands changes as a function of the fingerprinting method and of the number of clones studied. Lander and Waterman (1988) have presented such a mathematical analysis for physical mapping by fingerprinting random clones, which can be used to help design such projects.

Recently, interest has focused on an alternative method of physical mapping: linking random clones by anchoring. In this approach, one has a random genomic library of "clones" (which will contain large genomic inserts, as in phage, cosmids, or yeast artificial chromosomes (YACs)) and a random genomic library of "anchors" (which will contain very short genomic inserts, as in small plasmids). An anchoring method involves determining which clones contain a given anchor. Clones containing a common anchor are linked into islands.

The various laboratory methods that could be used for anchoring clones include the following:

(i) PCR screening. Anchors could consist of short PCR assays for unique regions of the genome, which have been dubbed sequence tagged sites (STSs). The clones containing an anchor could be found by PCR assays on appropriately nested sets of clones, as described by Green and Olson (1990). This approach is sometimes called STS content mapping.

(ii) Filter hybridization. Anchors could consist of short unique-sequence genomic DNA probes. The clones containing an anchor could be found by hybridizing the probe to a filter containing a gridded array of the clones.

(iii) Recombinational screening. One could even imagine an *in vivo* method. Anchors might consist of cells containing a plasmid with a short unique-sequence genomic DNA insert. The clones containing

0888-7543/91 $3.00

an anchor could be identified by mating an anchor to a gridded library of clones, under conditions where the progeny could grow only if the anchor and clone sequences underwent homologous recombination. Although this precise approach has not yet been experimentally implemented, related schemes have been developed (Seed, 1980).

Anchoring schemes are especially useful when the anchors consist of clones whose "top-down" position in the genome is known (such as DNA polymorphisms or markers located by *in situ* hybridization). Such anchors make it possible to order islands along the genome. For experimental purposes, physical maps of ordered islands are immediately useful even without complete long-range continuity.

The purpose of this paper is to provide a mathematical analysis of physical mapping by anchoring random clones, analogous to that presented by Lander and Waterman (1988) for fingerprinting random clones. Specifically, we describe the distribution of islands as a function of the number of anchors and number of clones used.

The paper is divided into two parts in an attempt to serve readers with different interests. The part labeled "Biological Results" (Sections 2.1–2.5) presents the key useful results for readers interested primarily in applications and omits mathematical technicalities and proofs. The section labeled "Mathmatical Results" (Sections 3.1–3.3) is aimed at the mathematical reader and contains the more technical results and all proofs.

The sections are organized as follows. Under "Biological Results," Section 2.1 presents the basic formulas for the distribution of islands for the case of clones of fixed length and discusses the strategic issues involved in designing a physical mapping experiment. Section 2.2 generalizes the formulas to the case of clones of variable length, examining the effect of such variability. Section 2.3 addresses the problem that arises when physical maps must be constructed from a clone library containing a substantial number of clones with chimeric inserts. Section 2.4 introduces the question of expectations from the standpoint of randomly chosen anchors and randomly chosen clones, in the case of clones of fixed length. Section 2.5 applies the results of the previous sections to the practical example of constructing a physical map of a mammalian chromosome or genome. Under "Mathematical Results," Section 3.1 discusses an elegant mathematical duality between clones and anchors, in the case of fixed length. Section 3.2 contains the mathematical proofs for Sections 2.1, 2.2, and 2.3. Section 3.3 provides the mathematical underpinnings for the discussion in Section 2.4 concerning anchor-biased and clone-biased sampling of islands.

After the acceptance of this paper, we became aware of the recent papers of Barillot *et al.* (1991) and Torney (1991), which contain some results related to ours below.

2. BIOLOGICAL RESULTS

2.1. *Properties of Islands: Constant Length Clones*

An anchoring scheme refers to any method for determining which clones contain a given anchor. From a purely mathematical standpoint, one can abstract away the experimental details and imagine placing paper strips (clones) randomly along a line (genome) and then placing staples (anchors) randomly along the line to join the strips into larger units (islands).

Throughout, we assume that we have a perfectly representative sample of genomic clones and a perfectly random collection of anchors. (More precisely, we mean that clones and anchors are distributed according to a homogeneous Poisson process along the genome.) In practice, these assumptions are likely to be violated: the collection of clones and anchors will be subject to (unknown) cloning bias; the collection of clones will often not be randomly distributed because the inserts are produced by partial digestion rather than random shearing (and thus will have systematic biases in regions rich or poor in the particular restriction site); and the collection of anchors will often be screened to avoid repetitive elements or, perhaps, to enrich for genes. Nevertheless, we employ these simple assumptions in this initial analysis because we feel that they capture the essential features of the problem and because we lack sufficient data to model the inhomogeneities.

We define the following symbols:

G, haploid genome length in basepairs (bp);

L, length of clone insert in basepairs (which may be a constant or variable);

N, number of clones in library;

M, number of anchors studied;

$a = LN/G$, expected number of clones covering a random point (also called the redundancy of coverage);

$b = LM/G$, expected number of anchors contained in a random clone.

(For the mathematically fastidious reader, we note that N and M are actually treated as random variables in the proofs, with expectation equal to the number of clones and anchors used. See Section 3.2.)

Clones are linked by anchors into *apparent islands*, consisting of one or more members. The islands are only *apparent* because some actual overlaps will go undetected; the actual islands that would result if all overlaps could be detected would be larger. A *singleton clone* refers to a clone anchored to no other clone.

Islands consisting of just a singleton clone are called *singleton islands*, while islands containing two or more clones are called *contigs*. With islands ordered on the genome, the apparent gaps between consecutive islands are called *oceans*. Oceans can be *actual oceans* if the islands do not overlap, or only *apparent oceans* if the islands overlap but the overlap has not been detected because it lacks an anchor.

It is useful to distinguish between *anchored islands* and *unanchored islands*—that is, islands that do or do not contain at least one anchor. In many respects, only anchored islands are interesting: anchored islands can often be ordered along the genome by using the anchors for top-down localization (as by genetic mapping or *in situ* hybridization) and thus become immediately useful for experimental purposes, while unanchored islands contain no information about bottom-up linkage to other clones or top-down position in the genome. The fraction of the genome covered by anchored islands is an important measure of the value of a physical map.

We begin by describing the properties of anchored islands.

PROPOSITION 1 (Anchored Islands). *With notation above, we have:*

(i) *The probability q_1 that a clone contains no anchors is e^{-b}. The expected number of unanchored islands is Ne^{-b}.*

(ii) *The probability p_1 that a clone is the rightmost clone of an anchored island is*

$$p_1 = \frac{b(e^{-a} - e^{-b})}{(b-a)} \quad \text{if } a \neq b$$

and

$$= ae^{-a} \quad \text{if } a = b.$$

The expected number of anchored islands is then Np_1.

(iii) *The expected number of clones in an anchored island is $(1 - q_1)/p_1$.*

(iv) *The probability p_2 that a clone is a singleton anchored island is*

$$p_2 = \frac{b(a^2 - ab - b)}{(a-b)^2} e^{-(a+b)} + \frac{b^2}{(a-b)^2} e^{-2a} \quad \text{if } a \neq b$$

and

$$= \frac{2a + a^2}{2e^{2a}} \quad \text{if } a = b.$$

The expected number of singleton anchored islands is Np_2.

(v) *The expected length of an anchored island is λL, where*

$$\lambda = \frac{(a-b)^2 e^{a+b} + a(ab - b^2 - a)e^a + b(2a - b)e^b}{ab(a-b)(e^a - e^b)} \quad \text{if } a \neq b$$

and

$$= \frac{2e^a + a^2 - 2a - 2}{2a^2} \quad \text{if } a = b.$$

(vi) *The expected proportion r_0 of the genome not covered by anchored islands is*

$$r_0 = e^{-a} + \frac{a(b^2 - ab - a)}{(b-a)^2} e^{-(a+b)} + \frac{a^2}{(b-a)^2} e^{-2b} \quad \text{if } a \neq b$$

and

$$= \frac{2e^a + 2a + a^2}{2e^{2a}} \quad \text{if } a = b.$$

(vii) *The expected number of anchors in an anchored island is $b(1 - e^{-a})/ap_1$.*

(viii) *For any $x \geq 0$, the probability that an anchored island is followed by actual ocean of length at least xL is $e^{-a(x+1)}(1 - q_1)/p_1$. In particular, taking $x = 0$, the formula gives the probability that an anchored island is followed by an actual ocean rather than an undetected overlap.*

(ix) *As $b \to \infty$, every island is anchored and all overlaps are detected. The results coincide with those given by Lander and Waterman (1988) for fingerprinting schemes with $\theta = 0$.*

Results about anchored islands can be easily converted into results about all islands by accounting for the effect of unanchored islands, which simply requires taking appropriate weighted averages.

COROLLARY 2 (All Islands). *With the notation above, we have:*

(i) *The expected number of islands is $N(p_1 + q_1)$.*

(ii) *The expected number of clones in an island is $1/(p_1 + q_1)$.*

(iii) *The expected number of singleton islands is $N(p_2 + q_1)$.*

(iv) *The expected length of an island is $L(p_1\lambda + q_1)/(p_1 + q_1)$.*

(v) *The expected proportion of the genome not covered by islands is e^{-a}.*

(vi) *The expected number of anchors in an island is $b(1 - e^{-a})/a(p_1 + q_1)$.*

(vii) *The probability that an actual ocean of length at least xL occurs at the end of an island is $e^{-a(x+1)}/(p_1 + q_1)$. In particular, taking $x = 0$, the formula gives the probability that an apparent ocean is real (as opposed to an undetected overlap occurring).*

Figure 1 shows some of the results in graph form. Because anchoring methods usually involve screening a fixed clone library in parallel, the graphs show curves for fixed clone libraries with coverage $a = 1, 2, \ldots, 10$ as the coverage b in anchors increases along the horizontal axis. In addition, the figure shows values of G/L for various representative situations.

It is particularly instructive to consider the proportion of genome covered by anchored islands. If the anchors can be ordered in a top-down fashion (as by genetic mapping), then anchored islands can be positioned in the genome. Such a physical map will be extremely useful as soon as the majority of the genome is covered with anchored islands, even if the islands are relatively short. For many purposes (such as cloning genes known by genetic location), direct experimental access to most regions will be more important than long-range continuity.

Figure 1A shows that the genome is rapidly covered by anchored islands even for relatively small b. Indeed, one rapidly reaches the practical point of diminishing returns around $b = 3$. For a clone library with three-fold coverage, the proportion of genome covered is 67, 89, and 93% for $b = 1, 2, 3$ and 95% for $b = \infty$. For a clone library with five-fold coverage, the proportion of genome covered is 79, 95, and 98% for $b = 1, 2, 3$, and 99% for $b = \infty$. More generally, Proposition 1(vi) makes clear that the proportion of genome covered for a clone library with a-fold coverage rapidly approaches $1 - e^{-a}$.

While the proportion of genome covered by anchored islands already begins to approach saturation for $(a, b) = (3, 3)$, significant gains in long-range continuity are achieved for larger a and b—as reflected in the decrease in expected number and increase in expected length of anchored islands shown in Figs. 1C and 1D. The expected number of anchored islands reaches a maximum in the range $1 \leq b \leq 2$ for the cases of interest and falls steadily thereafter. The expected length of anchored islands is only about $2L$ for $(a, b) = (3, 3)$, but grows to exceed $5L$ for $(a, b) = (5,5)$. Greater coverage in clones and anchors is thus needed to achieve long-range continuity than to achieve a high degree of genomic coverage by anchored islands.

It is useful for some purposes to know about the full distribution of the length S of an anchored island, not just about the average length $ES := \lambda L$ given in Proposition 1(v). Although we do not have a closed form expression for the length distribution of anchored islands, we can offer the following heuristic approximation. Anchored islands fall into two classes: (i) singleton anchored islands, comprising an expected proportion $p_3 := p_2/p_1$ of the total and having length L; and (ii) nonsingleton anchored islands or contigs, comprising a proportion $1 - p_3$ of the total and having average length $\lambda^* L$, where $\lambda^* = (\lambda - p_3)/(1 - p_3)$. Reasoning that the occurrence of an ocean is a rare event (for a and b large), we would expect that the length S of nonsingleton anchored islands should be approximately distributed as $S = L(1 + X)$, where X is an exponentially distributed random variable with mean $\lambda^* - 1$. This heuristic turns out to provide an excellent approximation for a and b in the range of interest for physical mapping, as shown by computer simulations (see Fig. 2).

Finally, the properties of *all* islands (both anchored and unanchored) can be computed from Corollary 2, but these averages are somewhat less informative because they are often dominated by the contribution from the unanchored islands. We elaborate on this point in Section 3.2.

2.2. *Properties of Islands: Variable Length Clones*

In the previous section, we assumed that clones have constant length. This assumption may not be too bad for phage and cosmid libraries, which have a fairly strict size range due to packaging constraints, but is considerably worse for some YAC libraries, which can have substantial length variation. We now extend the results to the case of clones of variable length. (Although the generalization obviously includes the results of the previous section as a special case, we stated the special case first because these simpler results are more user friendly for most experimentalists.)

We suppose that clone lengths L are independent identically distributed random variables with mean length EL. The probability density function of the normalized length L/EL is denoted by $f(l)$. We define the auxiliary functions

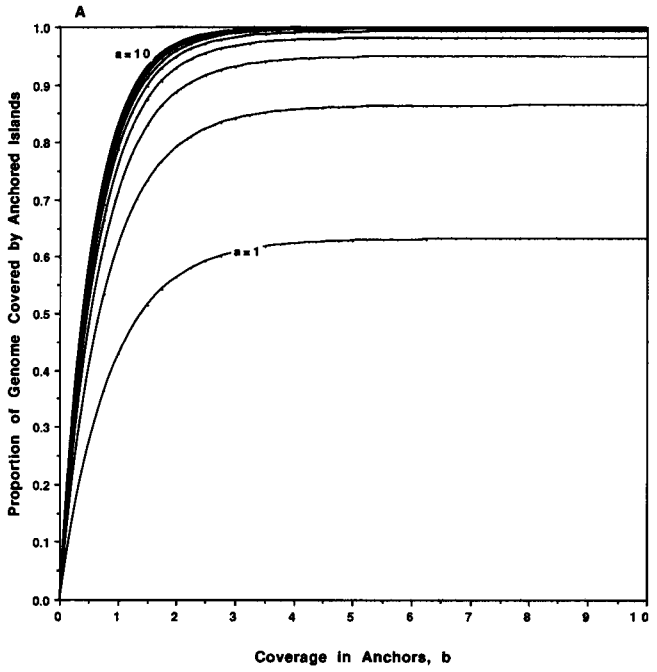
$$\mathcal{F}(x) = \text{Prob}(L/EL > x) = \int_x^\infty f(l)dl$$

and

$$J(x) = \exp\left\{-a \int_x^\infty \mathcal{F}(l)dl\right\},$$

which depend on a and f. As described in Section 3.2, the function $J(x)$ has the nice interpretation that it is the probability that two points separated by distance x are not covered by a common clone.

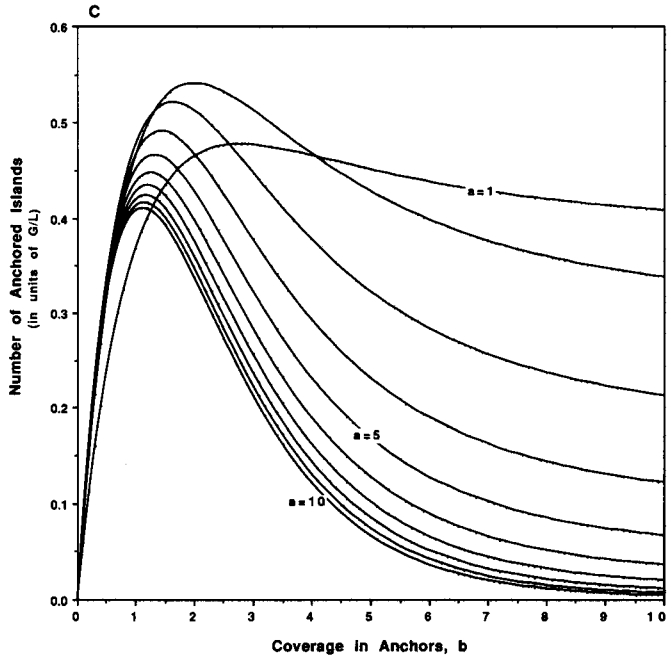
It is also convenient to redefine the notion of singleton clones and islands. A *singleton clone* refers to a clone that is anchored to no other clone extending beyond its boundary (but may be anchored to clones completely contained within it). A *singleton island* refers to an island completely covered by a singleton clone (but may contain other clones completely contained within this clone). In the case of fixed clone



B

Approximate Values of G/L

	Phage (15 kb)	Cosmid (40 kb)	YAC (300 kb)	YAC (1 Mb)
S. cerevisiae (13 Mb)	867	325	43	13
A. thaliana (80 Mb)	5333	2000	267	20
Avg. Human Chrom. (150 Mb)	10,000	3,750	500	150
Human Genome (3 Gb)	200,000	75,000	10,000	3,000



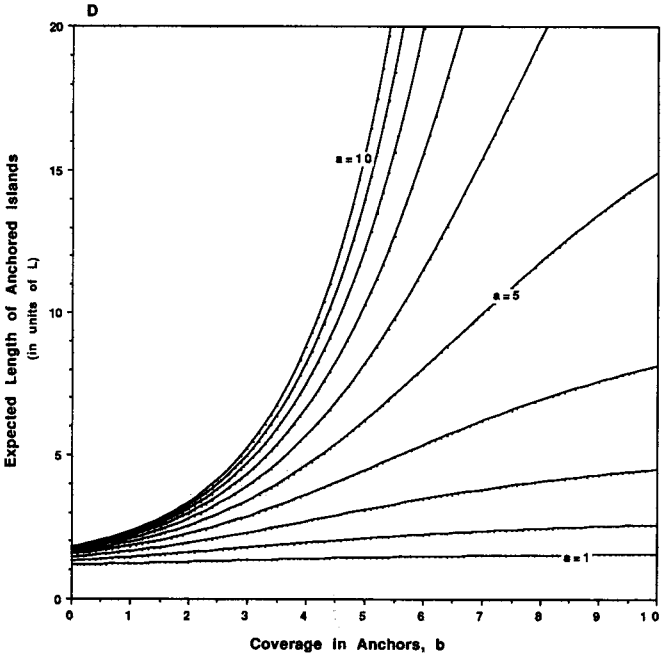


FIG. 1. For the situation of clones of constant length, the graphs show (**A**) the expected proportion of the genome covered by anchored islands, (**C**) the expected number of anchored islands, and (**D**) the expected length of an anchored island, as a function of coverage b in anchors for the values $a = 1, 2, \ldots, 10$. (Only a few of the curves are marked by the corresponding value of a, with the others occurring in the expected order.) To make the graphs independent of genome and clone size, the expected number of islands is expressed in units of G/L (the number of clones needed to cover the genome) and the expected length of islands is expressed in terms of L (the length of a clone). The table (**B**) lists the value of G/L for certain representative genomes and cloning vectors, including two different sizes of YACs.

length, these definitions coincide with those given earlier. Proposition 1 can now be generalized as follows.

PROPOSITION 3. *With notation as above, we have:*

(i) *The probability q_l that a clone contains no anchors is*

$$q_l = \int_0^\infty e^{-bl} f(l)\,dl.$$

The expected number of unanchored clones is Nq_l.

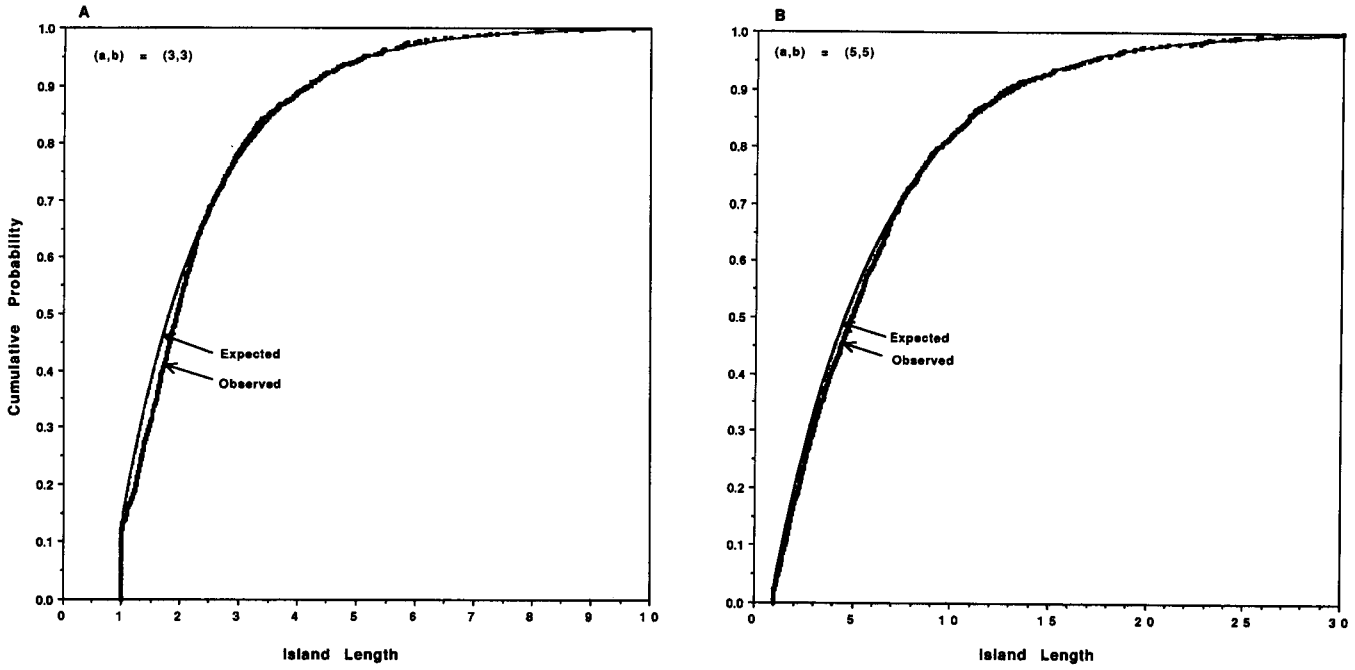


FIG. 2. Distribution of lengths of anchored islands (in units of L) for (**A**) the case $(a, b) = (3, 3)$ and (**B**) the case $(a, b) = (5, 5)$. Graphs show the "observed" results for the cumulative distribution function for length for 1000 anchored islands in a computer simulation and the "expected" results based on the prediction that anchored islands consist of a proportion p_3 of singletons with length L and a proportion $1 - p_3$ of nonsingletons with lengths distributed as $L(1 + X)$, where X is an exponentially distributed random variable with mean $\lambda^* - 1$. (See text for definition of p_3 and λ^*.) Computer simulations were performed as described at the end of Section 2.4.

(ii) *The probability p_1 that a clone is the rightmost clone of an anchored island is*

$$p_1 = \int_0^\infty be^{-bu}J(u)\mathcal{F}(u)du.$$

The expected number of anchored islands is then Np_1.

(iii) *The expected number of clones in an anchored island is $(1 - q_1)/p_1$.*

(iv) *The probability p_2 that a clone is a singleton anchored island is*

$$p_2 = \int_0^\infty \int_0^l \int_0^{l-v} b^2e^{-b(u+v)} \frac{J(u)J(v)}{J(l)} f(l)dudvdl + \int_0^\infty \int_0^l be^{-bl} \frac{J(u)J(l-u)}{J(l)} f(l)dudl.$$

The expected number of singleton anchored islands is Np_2.

(v) *The expected length of an anchored island is λEL, where*

$$\lambda = \left\{1 + \int_0^\infty (b^2u - 2b)e^{-bu}J(u)du\right\} \Big/ ap_1.$$

(vi) *The expected proportion r_0 of the genome not covered by anchored islands is*

$$r_0 = \int_0^\infty \int_0^\infty b^2e^{-b(u+v)} \frac{J(u)J(v)}{J(u+v)} dudv.$$

(vii) *The expected number of anchors in an anchored island is $b(1 - e^{-a})/ap_1$.*

(viii) *The probability that an anchored island is followed by an actual ocean of length at least $x(EL)$ is $e^{-a(x+1)}(1 - q_1)/p_1$. In particular, taking $x = 0$, the formula gives the probability that an anchored island is followed by an actual ocean rather than an undetected overlap.*

(ix) *As $b \to \infty$, every island is anchored and all overlaps are detected. The formulas reduce to:* (i) *0*; (ii) e^{-a}; (iii) e^a; (iv) $e^{-2a}Y$; (v) $EL(e^a - 1)/a$; (vi) e^{-a}; (vii) ∞; (viii) e^{-ax}, *where*

$$Y = \int_0^\infty \frac{f(l)}{J(l)} dl.$$

Results about anchored islands can be converted into results about all islands by accounting for the effects of unanchored islands. The result is identical to Corollary 2, except for a slight modification in part (iv).

COROLLARY 4 (All Islands). *With notation above, we have:*

(i) *The expected number of islands is $N(p_1 + q_1)$.*

(ii) *The expected number of clones in an island is* $1/(p_1 + q_1)$.

(iii) *The expected number of singleton islands is* $N(p_2 + q_1)$.

(iv) *The expected length of an island is* $EL(p_1\lambda + q_1\lambda')/(p_1 + q_1)$, *where* λ' *is the normalized expected length of an unanchored island,*

$$\lambda' = \left\{\int_0^\infty le^{-bl} f(l)\,dl\right\} \Big/ q_1.$$

(v) *The expected proportion of the genome not covered by islands is* $J(0) = e^{-a}$.

(vi) *The expected number of anchors in an island is* $b(1 - e^{-a})/a(p_1 + q_1)$.

(vii) *The probability that an actual ocean of length at least* $x(EL)$ *occurs at the end of an island is* $e^{-a(x+1)}/(p_1 + q_1)$. *In particular, taking* $x = 0$, *the formula gives the probability that an apparent ocean is real (as opposed to an undetected overlap occurring).*

With the results above, it is possible to calculate expected values for any given distribution of clone lengths by simply using the appropriate functions $\mathcal{F}$ and J. The case of fixed-length clones corresponds to

$$J(x) = e^{-a(1-x)}, \quad \text{if } x \leq 1.$$
$$= 1, \quad \text{if } x > 1.$$

We mention two other important special cases.

PROPOSITION 5. *Suppose that normalized clone lengths* L/EL *are distributed uniformly in the interval* $[(1 - s), (1 + s)]$. *Then*

$$J(x) = \begin{cases} \exp(-a(1-x)) & 0 \leq x \leq (1-s) \\ \exp(-a(1+s-x)^2/4s) & (1-s) < x \leq (1+s) \\ 1 & (1+s) < x. \end{cases}$$

PROPOSITION 6. *Suppose that normalized clone lengths* L/EL *are exponentially distributed with mean 1. Then* $J(x) = \exp(-a\exp(-x))$.

In the first case, the integrals in Proposition 4 can all be expressed in closed form in terms of the cumulative distribution function of the normal distribution, although the precise formulas are unenlightening and we omit them. In the second case, the integrals cannot be expressed in closed form. In any case, numerical results can always be easily obtained by straightforward numerical integration.

For example, Fig. 3 shows the situation of normalized clone lengths L/EL uniformly distributed in the interval $[(1 - s), (1 + s)]$ as in Proposition 5, for the case $a = 3$ and $s = 0.0, 0.5, 1.0$. As clone length becomes more variable, we see that (i) the expected number of anchored islands decreases, (ii) the expected length of anchored islands increases, and (iii) the expected proportion of genome covered by anchored islands increases compared to the fixed length case.

In fact, the first observation turns out to be true for any distribution of clone lengths.

PROPOSITION 7. *The expected number of anchored islands is always larger for the case of clones having constant length than for the case of clones having variable length with the same mean.*

By contrast, the second and third observations concerning expected length of anchored islands and expected proportion of genome covered are not true for all length distributions. (For example, one can construct counterexamples to both for the rather uninteresting case of very small a.) However, the observations do appear to hold for all practical situations that we have examined—for example, for uniform length distribution and $2 \leq a, b \leq 8$. Thus, we suggest them as reasonable heuristics in applications.

2.3. *Coping with Chimeras*

Clone libraries sometimes have the problem that clones may contain chimeric inserts—that is, inserts that consist of DNA from two or more regions of the genome (which have been juxtaposed by virtue of an artifact occurring *in vitro* or *in vivo*). This has been a particular problem for some first-generation YAC libraries of mammalian genomes, in which the frequency of chimeric clones may approach 50% (E. Green, personal communication). What problems do chimeras pose for constructing physical maps?

If all anchors were ordered in a top-down fashion in the genome (as by genetic mapping or *in situ* hybridization), chimeric clones would cause no problem because false linkages would be detected as apparently joining anchors known to lie in different genomic regions.

For many physical mapping projects of mammalian genomes, however, it may not be practical to assign top-down positions to the vast majority of anchors. In this case, the accuracy of the map rests precariously on the linkages inferred from the anchors. Two islands from different regions of the genome might be unknowingly joined, based on a single chimera that appeared to link them.

One sensible solution would be to construct islands by requiring that consecutive anchors be linked by at least *two* clones. A region of the genome covered in this fashion could be called *double-linked islands*. Because chimeric junctions are thought to be (relatively) random occurrences, it is unlikely that one would encounter two independent clones represent-

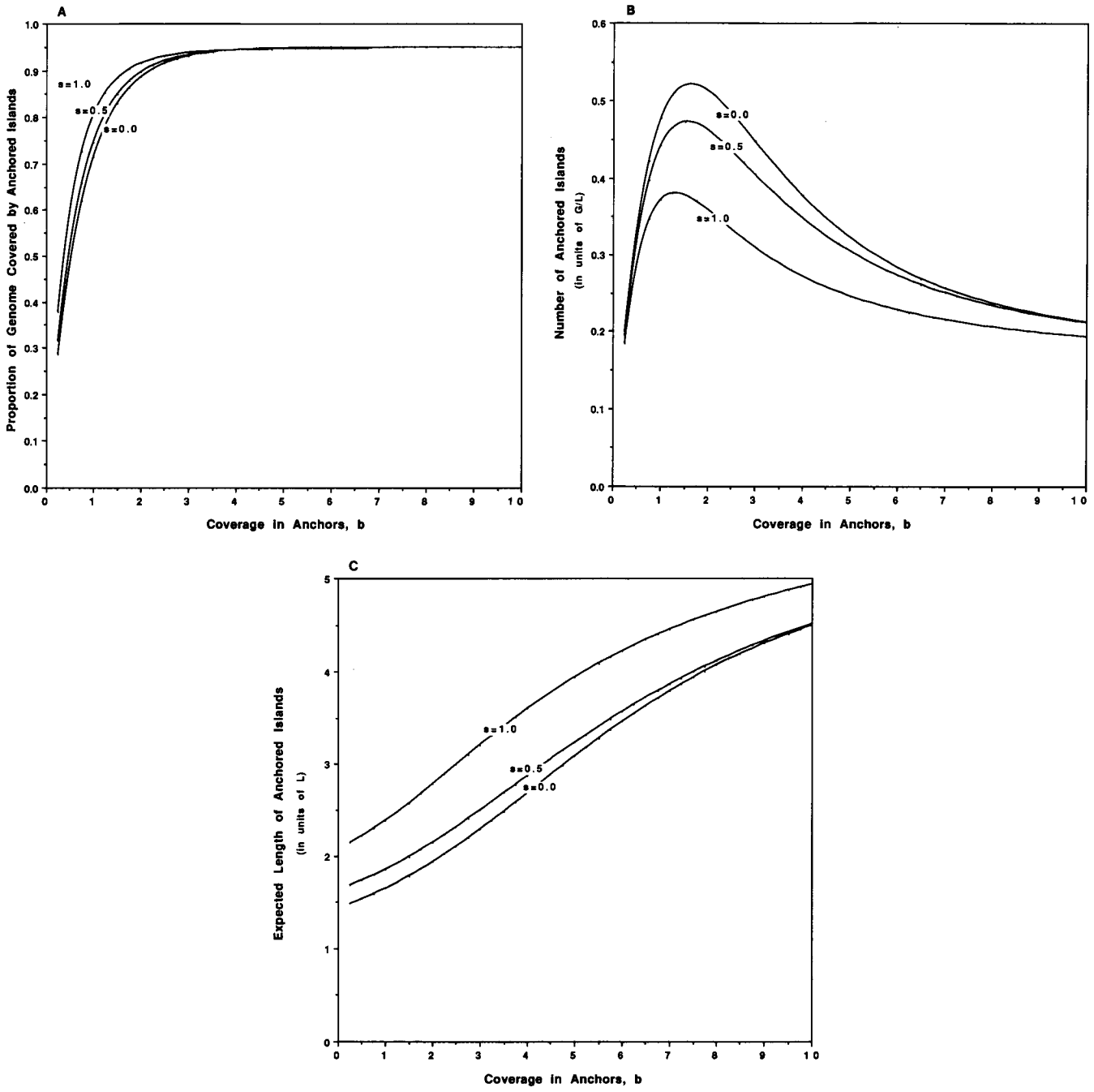


FIG. 3. For the situation of clones with normalized length L/EL uniformly distributed in the interval $[(1-s),(1+s)]$, the graphs show (**A**) the expected proportion of the genome covered by anchored islands, (**B**) the expected number of anchored islands, and (**C**) the expected length of an anchored island, as a function of the coverage b in anchors for the cases $a = 3$ and $s = 0.0$, 0.5, and 1.0.

ing the same chimeric junction. Thus, double-linked islands should almost always be correct.

Alternatively, one could construct preliminary islands by using single linkages, but then perform detailed restriction mapping of each island to determine overlap among clones. One could then break each preliminary island wherever it was covered by only a single clone (since this might be a chimeric junction). Thus, one would only believe contiguous regions covered throughout their length by at least two clones, regions that could be called *double-covered islands*. Double-linked islands are always double-covered, but the converse is not true: the region between two anchors may be covered at every point by multiple

clones without there being multiple clones connecting the anchors. Double-covered islands take considerably more work to assemble than double-linked islands, but they will be somewhat longer in general. These definitions are illustrated in Fig. 4.

In either case, it is important to know: How would the progress toward the completion of such physical map be affected by the need to build either double-linked islands or double-covered islands?

It is helpful to consider slightly more general notions because they make the mathematical results more transparent. An *anchored k-linked island* will be defined as a maximal interval I of the genome with the properties that (i) I contains at least one anchor; (ii) every pair of consecutive anchors in I is connected by at least k clones; and (iii) every point in I is connected to some anchor in I by at least k clones. An *anchored k-covered* island will be defined as a maximal interval I of the genome with the properties that (i) I contains at least one anchor; (ii) every pair of consecutive anchors in I is connected by at least one clone; and (iii) every point in I is covered by at least k clones, each connected to an anchor in I. (Note that both anchored 1-linked islands and anchored 1-covered islands reduce to the usual notion of anchored islands.)

The following results now generalize Proposition 3 to the case of anchored k-linked and k-covered islands.

PROPOSITION 8. *For anchored k-linked islands, the following results hold:*

(i) *The expected number of anchored k-linked islands is* Np_1.

(ii) *The expected length of an anchored k-linked island is* $EL(1 - r_0 + r_2)/ap_1$.

(iii) *The expected proportion of the genome not covered by k-linked anchored islands is* r_0.

The quantities p_1, r_0 *and* r_2 *are defined as follows. Let* V *and* W *be independent exponential random variables with parameter* b. *Let* R, S, *and* T *be random variables that, conditional on* V *and* W, *are independent Poisson variables with means*

$$E(R|V, W) = a\int_{V+W}^{\infty} \mathcal{J}(t)dt$$

$$E(S|V, W) = a\int_{W}^{V+W} \mathcal{J}(t)dt$$

$$E(T|V, W) = a\int_{V}^{V+W} \mathcal{J}(t)dt$$

and

$$p_1 = \int_0^\infty\int_0^\infty\int_0^\infty 1\{v < l\}[P(T + R = k - 1|V = v, W = w)] \times b^2e^{-b(v+w)}f(l)dwdvdl$$

$$= \int_0^\infty\int_0^\infty [P(T + R = k - 1|V = v, W = w)] \times b^2e^{-b(v+w)}\mathcal{J}(v)dwdv$$

$$r_0 = \int_0^\infty\int_0^\infty [P(R + S < k, R + T < k|V = v, W = w)] \times b^2e^{-b(v+w)}dwdv$$

$$r_2 = \int_0^\infty\int_0^\infty [P(R < k, S \geq k, T \geq k|V = v, W = w)] \times b^2e^{-b(v+w)}dwdv,$$

where $1\{X\}$ *is the indicator function for the event* X *and* $P(X)$ denotes the probability of the event X.

PROPOSITION 9. *For anchored k-covered islands, the following results hold:*

(i) *The expected number of anchored k-covered islands is* Np_1.

(ii) *The expected length of an anchored k-covered island is* $EL(1 - r_0 + r_2)/ap_1$.

(iii) *The expected proportion of the genome not covered by k-covered anchored islands is* r_0,

where V, W, R, S, *and* T *are defined as in the previous proposition and*

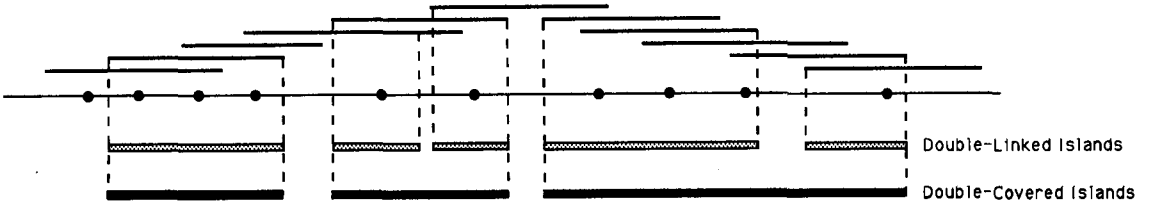


FIG. 4. Illustration of concept of double-linked and double-covered islands, useful for dealing with the problem of chimeric clones. The diagram shows the genome (horizontal line), anchors (circles on the genome), and clones (line segments drawn above the genome). While the clones lie in a single ordinary anchored island, the island is broken into five double-linked islands and three double-covered islands (indicated below the genome).

$$p_1 = \int_0^\infty \int_0^\infty \int_0^\infty 1\{v < l\} \times [P(R = 0, T = k - 1 \mid V = v, W = w) + P(R > 0, R + S + T = k - 1 \mid V = v, W = w)] \times b^2 e^{-b(v+w)} f(l)\, dw dv dl$$

$$= \int_0^\infty \int_0^\infty [P(R = 0, T = k - 1 \mid V = v, W = w) + P(R > 0, R + S + T = k - 1 \mid V = v, W = w)] \times b^2 e^{-b(v+w)} \mathcal{F}(v)\, dw dv$$

$$r_0 = \int_0^\infty \int_0^\infty [P(R = 0, S < k, T < k \mid V = v, W = w) + P(R > 0, R + S + T < k \mid V = v, W = w)] \times b^2 e^{-b(v+w)}\, dw dv$$

$$r_2 = \int_0^\infty \int_0^\infty [P(R = 0, S \geq k, T \geq k \mid V = v, W = w)] \times b^2 e^{-b(v+w)}\, dw dv.$$

These formulas can be straightforwardly computed for any length distribution since the bracketed terms simply involve probabilities for independent Poisson random variables. For physical mapping, we are primarily interested in the case $k = 2$, corresponding to anchors double-linked and double-covered islands.

Figure 5 shows the properties of anchored double-linked islands and anchored double-covered islands, for the case of fixed-length clones. Comparing these graphs with the results for ordinary anchored islands shown in Fig. 1, we can assess the consequences of coping with chimeras.

We observe that the proportion of genome covered is considerably less than that for ordinary anchored islands. Indeed, it is easy to see (by considering the Poisson probability of a point being covered by no clones or by exactly one clone) that the proportion of genome *not* covered by single-covered anchored islands tends to e^{-a} as $b \to \infty$, while the proportion *not* covered by anchored double-covered islands (or anchored double-covered islands) tends to $(1 + a)e^{-a}$ as $b \to \infty$. For both double-linked and double-covered islands, the proportion covered rises steeply in the interval $0 \leq b \leq 2$ and is already very close to the asymptotic limit by $b = 4$.

If we assume that it is practical to achieve 3-fold coverage in anchors, the following conclusions emerge. To cover about 95% of the genome, one needs a 3.5-fold clone library using ordinary anchored islands, but a 6-fold library using either anchored double-linked islands or anchored double-covered islands. To cover 99% of the genome, one needs a 6-fold library using ordinary anchored islands, but about a 10-fold library using either anchored double-linked islands or anchored double-covered islands. For a and b in the range of interest, one thus requires about three additional genome equivalents to achieve the same degree of coverage.

Although double-linked and double-covered islands do not differ greatly with respect to the total proportion of genome covered, there is a striking difference in the expected number of islands and in their expected length. As suggested by the illustration in Fig. 4 and confirmed by the graphs in Fig. 5, double-linked islands are considerably shorter and more numerous than double-covered islands. It is easy to understand why: many double-linked islands are linked to one another through single clones; such islands can be joined into double-covered islands provided that one is willing to carefully inspect the putative linking region (e.g., by restriction mapping) to confirm that every point is supported by information from at least two independent clones.

Whether one uses double-linked islands or double-covered islands, substantial extra work will be required. Moreover, the resulting physical map is not as useful because all that one is certain about is the order of the anchors. Clones that are apparently nonchimeric may nevertheless actually be undetected chimeras, and thus the island ends cannot be used for chromosome walking without confirmation that they map to the correct location in the genome. Clearly, it would be far better to construct chimera-free clone libraries. Several groups are currently working toward this goal.

2.4. *Anchor-Biased and Clone-Biased Sampling of Islands*

We should emphasize that there is a distinction between a randomly chosen island and the island containing a randomly chosen clone or a randomly chosen anchor. For example, an experimentalist will often have a particular gene in hand (which can be converted into an anchor) and may wish to know the expected size of the island that contains it. On average, the island that contains a randomly chosen anchor will be larger than a randomly chosen island—because randomly chosen anchors will be more likely to fall in large islands than in small islands. That is, the sampling of islands is biased (e.g., by number of anchors or clones). This phenomenon is related to the well-known "waiting time" paradox in probability: Although the bus company knows that the mean arrival time between buses is 10 min, a passenger arriving at the bus stop at a random time will experience a wait of more than 5 min, on average, because random arrivals are more likely to fall in longer interbus intervals than in shorter ones.

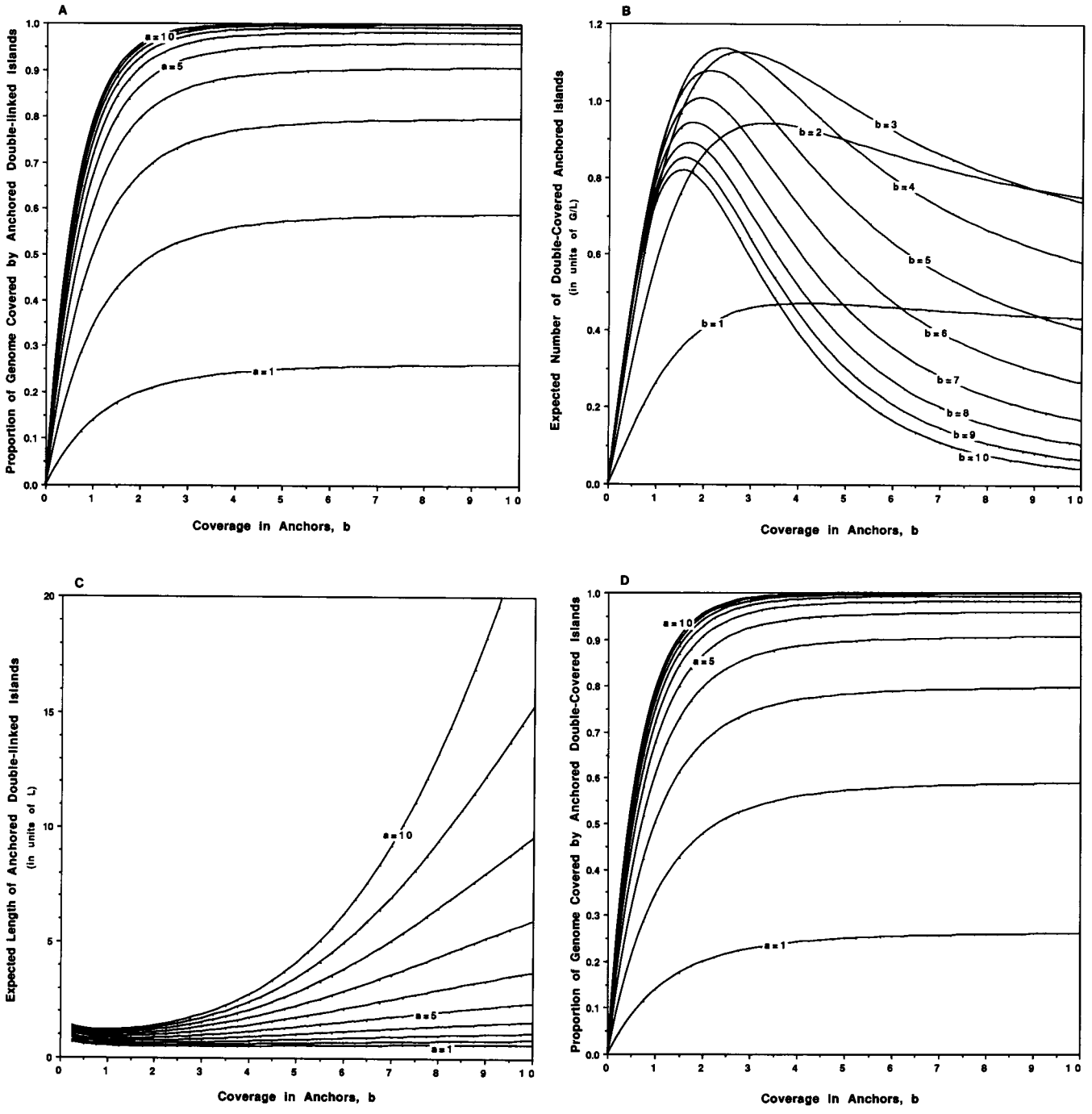


FIG. 5. For the situation of clones of constant length, the graphs show (A) the expected proportion of the genome covered by double-linked anchored islands, (B) the expected number of double-linked anchored islands, (C) the expected length of a double-linked anchored island, (D) the expected proportion of the genome covered by double-covered anchored islands, (E) the expected number of double-covered anchored islands, and (F) the expected length of a double-covered anchored island as a function of the coverage b in anchors for the values $a = 1, 2, \ldots, 10$. (Note different ranges in (B) and (E).)

The expected values under clone-biased and anchor-biased sampling can be computed by defining particular renewal processes discussed below in Section 3.3. Broadly, the key results can be summarized as follows: The length of the island containing a randomly chosen point in the genome is approximately 50% larger than the length of a randomly chosen island, for typical a and b in the range of interest (e.g., $3 \leq a, b \leq 5$).

Finally, we should note that additional properties of anchoring schemes beyond those discussed in this paper (e.g., complete distributions rather than simple

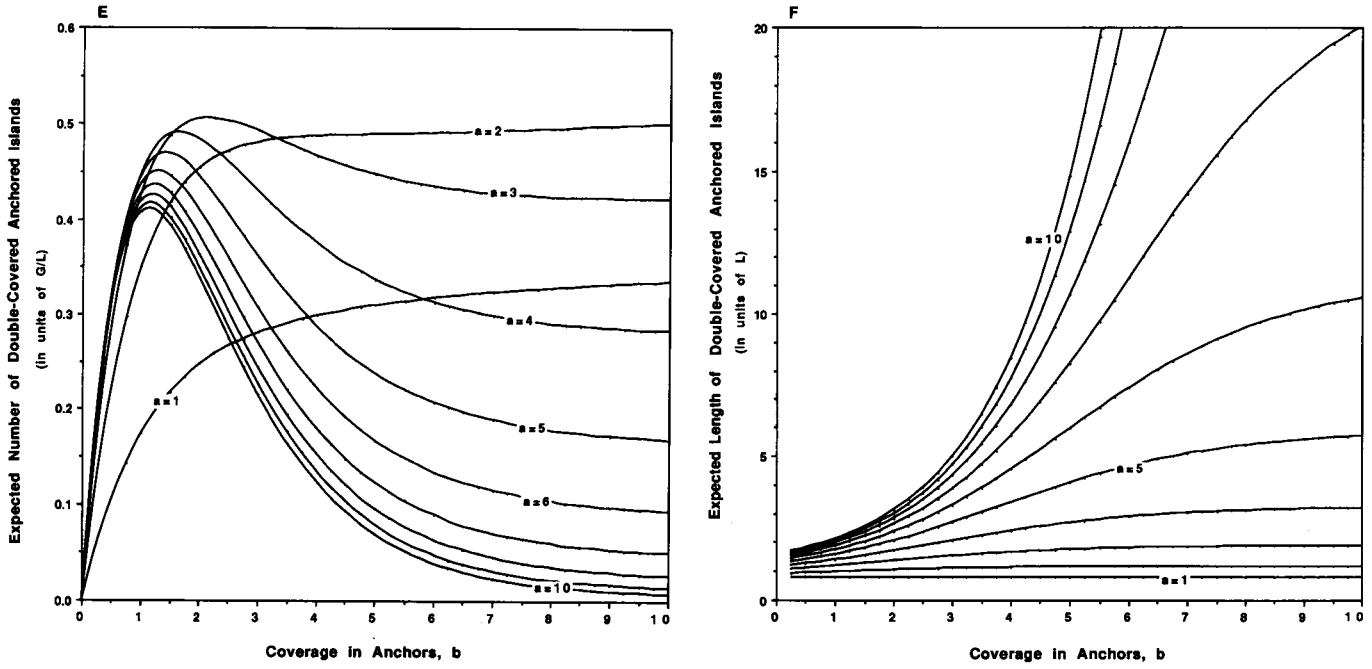


FIG. 5—*Continued*

means) can be obtained by computer simulation. We sketch a particularly efficient algorithm: (i) Generate an ordered sequence S of left-hand clone starts and anchors by choosing exponentially distributed arrival times with mean $1/(a + b)$ and assigning them to be clone starts or anchors by flipping a coin with head probability $a/(a + b)$; (ii) generate random lengths for each clone; (iii) intercalate the right clone ends into the ordered sequence S; and (iv) for each successive $p \in S$, keep track of the clones that have started but not yet ended at p and of the anchored island to the left of p. The expected running time is $O(abg)$, where $g = G/EL$.

2.5. *Practical Implications*

Consider the problem of mapping a typical mammalian chromosome of length about 150 Mb with YACs having a constant length of 500 kb so that $G/L = 300$. With 900 clones and 900 anchors, one expects to cover about 93% of the genome with about 135 islands having average length of about 1.1 Mb. With 1500 clones and 1500 anchors, one expects to cover about 99% of the genome with about 50 islands having average length of about 3 Mb. If a randomly chosen gene is used as an anchor, the expected length of the island containing it will be about 50% bigger in both cases.

Suppose, instead, that YACs have length distributed uniformly in the range 250–750 kb. With 900 clones and 900 anchors, one expects to cover about 94% of the genome with about 125 islands having average length of about 1.2 Mb. With 1500 clones and 1500 anchors, one expects to cover about 99% of the genome with about 40 islands having average length of about 4 Mb.

If chimeras are a problem, the situation is rather worse because one must achieve double linkage or double coverage (with the latter involving considerably more work). We consider only the case of clones of fixed length. With 900 clones (excluding those discovered to be chimeras during the mapping) and 900 anchors, one expects to cover about 68% of the genome with about 340 anchored double-linked islands having average length of about 340 kb. With 1500 clones and 1500 anchors, one expects to cover about 95% of the genome with about 220 anchored double-linked islands having average length of about 660 kb. With anchored double-covered islands, the proportion of genome covered is about the same, but the islands are fewer and longer.

Finally, constructing a physical map of an entire mammalian genome requires multiplying the numbers of clones and anchors by a factor of 20. Some 30,000 anchors might be required when using large YACs, which somewhat exceeds the limits of current practice (in which a high throughput lab might work with a few hundred anchors per year) but may not be beyond possibility. Some 500,000 would be needed if cosmids were used, which is so large as to be safely declared impractical. In summary, some improvement in efficiency and economy is needed to make this approach easily applicable to entire mammalian genomes.

3. MATHEMATICAL RESULTS

3.1. *Mathematical Duality between Clones and Anchors*

In the case of clones of constant length L, there is a fundamental symmetry between clones and anchors. Aside from being elegant, the symmetry implies that every formula $f(a, b)$ for some property of contigs can be interpreted as the formula for a dual property with the roles of clones and anchors reversed. To see the symmetry between clones and anchors, we must represent both elements by points: each clone should be labeled by the point at its center and each anchor by the point it occupies. A clone and an anchor are *adjacent* (that is, overlap) if the corresponding points lie within distance $L/2$. This adjacency relation defines a bipartite graph, whose connected components are of three types:

(i) isolated anchors,

(ii) isolated clones, and

(iii) islands containing at least one clone and at least one anchor.

The second class corresponds to the unanchored islands, while the third class corresponds to the anchored islands.

From the standpoint of symmetry, it is best to focus on anchored islands because this class is closed under duality—i.e., interchanging the role of clones and anchors. (As noted earlier, it is straightforward to convert properties of all islands into properties of anchored islands and *vice versa*.)

Some asymmetry between clones and anchors arises because the length of islands is naturally measured from the left end of the leftmost clone to the right end of the rightmost clone. To achieve symmetry, one should measure distances from the center of the leftmost clones to the center of the rightmost clone—which is smaller by L than the natural distance.

The following examples illustrates the use of duality. As before, let a be the intensity of clones and b be the intensity of anchors.

(i) Clearly, the expected number of clones in an anchored island is dual to the expected number of anchors in an anchored island. To confirm this, note that the formula for the first quantity is $(1 - e^{-b})(b - a)/b(e^{-a} - e^{-b})$ by Proposition 1(iii) and the formula for the second is $(1 - e^{-a})(b - a)/a(e^{-a} - e^{-b})$ by Proposition 1(vii). These are indeed related by interchange of a and b.

(ii) The expected proportion of the genome not covered by anchored islands (Proposition 1(v)) turns out to be the mathematical dual of the probability that a clone lies in a singleton island (Proposition 2(iii)). (Verify this by writing out both formulas explicitly in terms of a and b.) To see why, observe that a random point x_1 lies in an anchored island if and only if there exists a clone C and an anchor x_2 such that x_1 overlaps C and C overlaps x_2. Now, the dual situation involves adding a random clone C_1 and asking whether there exist an anchor x and a clone C_2 such that C_1 overlaps x and x overlaps C_2. This is exactly the condition describing whether C_1 lies in an island with at least two clones.

(iii) If $h(a, b)$ represents the expected length of an anchored island, then $h(a, b) - L \geqslant 0$ is the expected distance between the leftmost and rightmost clone centers in an anchored island. By duality, $h(b, a) - L$ is the expected distance between the leftmost and rightmost anchors in an anchored island (still using a as the intensity of clones). Thus, $h(a, b) - (h(b, a) - L)$ is the expected length of that part of an anchored island outside its outermost anchors. By the symmetry of left and right, one-half of this distance is the expected distance from the left end of the anchored island to its leftmost anchor.

Regrettably, the duality no longer holds in the variable length case.

3.2. *Mathematical Proofs of Main Results*

This section contains the proofs of the key results. We begin with Proposition 3 about variable length clones, since the results about fixed length clones follow as an easy special case.

We begin by specifying the mathematical model to which our propositions and corollaries apply as rigorous results. This model differs from the precise situation described in section 2.1 in three minor ways: the genome is modeled as continuous rather than discrete; boundary effects at the end of chromosomes are ignored; and N and M, the number of clones and anchors, are modeled as random variables rather than observed values. (As a consequence of this last point, expectations should be inserted into various formulas to make them mathematically strictly correct. For example, we should properly write $a = (EL)(EN)/G$ rather than $a = LN/G$, etc. In writing for a mixed audience, however, we have chosen to write the formulas using N rather than EN in the belief that this will be clearer to the biologist and that the reader versed in probability can insert E when needed.) It is easy to show that, for practical purposes, these differences have negligible effect for $G \gg EL$ and $EL \gg 1$.

We suppose that clones have independent and identically distributed lengths L, with mean EL. Let $Q = L/EL$ denote rescaled clone length, so that clones have average length 1. We suppose that clones occur on the real line $R \equiv (-\infty, \infty)$ according to a Poisson process of rate a, and anchors (modeled by points) occur according to a Poisson process of rate b. The genome corresponds to the interval $(0, G/EL)$, and we write g for the normalized genome length, $g = G/$

EL. (A convenient reference for the techniques pertaining to stationary processes is Breiman, 1968.)

We assume that the process of left clone ends is a homogeneous Poisson process of rate a, and that (rescaled) clone lengths are independent and identically distributed with mean 1. It follows that the process of right clone ends is also a Poisson process of rate a, and we label these points $\{A_i, i \in Z\}$ (where Z denotes the set of integers) in such a way that

$$\cdots A_{-2} < A_{-1} < A_0 \leqslant 0 < A_1 < \cdots < A_N < g \leqslant A_{N+1} < \cdots.$$

Here N is the number of clones whose right ends fall in the genome $(0, g)$. We recall some basic facts about a Poisson process. First, N has a Poisson distribution with mean $EN = ag$ and variance ag,

$$P(N = n) = \frac{(ag)^n e^{-ag}}{n!}, \quad n = 0, 1, \ldots$$

and

$$N/g \to a, \text{ almost surely as } g \to \infty.$$

The interarrival times $A_i - A_{i-1}$ are mutually independent exponentially distributed random variables with mean $1/a$ and probability density ae^{-ax}, for $x > 0$.

It is a remarkable fact about a Poisson process that the distribution of the set $\{A_i, i \in Z, i \neq 0\}$ conditioned on the event $A_0 = 0$ is the same as the distribution of the set $\{A_i, i \in Z\}$ before conditioning. We use this in what follows, referring informally to a "given clone" when we argue by conditioning on having a clone at a given location.

The process of anchors is described by a stationary time-homogeneous Poisson process with rate b, and we assume that the process of anchors and the process of clones are independent.

Let N_u and N_a denote, respectively, the number of unanchored and anchored clones whose right-hand ends falls in $(0, g)$, so that $N = N_u + N_a$. Define q_1 to be the probability that a clone is unanchored. Since the probability that a clone of length Q has no anchors is e^{-bQ}, we have

$$q_1 = Ee^{-bQ} = \int_0^\infty e^{-bl} f(l)\, dl,$$

where $f(l)$, $l \geqslant 0$ is the density function of Q, the normalized clone length. It follows that the intensity of unanchored clones is aq_1, while the intensity of anchored clones is $a(1 - q_1)$. In addition,

$$\frac{N_u}{g} \to aq_1 \quad \text{and} \quad \frac{N_a}{g} \to a(1 - q_1)$$

almost surely as $g \to \infty$.

Next we form the process of right-hand ends of anchored islands. Anchored islands have the property that one anchored island cannot be completely contained within another anchored island. Therefore we can label anchored islands in order, taking the order either from the right or the left ends. We label their right-hand ends in such a way that

$$\cdots C_{-2} < C_{-1} < C_0 \leqslant 0 < C_1 < \cdots < C_K < g \leqslant C_{K+1} < \cdots$$

so that K is the number of anchored islands that have their right-hand ends in the genome $(0, g)$. Clearly, $\{A_i, i \in Z\} \supseteq \{C_j, j \in Z\}$ and the set of locations $\{C_j, j \in Z\}$ is stationary.

The intensity ap_1 of the process $\{C_j, j \in Z\}$, that is, the expected number of right-hand ends of anchored islands per unit length, may be calculated as

$$\begin{aligned} &E(\text{number of anchored island ends in } (0, g)) \\ &= E\int_0^g 1\{\text{clone at } t \text{ is right end of} \\ &\qquad \text{anchored island} \mid \text{clone at } t\} P(\text{clone at } t) \\ &\equiv \int_0^g p(t)\, adt \\ &= gap_1, \end{aligned}$$

since stationarity shows that $p(t)$ is independent of t, the common value being

$$p_1 = P(\text{clone at } t \text{ is right-hand end of anchored island} \mid \text{clone at } t\}.$$

We define the function $\mathcal{F}$ by

$$\mathcal{F}(l) = P(Q > l) = \int_l^\infty f(r)\, dr, \quad l > 0,$$

and define the function J by

$$J(x) = \exp\left\{-a\int_x^\infty \mathcal{F}(l)\, dl\right\}, \quad x \geqslant 0.$$

We can interpret $J(x)$ as the probability that two points at distance x apart are not covered by a common clone. Observe that $J(0) = e^{-a}$, since rescaled clones have average length 1.

To calculate p_1, we proceed as follows. Let E be the event that the right-hand end of a given clone is the end of its anchored island, and suppose that this clone has length Q, with the right-hand end of the clone being located at 0. Let V be the distance back from 0 to the previous anchor. The conditional probability that this clone is the rightmost member of an anchored island is the probability that any clone that starts before $-V$ ends before 0 if $V \leq Q$ and is 0 if $V > Q$. The number of clones that start to the left of $-V$ and end to the right of 0 has a Poisson distribution with mean $a\int_V^\infty \mathcal{F}(t)dt$, so that the probability that there are no such events is $J(V)$. Hence given Q and V, the probability that a given clone is rightmost in its anchored island is $J(V)1\{V \leq Q\}$. Averaging over the distribution of Q and V gives

$$p_1 = EJ(V)1\{V \leq Q\} = \int_0^\infty be^{-bu}J(u)\mathcal{F}(u)du.$$

Since the intensity of clones to which this argument applies is a, the intensity of anchored islands is ap_1. With this background, we now turn to the proof of Proposition 3.

Proof of Proposition 3. To establish part (i), we observe that the number N_u of unanchored clones in the genome $(0, g)$ has expectation $EN_u = gaq_1$. The ergodic theorem implies that $N_u/g \to aq_1$ almost surely as $g \to \infty$. Since $N/g \to a$ almost surely, the number of unanchored islands is $N_u \sim gaq_1 \sim Nq_1$ for large g.

To establish part (ii), argue similarly concerning the number K of right-hand ends of anchored islands in the genome $(0, g)$, which has expectation $EK = gap_1$. Since $K/g \to ap_1$ and $N/g \to a$ almost surely as $g \to \infty$, the number of anchored islands is $K \sim gap_1 \sim Np_1$ for large g.

To establish part (iii), we introduce the process of anchored island masses. Let M_j be the number of clones in the jth anchored island. $M_1, \ldots, M_K$ are the masses of the K anchored islands whose right-hand ends fall in the genome $(0, g)$. The labeling of the points of the process destroys the stationarity of the property, in that none of $M_1, \ldots, M_K$ is "typical." In particular, the island of mass M_1 tends to straddle 0 and to be larger than typical. (What is stationary is the random measure $\sum_i M_i\delta_{C_i}$ with masses M_i at the random locations C_i, $i \in Z$.) To define the average mass EM of a typical anchored island, we proceed by a limiting argument.

Among anchored islands that have right-hand ends in $(0, g)$, the average number of clones per island is

$$\bar{M}_g = \frac{1}{K}\sum_{j=1}^{K} M_j.$$

Consider M_g', the difference between $\sum_1^K M_j$ and N_a, the number of anchored clones whose right-hand ends fall in $(0, g)$. Observe that the absolute value of M_g' is at most M'', the number of clones in the anchored islands that cover 0 or g, so that M'' has finite mean, and $M_g'/g \to 0$ almost surely as $g \to \infty$. Hence

$$\bar{M}_g = \frac{g}{K}\frac{1}{g}\sum_{j=1}^{K} M_j = \frac{g}{K}\left(\frac{N_a}{g} + \frac{M_g'}{g}\right).$$

As $g \to \infty$, $K/g \to ap_1$, $N_a/g \to a(1 - q_1)$ and $M_g'/g \to 0$ almost surely, so that

$$EM \equiv \lim_{g\to\infty} \bar{M}_g = \frac{(1 - q_1)}{p_1}. \qquad [1]$$

Hence we see that $\bar{M}_g \sim (1 - q_1)/p_1$, establishing part (iv).

We remark that the function $F(t)$, $t \geq 0$ defined by

$$F(t) = \lim_{g\to\infty} \frac{1}{K}\sum_{j=1}^{K} 1\{M_j \leq t\}$$

(the limit being almost sure and nonrandom) can be thought of as the distribution function of a random variable, which we denote by M. We will think of M as the "mass of a typical island." It can be shown that EM, computed from this distribution, agrees with the value above. In what follows, we implicitly use analogous constructions of typical quantities of interest.

To establish (iv), let E be the event that a given clone is a singleton anchored island. Suppose that the given clone has length Q and is located between $-Q$ and 0. Define V to be the distance forward from $-Q$ to the first anchor. Define W to be the distance back from 0 to the previous anchor (so that $W = Q - V$ if $V < Q$ and there are no other anchors in $(-Q, 0)$). Given (V, W) and $V + W < Q$, the conditional probability of E is the probability that no clones that start to the left of $-Q$ cover the anchor at $-Q + V$ and no clones that start in $(-Q, -W)$ cover 0. This probability is $J(V)J(W)/J(Q)$. If $V + W = Q$, the probability is $J(V)J(Q - V)/J(Q)$. One way to realize the random variable (V, W) is to construct (V', W') as independent and identically distributed exponential random variables with parameter b and define

$$(V, W) = (V', W') \quad \text{if } V' + W' \leq Q$$
$$= (V', 1 - V') \quad \text{if } V' < Q, \quad V' + W' > Q.$$

With this notation, the probability p_2 that a given clone is a singleton anchored island is

$$p_2 = E \frac{1\{V' < Q\} J(V') J(\min W', Q - V')}{J(Q)},$$

which can be evaluated as in (iv).

Next we define the process of anchored island lengths. Suppose that the jth anchored island has length S_j. We define the average ES of the length S of a typical island as an ergodic limit. Define

$$\bar{S}_g = \frac{1}{K} \sum_{j=1}^{K} S_j,$$

the average length of the anchored islands that have their right ends in the genome $(0, g)$. Since a point can fall at most in two anchored islands, it follows that

$$\frac{1}{g} \sum_{j=1}^{K} S_j \to r_1 + 2r_2, \qquad [2]$$

almost surely as $g \to \infty$, where r_i is the probability that a point is covered by precisely i anchored islands. Since $r_0 + r_1 + r_2 = 1$, we need only calculate r_0 and r_2 to calculate [2].

To calculate r_0 and r_2, let t be a point in the genome $(0, g)$, let W be the distance forward from t to the first anchor to the right, and let V be the distance back from t to the previous anchor to the left.

To calculate r_0, let E be the event that t is not covered by an anchored island, so that $r_0 = P(E)$. The event E occurs as long as any clones that start to the left of $t - V$ end before t and any clones that start in $(t - V, t)$ end before $t + W$. Given (V, W), the conditional probability of this is

$$P(E \mid V, W) = \exp\left(-a\left(\int_V^\infty \mathcal{J}(r)\,dr + \int_W^{V+W} \mathcal{J}(r)\,dr\right)\right)$$
$$= J(V)J(W)/J(V + W).$$

Averaging over the distribution of V and W gives

$$r_0 = E \frac{J(V)J(W)}{J(V + W)}, \qquad [3]$$

which proves part (vi).

To calculate r_2, let E be the event that t is covered by precisely two anchored islands so that $r_2 = P(E)$. The event E occurs if at least one clone that starts to the left of $t - V$ ends in $(t, t + W)$, no clones that start to the left of $t - V$ end after $t + W$, and at least one clone that starts in $(t - V, t)$ ends after $t + W$. Elementary manipulations with Poisson processes then show that

$$P(E \mid V, W) = \left(1 - \frac{J(V)}{J(V + W)}\right) J(V + W) \left(1 - \frac{J(W)}{J(V + W)}\right)$$
$$= \frac{(J(V) - J(V + W))(J(W) - J(V + W))}{J(V + W)}.$$

Averaging over the distribution of V and W shows that

$$r_2 = E \frac{(J(V) - J(V + W))(J(W) - J(V + W))}{J(V + W)}. \qquad [4]$$

It now follows from [2] that

$$\bar{S}_g \to ES \equiv \frac{r_1 + 2r_2}{ap_1} \qquad [5]$$

almost surely as $g \to \infty$. Since $r_1 + 2r_2 = 1 - r_0 + r_2$, it follows from [3] and [4] that

$$ap_1 ES = 1 - E(J(V) + J(W) - J(V + W))$$
$$= 1 + EJ(V + W) - 2EJ(V),$$

which establishes (v).

To establish (vii), let H_j denote the number of anchors in the jth anchored island and let R denote the number of anchors that fall in anchored islands whose right-hand ends fall in $(0, g)$. We define the average EH of the number H of anchors in a typical anchored island as an ergodic limit:

$$EH = \lim_{g \to \infty} \frac{1}{K} \sum_{j=1}^{K} H_j = \lim_{g \to \infty} \frac{g}{K} \frac{R}{g}.$$

Since the probability that an anchor is not covered by an island is $J(0) = e^{-a}$, the intensity of anchors involved in islands is $b(1 - e^{-a})$. Thus $R/g \to b(1 - e^{-a})$, so that $EH = b(1 - e^{-a})/ap_1$. This establishes (vii).

Finally, we prove (viii). Suppose that we are given an island of length Q, whose right-hand end is located at 0. Let V be the distance back from 0 to the first anchor. The conditional probability that this island is (a) anchored, (b) the end of its island, and (c) fol-

lowed by an actual ocean of length at least k is the probability that all clones that start to the left of 0 end to the left of 0 as long as $V < Q$, and is zero otherwise. This conditional probability is $J(0)\ 1\{V < Q\}e^{-ka}$, so that the probability we seek is $e^{-a(k+1)}P(V < Q) = e^{-a(k+1)}(1 - q_1)$, and the proof is completed by dividing by the probability p_1 that a given clone is the end of an anchored island. This completes the proof of Proposition 3.□

Corollary 4 follows by accounting for the effect of unanchored islands on the various average quantities in Proposition 3. Proposition 1 and Corollary 2 are the special case in which the rescaled clone length Q is identically 1. In this case, $\mathcal{F}(l) = 1$ if $l < 1$ and $\mathcal{F}(l) = 0$ if $l \geqslant 1$. It follows that

$$J(x) = e^{-a(1-x)}, \quad \text{if } x \leqslant 1$$
$$= 1, \quad \text{if } x > 1.$$

Similarly, Corollaries 5 and 6 represent special cases of particular probability density functions.

We should note that the expected values for all islands (Corollaries 2 and 4) may be less informative because they can be dominated by the effects of the unanchored islands. For $a \to \infty$ and any b fixed, the proportion of unanchored islands exhibits somewhat troublesome behavior: the expected proportion of unanchored islands tends to 1, even if b is so large that there is a high probability that there are no unanchored islands in the genome! It is not hard to see why. As $a \to \infty$, the anchored islands become few in number and large in size. However, there is a chance that there are some regions of the genome devoid of anchors and large enough to contain a clone. If so, the clones falling in such regions will contribute an infinite number of unanchored islands as $a \to \infty$. If not, there will be no unanchored islands whatsoever. In such cases, averages can be deceptive. In general, it is more helpful to focus attention on anchored islands.

Proof of Proposition 7. Let $Q = L/EL$ denote the normalized length of a random clone, so that $EQ = 1$. The function $J(u)$ can be expressed as

$$J(u) = \mathrm{e}^{-aE(\max(Q-u,0))}$$

and, because $\max(x, 0)$ is concave up, Jensen's inequality implies that

$$J(u) = \mathrm{e}^{-aE(\max(Q-u,0))} \geqslant e^{-a\max(EQ-u,0)} = J_f(u), \quad [6]$$

where $J_f(u)$ is the corresponding function for the case of constant normalized clone length $Q = 1$. We next note that Proposition 3 (ii) can be written as

$$ap_1 = \int_0^\infty be^{-bu}J'(u)du,$$

where J' is the derivative of J with respect to u. Integrating by parts and using the fact that $J(0) = e^{-a}$, we have

$$ap_1 = -be^{-a} + \int_0^\infty b^2e^{-bu}J(u)du. \quad [7]$$

The results now follow from [6] and [7].□

Proof of Proposition 8. We only sketch the proof, since it closely follows the arguments for Proposition 3 (ii), (v), and (vi). Let P be a point, V the distance from P to the closest anchor A_1 to the left, and W the distance to the closest anchor A_2 to the right. Let the random variables R, S, and T represent, respectively, the number of clones that cover A_1, P, and A_2; the number of clones that cover A_1 and P but not A_2; and the number of clones that do not cover A_1 but do cover A_2 and P. Conditional on V and W, the variables R, S, and T are independent Poisson variables with means:

$$E(R|V, W) = a\int_{V+W}^\infty \mathcal{F}(t)dt = -\log[J(V + W)],$$

$$E(S|V, W) = a\int_W^{V+W} \mathcal{F}(t)dt$$
$$= -\log[J(W)/J(V + W)],$$

and

$$E(T|V, W) = a\int_V^{V+W} \mathcal{F}(t)dt$$
$$= -\log[J(V)/J(V + W)].$$

In part (i), the point P is taken to be the right end of a clone C. The first term is the probability that A_1 lies within C. The bracketed term in the integrand represents the probability that C is the rightmost clone of an anchored k-linked island conditional on A_1 lying within C. The integral averages over the distribution V, W, and clone length. In parts (ii) and (iii), the point P is taken to be a random point in the genome. The bracketed term in the integrand is the probability that P lies in no (respectively, 2) anchored k-linked islands. The integrals average over the distribution of V and W.□

Proof of Proposition 9. The proof is analogous to that of Proposition 8, with the definitions of p_1, r_0, and r_2 modified to reflect the notion of k-covered as opposed to k-linked islands.□

3.3. *Mathematical Treatment of Anchor-Biased and Clone-Biased Sampling*

In this section, we provide a mathematical discussion of the question of anchor-biased or clone-biased sampling. Such sampling turns out to be interesting not only from an experimental standpoint but also from a mathematical standpoint because it allows one to calculate variances and covariances associated with quantities of interest. The following proposition explains this.

PROPOSITION 10. *Let S', M', and H' denote, respectively, the length, number of clones, and number of anchors contained in a randomly chosen island (which may be anchored or unanchored). The expectation of these quantities is given in Corollary 4.*

(i) *Let S_c and M_c be, respectively, the length and number of clones in the island containing a randomly chosen clone. Then $EM'S' = EM'ES_c$ and $EM'^2 = EM'EM_c$.*

(ii) *Let S_a and M_a be the analogous quantities for the island containing a randomly chosen anchor (which may be an empty island if the anchor does not fall in an island). Then $EH'S' = EH'ES_a$ and $EH'M' = EH'EM_a$.*

Proof. To establish (i), let K' be the number of islands with right-hand ends in $(0, g)$ (so that $K' = K + N_u$), let S'_j be the length of the jth such island, and let M'_j be its mass (i.e., number of clones). Note that

$$\frac{1}{K'}\sum_{j=1}^{K'} M'_j S'_j = \frac{1}{K'}\sum_{i=1}^{N} S'_{Ji} + \epsilon_g,$$

where J_i is the label of the island containing the ith clone, S'_{Ji} is its length, and ϵ_g is an end-effects term. Methods analogous to those used to calculate ES in [5] show that $\epsilon_g/g \to 0$ almost surely as $g \to \infty$. Since $\{S'_{Ji},\ i \in Z\}$ is a stationary process, the ergodic theorem guarantees that

$$\frac{1}{N}\sum_{i=1}^{N} S_{Ji} \to ES_c$$

as $g \to \infty$. Since $N/K' \to EM'$, we see that

$$\frac{1}{K'}\sum_{j=1}^{K'} M'_j S'_j \to EM'ES_c$$

almost surely as $g \to \infty$. Hence $EM'S' = EM'ES_c$.

Similarly, we observe that

$$\sum_{j=1}^{K'} M'^2_j = \sum_{j=1}^{K'} \Big(\sum_{i=1}^{N} 1\{J_i = j\}\Big)^2 + \epsilon'_g,$$

where once more the end-effects term $\epsilon'_g/g \to 0$ almost surely as $g \to \infty$. Therefore

$$EM'^2 = \lim_{g\to\infty} \frac{1}{K'}\sum_{j=1}^{K'} M'^2_j = EM'EM_c$$

almost surely. This establishes part (i). Analogous arguments establish part (ii).□

Remark. These results may be converted into results about anchored islands in the following way. An unanchored island is a clone containing no anchors, with expected length λ' given in Proposition 4 (iv), and mass identically 1. The proportion of anchored islands among all islands is $p_1/(p_1 + q_1)$. Write S, M, and H, without primes, to refer to randomly chosen anchored islands. Then

$$(p_1 + q_1)ES' = p_1ES + q_1 1$$

gives the relation between Proposition 3 (iv) and Corollary 4 (iv), while

$$(p_1 + q_1)EM' = p_1EM + q_1 1$$

gives the relation between Proposition 3 (ii) and Corollary 4 (ii). The relations

$$(p_1 + q_1)EM'S' = p_1EMS + q_1\lambda'$$

and

$$(p_1 + q_1)EM'^2 = p_1EM^2 + q_1 1$$

are needed to convert Proposition 10 into a statement about anchored islands.

The expected values under clone-biased and anchor-biased sampling can be computed by defining particular renewal processes. As it happens, the results cannot be expressed in closed form but rather in terms of functions defined on the set $\Delta = \{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}$ that are the solution to certain integral equations. Although they may look daunting, they can be computed by numerical methods. For simplicity, we state the results only for the case of fixed-length clones.

PROPOSITION 11. *With notation as above, we have:*

(i) *Given a randomly chosen clone, the expected length of the island containing it is $ES_c = 1 + 2f_1(0, 0)$, where $f_1(x, y)$ is the function defined on Δ satisfying the integral equation*

$$f_1(x, y) = \iint_{\mathcal{R}(x,y)} [1 - u - v + f_1(u, v)] \times ae^{-au}be^{-bv}dudv,$$

where $\mathcal{R}(x,y) = \{(u,v): u,v \geq 0, u+v \leq 1-x, v \leq 1-x-y\}$. *Note that this expectation includes unanchored clones.*

(ii) *Given a randomly chosen clone, the expected number of clones in the island containing it is* $EM_c = 1 + 2f_2(0,0)$, *where* $f_1(x,y)$ *is the function defined on* Δ *satisfying the integral equation*

$$f_2(x,y) = \iint_{\mathcal{R}(x,y)} [1 + a(1-u-v-x) + f_2(u,v)]ae^{-au}be^{-bv}dudv.$$

Note that this expectation includes unanchored clones.

(iii) *Given a randomly chosen anchor, the expected length of the island containing it is*

$$ES_a = 2\int_0^1 \int_0^\infty [\max(x-y,0) + f_3(0, \min(x,y))]ae^{-ax}be^{-by}dxdy,$$

where $f_3(x,y)$ *is the function defined on* Δ *satisfying the integral equation*

$$f_3(x,y) = \iint_{\mathcal{S}(x,y)} [1-u-v-x + f_3(u,v)]ae^{-au}be^{-bv}dudv + \iint_{\mathcal{S}'(x,y)} (1-u)ae^{-au}be^{-bv}dudv + \int_{\mathcal{S}''(x,y)} yae^{-au}du,$$

where

$$\mathcal{S}(x,y) = \{(u,v): u,v \geq 0, 1-u-v \geq y, u \leq 1-x-y\},$$

$$\mathcal{S}'(x,y) = \{(u,v): u,v \geq 0, 1-u > x+y, 1-u-v \leq y\},$$

$$\mathcal{S}''(x,y) = \{(u,v): u \geq 0, u > 1-x-y\}.$$

Note that the length of the island containing a random anchor is zero if that anchor is not covered by any clones.

(iv) *Given a random anchor, the expected number of clones in the island containing it is*

$$EM_a = -a + 2\int_0^1 \int_0^\infty f_4(0, \min(x,y)) \times ae^{-ax}be^{-by}dxdy,$$

where $f_4(x,y)$ *is the function defined on* Δ *satisfying the integral equation*

$$f_4(x,y) = \iint_{\mathcal{S}(x,y)} [1 + a(1-u-x) + f_4(u,v)] \times ae^{-au}be^{-bv}dudv + \iint_{\mathcal{S}'(x,y)} (2 + a(1-u-x))ae^{-au}be^{-bv}dudv + \int_{\mathcal{S}''(x,y)} (1+ay)ae^{-au}du.$$

Note that the number of clones in the island containing a random anchor is zero if that anchor is not covered by any clones.

(v) *Given a random anchor that is contained within at least one clone, the expected length of the island containing it is* $ES_a/(1-e^{-a})$ *and the expected number of clones in the island containing it is* $EM_a/(1-e^{-a})$, *where* ES_a *and* EM_a *are defined in* (iii) *and* (iv), *respectively.*

Proof of Proposition 11. Let C_1 be a random clone, which will be arbitrarily taken to correspond to the interval $[0, 1]$. Let $(x, y) \in \Delta$. Suppose that we are given the additional information that there are no left clone ends in $(0, x]$, that there is an anchor at x, that there are no anchors in $(x, x+y]$, and that there is a right clone end at $x+y$. (See Fig. 6.) Define $L(x,y)$ to be the distance from the right end of the clone C_1 to the end of its island. We derive a renewal equation satisfied by $f_1(x,y)$, the mean of $L(x,y)$.

Define V to be the (random) distance back from the right end of C_1 to the rightmost anchor in C_1, and let U be the distance back from this anchor to the rightmost clone end occurring to the left of the anchor. Define the region $\mathcal{R}(x,y)$ as in part (i) above and suppose that $(U,V) = (u,v) \in \mathcal{R}(x,y)$. From Fig. 6, we see that the clone C_2 that starts at $1-u-v$ extends the island containing C_1, and that given (U,V)

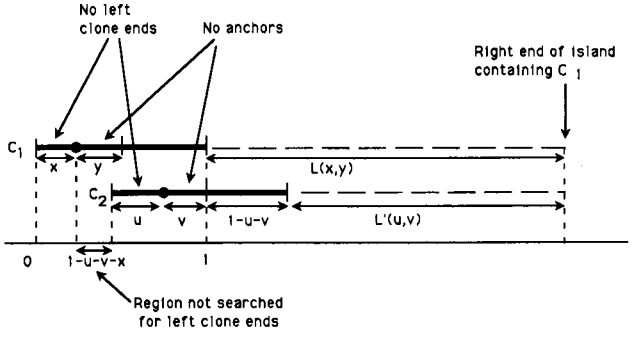


FIG. 6. Diagram illustrating renewal process used in the Proof of Proposition 11 (i) and (ii).

$= (u, v)$ we have $L(x,y) = 1 - u - v + L'(u,v)$, where $L'(u, v)$ has the same probability law as $L(u, v)$. If, on the other hand, $(U, V) \notin \mathcal{R}(x,y)$, then clone C_1 is the rightmost clone in its island, and so $L(x, y) = 0$. Restricted to the set $\mathcal{R}(x,y)$, U and V have the same distribution as independent exponential random variables with parameters a and b, respectively. Averaging over the distribution of (U, V) shows that

$$f_1(x,y) = \iint_{\mathcal{R}(x,y)} [1 - u - v + f_1(u,v)]ae^{-au}be^{-bv}dudv$$

as required. Note that a typical clone contains no additional information of the sort described above for C_1—that is, we should take $x = y = 0$ to obtain the expected additional length from the right end of the typical clone to the right end of its island. Since the process is symmetrical, this is also the expected additional length from the left end of the clone to the left end of its contig. This establishes part (i) of the proposition.

The same renewal argument may be used to calculate $f_2(x,y)$, the mean of the number $N(x,y)$ of additional clones in the island containing C_1 that end to the right of C_1. Once again, we condition on the event $U = u$, $V = v$, and suppose first that $(u,v) \in \mathcal{R}(x,y)$. Conditional on this, we see from Fig. 6 that $N(x,y) = 1 + P(u,v) + N'(u,x)$, where $P(u,v)$ is the number of clones that start in the interval $(x, 1 - u - v)$ (and so has a Poisson distribution with mean $a(1 - u - v - x)$) and $N'(u,v)$ has the same probability law as $N(u,v)$. On the other hand, if $(U, V) \notin \mathcal{R}(x,y)$, then clone C_1 is the rightmost in its island, and so $N(x,y) = 0$. Averaging the conditional expectations over the distribution of (U, V) shows that

$$f_2(x,y) = \iint_{\mathcal{R}(x,y)} [1 + a(1 - u - v - x) + f_2(u,v)] \times ae^{-au}be^{-bv}dudv.$$

The results (ii) now follow from the same argument that verified (i).

To establish (iii) and (iv), a different renewal argument is needed as depicted in Fig. 7. Let A be a random anchor, which will be arbitrarily assumed to lie at 0. For $(x,y) \in \Delta$, suppose that there is a right clone end at y, that there are no anchors in $(0,y)$, and that no further right clone ends occur in $(y, y + x)$. Let C_1 denote the clone that ends at y (see Fig. 7). Define $L(x,y)$ to be the distance from A to the right end of the island containing C_1.

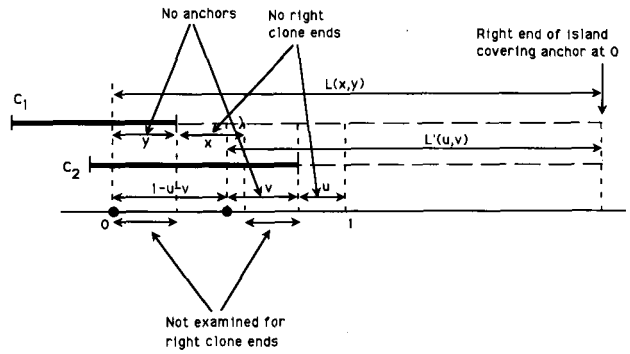


FIG. 7. Diagram illustrating renewal process used in the Proof of Proposition 10 (iii) and (iv).

Starting from 1, we look back a distance U to the first right clone end, and a further distance V back to the first anchor. Define the sets $\mathcal{S}(x,y)$, $\mathcal{S}'(x,y)$, and $\mathcal{S}''(x,y)$ as in part (iii) above. Let U' and V' be independent exponential random variables with parameters a and b, respectively. Using Fig. 7, we see that the random variables U and V may be defined in terms of U' and V' as follows:

$$(U, V) = (U', V') \quad \text{if } (U', V') \in \mathcal{S}(x,y)$$
$$(U, V) = (U', 1 - U') \quad \text{if } (U', V') \in \mathcal{S}'(x,y)$$
$$U = 1 - y \quad \text{if } U' \in \mathcal{S}''(x,y)$$

Observe, from Fig. 7, that if $(U', V') = (u,v) \in \mathcal{S}(x, y)$, then $U = u$ and $V = v$ and $L(x,y) = 1 - u - v + L'(u,v)$, where $L'(u,v)$ has the same probability law as $L(u,v)$. If, however, $(U', V') = (u,v) \in \mathcal{S}'(x, y)$, then $U = u$ and the island containing the anchor A ends at $1 - u$, so that $L(x,y) = 1 - u$. Finally, if $U' = u \in \mathcal{S}''(x,y)$, then $U = 1 - y$ and the island containing the anchor A ends at y. Hence $L(x,y) = y$. Averaging over the distribution of (U, V) shows that $f_3(x,y)$, the mean of $L(x,y)$, satisfies the integral equation

$$f_3(x,y) = \iint_{\mathcal{S}(x,y)} [1 + a(1 - u - x) + f_4(u,v)] \times ae^{-au}be^{-bv}dudv + \iint_{\mathcal{S}'(x,y)} (2 + a(1 - u - x))ae^{-au}be^{-bv}dudv + \int_{\mathcal{S}''(x,y)} (1 + ay)ae^{-au}du.$$

To complete the proof of (iii), observe that the situation of a given anchor can be described by $x = 0$. The condition involving y says that there is a clone covering the anchor A such that A is the rightmost anchor in the clone (see Fig. 7). For a random anchor, y is the

value of a random variable Y' with the following structure. Start looking for right clone ends starting from an anchor at A, say. Let X' be the distance to the first such clone end. Let Y' be the distance back from this right clone end to the first anchor (which may be A). One way to realize the pair (X', Y') is to take X, Y as independent exponential random variables with parameters a and b, respectively, and then set $(X', Y') = (X, \min(X, Y))$. The length to the right of A of the island containing A is then $X' - Y' + f_3(0, Y')$ if $X' \leq 1$ (in which case there is a clone covering A). In the event that $X' > 1$, there is no clone covering A and we define the length of the "island" containing A to be zero. Written in terms of (X, Y), we see that the expected length of the island containing a randomly chosen anchor is given by $E((\max(X, Y, 0) + f_3(0, \min(X, Y)\, 1\{X \leq 1\})$. By symmetry, the expected length of the portion of the island to the left of A is the same. Writing this expectation as an integral gives the result in (iii).

Next, we prove (iv). We use the same renewal structure as that for (iii). Define $L(x, y)$ to be the number of clones that belong to the island containing C_1 and that lie (in whole or in part) to the right of A. From Fig. 7, observe that if $(U', V') = (u, v) \in \mathcal{S}(x, y)$, then $U = u$ and $V = v$ and $L(x, y) = 1 + P(u, v) + L'(u, v)$, where $L'(u, v)$ has the same probability distribution as $L(u, v)$, and $P(u, v)$ is the number of clones that have right ends in the intervals $(0, y)$ and $(x + y, 1 - u)$ (so that $P(u, v)$ has a Poisson distribution with mean $a(1 - u - x)$). If, however, $(U', V') = (u, v) \in \mathcal{S}'(x, y)$, then $U = u$ and the island containing the anchor A ends at $1 - u$, so that $L(x, y) = 2 + P(u, v)$. Finally, if $U' = u \in S''(x, y)$, then $U = 1 - y$ and the island containing the anchor A ends at y. Hence, $L(x, y) = 1 + P'(u, v)$, where $P'(u, v)$ is the number of clones with right ends in $(0, y)$, which is Poisson distributed with mean ay. Averaging over the distribution of (U, V) shows that $f_4(x, y)$ satisfies the integral equation

$$f_4(x, y) = \iint_{\mathcal{S}(x,y)} [1 + a(1 - u - x) + f_4(u, v)] \times ae^{-au}be^{-bv}dudv + \iint_{\mathcal{S}'(x,y)} [(2 + a(1 - u - x))ae^{-au}be^{-bv}dudv + \int_{\mathcal{S}''(x,y)} (1 + ay)ae^{-au}du.$$

To show the results in (iv), observe that a typical anchor has $x = 0$ and variable Y' with the following structure. As in the previous part, take X, Y independent exponential random variables with parameters a and b, respectively, and set $(X', Y') = (X, \min(X, Y))$. The number of clones ending to the right of 0 in the island containing the staple at 0 is then $f_4(0, Y')$ if $X' \leq 1$ and 0 otherwise. Written in terms of (X, Y), we see that the expected number of clones that end to the right of 0 in the island containing an anchor at 0 is given by $E(f_4(0, \min(X, Y)1\{X \leq 1\})$. By symmetry, the mean number of clones to the left of 0 is the same. We then need to subtract the expected number of clones that cover 0 (as these will be overcounted otherwise). This is just a. Combining these results and writing the expectation as an integral give the result in (iv). Finally, part (v) follows by conditioning. This completes the proof of Proposition 11.□

The implicit definitions in Proposition 11 are less convenient than explicit closed-form formulas, but the results can be calculated using standard methods for solving integral equations (Delves and Mohamed, 1985).

4. CONCLUSION

Considerable effort is already being focused on the construction of complete physical maps of the human and mouse genomes and further projects are being considered for various plant and animal species. Although many nonmathematical factors bear on the design of such projects (including the relative merits of different cloning vectors and anchoring techniques, uneven representation of clone and anchor libraries, etc.), a detailed mathematical analysis must also be a prerequisite. The results above should be of some value in designing and monitoring the progress of physical mapping projects.

ACKNOWLEDGMENTS

We are grateful to Mark Eggert for assistance in computer programming. This work was supported in part by grants from the National Science Foundation (DMS 90-05833 to R.A.A., S.T., and M.S.W. and DCB-8611317 to E.S.L.), the National Institutes of Health (HG00198 to E.S.L., GM41746 to S.T., and GM36230 to M.S.W.), the Markey Foundation (to E.S.L.), and the System Development Foundation (to E.S.L.).

REFERENCES

1. BARRILOT, E., DAUSSET, J., AND COHEN, D. (1991). Theoretical analysis of a physical mapping strategy using random single copy landmarks. *Proc. Natl. Acad. Sci. USA* **88**: 3917–3921.
2. BREIMAN, L. (1968). "Probability," Addison–Wesley, Reading, MA.
3. COULSON, A., SULSTON, J., BRENNER, S., AND KARN, J. (1986). Toward a physical map of the genome of the nema-

tode Caenorhabditis elegans. *Proc. Natl. Acad. Sci. USA* **83:** 7821–7825.

4. DELVES, L. M., AND MOHAMED, J. L. (1985). "Computational Methods for Integral Equations," Cambridge Univ. Press, London/New York.
5. GREEN, E. D., AND OLSON, M. V. (1990). Systematic screening of yeast artificial-chromosome libraries using the polymerase chain reaction. *Proc. Natl. Acad. Sci. USA.*
6. KOHARA, Y., AKIYAMA, A., AND ISONO, K. (1987). The physical map of the whole E. coli chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50:** 495–508.
7. LANDER, E. S., AND WATERMAN, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.
8. OLSON, M. V., DUTCHIK, J. E., GRAHAM, M. Y., BRODEUR, G. M., HELMS, C., FRANK, M., MACCOLLIN, M., SCHEINMAN, R., AND FRAND, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83:** 7826–7830.
9. SEED, B. (1980). Purification of genomic sequences from bacteriophage libraries by recombination and selection *in vivo*. *Nucleic Acids Res.* **11:** 2427–2445.
10. TORNEY, D. (1991). Mapping using unique sequences. *J. Mol. Biol.* **217:** 259–264.