

Sequencing a Genome by Walking with Clone-End Sequences: A Mathematical Analysis

Serafim Batzoglou,¹ Bonnie Berger,^{1,2} Jill Mesirov,⁴ and Eric S. Lander³⁻⁵

¹Laboratory for Computer Science and Departments of ²Mathematics and ³Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA; ⁴Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142 USA

One approach to sequencing a large genome is (1) to sequence a collection of nonoverlapping “seeds” chosen from a genomic library of large-insert clones [such as bacterial artificial chromosomes (BACs)] and then (2) to take successive “walking” steps by selecting and sequencing minimally overlapping clones, using information such as clone-end sequences to identify the overlaps. In this paper we analyze the strategic issues involved in using this approach. We derive formulas showing how two key factors, the initial density of seed clones and the depth of the genomic library used for walking, affect the cost and time of a sequencing project—that is, the amount of redundant sequencing and the number of steps to cover the vast majority of the genome. We also discuss a variant strategy in which a second genomic library with clones having a somewhat smaller insert size is used to close gaps. This approach can dramatically decrease the amount of redundant sequencing, without affecting the rate at which the genome is covered.

The complete DNA sequence of a genome is a powerful tool for studying an organism. Biological research in the 21st century will surely require obtaining the sequence of large numbers of important organisms, including many higher animals and plants with large genomes.

Various approaches have been proposed for sequencing large genomes, which broadly fall into two categories: (1) whole-genome shotgun sequencing and (2) clone-based shotgun sequencing.

Shotgun sequencing involves breaking a target into random fragments, sequencing these fragments, and reconstructing the full sequence from these pieces. Shotgun sequencing was invented by Sanger, who applied it to the genome of bacteriophage λ (Sanger et al. 1980, 1982). It was subsequently extended to genomes of large recombinant plasmids, large viruses, mitochondria, chloroplasts, and bacteria (Ohyama et al. 1986; Goebel et al. 1990; Oda et al. 1992; Fleischmann et al. 1995). More recently, Weber and Myers (1997) proposed that the approach could be extended to very large genomes, such as the human genome, by making greater use of long-range linking information, for example, sequences at the opposite ends of large-insert clones. There are many potential pitfalls, such as the possibility that the huge number of repeats (e.g., one million copies of the Alu repeat in the human genome) may result in many large-scale sequence misassemblies that are hard to detect and correct (Green 1997).

Clone-based shotgun-sequencing involves obtaining a collection of large-insert clones covering a ge-

nome and performing shotgun sequencing on each clone. This approach was used for sequencing the genomes of the yeast *Saccharomyces cerevisiae* (Oliver et al. 1992; Dujon 1996) and *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998) and is being used for sequencing of mammalian genomes such as the human and mouse. Clone-based sequencing has the advantage that it eliminates the possibility of large-scale misassemblies, because each clone is assembled individually, but it requires that one generate the collection of clones covering the genome. It also has the drawback that there is some redundant sequencing as a result of the overlap of adjacent clones. (There is no workable technique for shotgun sequencing, only the nonoverlapping portion of a clone. Each clone must be subjected to full shotgun sequencing, with the overlapping portion therefore being shotgun-sequenced twice. Such duplication can be avoided in the directed finishing phase, in which gaps and ambiguities in the shotgun sequence are resolved. Such overlaps thus increase the cost for the overlapping region by somewhat less than twofold.)

The current cloning system of choice for clone-based sequencing is the bacterial artificial chromosome (BAC) with inserts of ~200 kb. The discussion below will refer to BACs (rather than generic “large-insert clones”), but the results, of course, are completely general.

Two approaches have been proposed for generating the BAC clones to be sequenced: (1) physical mapping (or “map first, sequence second”) and (2) walking (or “map as you go”).

Physical mapping involves constructing a complete physical map covering the genome before begin-

⁵Corresponding author.
E-MAIL lander@genome.wi.mit.edu; FAX (617) 252-1902.

ning sequencing (map first, sequence second). One “fingerprints” a BAC library, using a technique such as restriction digestion to characterize each clone, and then attempts to use this information to infer the order of the clones. A “path” of clones is then selected for sequencing. This approach was successfully used to generate physical maps of cosmids used to sequence the *S. cerevisiae* and *C. elegans* genomes (Coulson et al. 1986; Olson et al. 1986). It is being applied to the human genome, although the problem is more challenging because of the larger clones sizes (yielding more complex fingerprints), the larger genome size, and the more complex repeat structure.

The walking approach proceeds directly to sequencing without a prior physical map: One starts by sequencing an initial collection of random clones and then “walks” the genome by iteratively selecting minimally overlapping clones. Venter et al. (1996) proposed an efficient method for selecting minimally overlapping clones, which is commonly referred to as “BAC-end sequencing” or, more formally, as “sequence-tagged connectors” (STCs). A BAC library is characterized by sequencing the inserts of each clone at its two ends (using primers located in the two vector arms), and a database of the resulting BAC-end sequences is created. Given a fully sequenced BAC clone *C*, one can walk by searching the database to identify all overlapping BACs. The location and orientation of each overlapping BAC is immediately apparent from the position of its end sequence within *C* (Fig. 1). One can also tell whether an overlapping BAC closes a gap in the genomic sequence by whether its other end sequence lies in another fully sequenced BAC clone *C'*.

The purpose of this paper is to analyze the strategy of sequencing a genome by clone-based walking. Al-

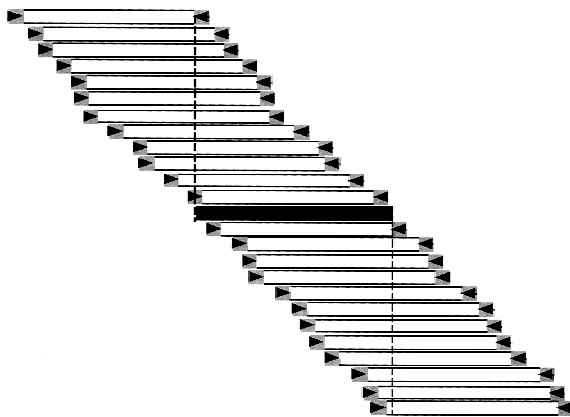


Figure 1 Overlapping BACs in library with depth $d = 12$ -fold. A typical BAC (shown in black) will tend to have 12 BACs overlapping its right end and 12 BACs overlapping its left end. Each overlapping clone can be identified because one of its end sequences (shown by arrowheads) is contained within the sequence of the target BAC. The precise position of each overlapping clone is inferred by the location and orientation of the end sequence within the sequence of the target BAC.

though the basic notion of walking is straightforward, key strategic questions have not been addressed: How many independent walks should be performed in parallel? With too few, one will need too many walking steps to cover the genome in a reasonable time. With too many, one may perform too much redundant sequencing as walks “bump” into one another with random—and therefore suboptimal—overlap. How can one maximize the speed of covering the genome while minimizing the amount of redundant sequencing?

The purpose of this paper is to present a mathematical analysis that expresses the amount of redundant sequencing in terms of the initial density of walks and the depth of the BAC library. We also describe and analyze a variant strategy in which one uses a second BAC library with smaller insert clones to close gaps efficiently; this strategy dramatically reduces the amount of redundant sequencing. The results provide direct guidance for planning a genome sequencing project.

Basic Model

We begin by describing the basic model to be analyzed.

BAC Library

We will initially consider the case of a single BAC library with the following properties: (1) The clones have constant size L . The clone size L will serve as our unit of distance, so that we may set $L = 1$. (Later, we will consider the use of a second BAC library with inserts of constant size $L' < 1$.) (2) The clones are random segments of the genome. We thus ignore the possibility of cloning bias. (3) The library covers the genome to depth d ; that is, the average number of clones covering a point in the genome is d . (4) The clones in the library have all been sequenced at both ends, yielding sufficient unique sequence to reliably detect overlap. We thus ignore the possibility that some BAC ends may not have been sequenced because of technical errors or may contain repeat sequences that make it impossible to recognize overlap.

Sequencing

We assume that the procedure for genomic sequencing satisfies the following conditions: (1) The cost of sequencing a BAC is directly proportional to the length of its insert. This agrees with current practice in large-scale genomic sequencing centers, which perform a fixed number of shotgun sequences per kilobase of insert size. [A constant number of reads per kb may not be precisely optimal. On one hand, it has been suggested that smaller clones may require slightly fewer reads per kb to reach closure (Roach 1995). On the other hand, smaller clones may require slightly more reads per kb of insert, because the cloning vector comprises a slightly larger proportion of the total clone length and thus of the reads from the shotgun library.

In any case, these effects are small and offsetting.] (2) The cost of producing a small-insert shotgun library from a BAC (which involves preparing, shearing, and cloning DNA) is negligible compared with the cost of performing the shotgun reads needed to sequence the BAC. This reflects current practice, for which the former is only ~2% of the latter.

In principle, one can sequence the genome with minimal overlap by sequencing a single initial seed clone C_0 and then walking successively outward by sequencing minimally overlapping clones, C_{-i} and C_i , on each side until one covers the entire genome (Fig. 2). The resulting sequence of clones is commonly called a minimal tiling path. (Strictly speaking, it should be referred to as a minimal tiling path for the given library and choice of starting clone.) At each step, minimally overlapping clone C_i is identified by comparing the database of BAC-end sequences with the completed sequence of C_{i-1} to find the BAC-end sequence that lies closest to the growing end and points outward (Fig. 1). We assume that there is always at least one overlapping clone pointing outward. This is a reasonable assumption, provided that the depth d is sufficiently large. For example, a BAC library with 200-kb inserts covering a mammalian genome of 3×10^9 bp to depth $d = 10$ should yield only 6.8 gaps—ignoring chromosome ends (Lander and Waterman 1988).

It is straightforward to analyze the expected amount of redundant sequencing. On average, each successive clone overlaps the previous sequence by $(1/d)$ of its length and provides $[1 - (1/d)]$ of new sequence. The ratio of redundant sequence to unique sequence is thus: $R = d^{-1} / (1 - d^{-1}) = 1 / (d - 1)$. A BAC library with depth $d = 10$ thus yields a minimal tiling path with redundant sequencing of 11.1%, whereas one with depth $d = 20$ entails redundant sequencing of 5.3%.

In reality, sequencing large genomes by serial walking is completely impractical owing to the cycle time to process each BAC clone: A mammalian genome would require 15,000 serial steps. Sequencing each BAC requires growing the clone, preparing DNA, shearing the DNA, constructing a small-insert shotgun library, performing shotgun sequencing of clones from the small-insert library, assembling the reads by computer, and closing remaining gaps. Various quality assessment steps are performed along the way to ensure high yield. Although each individual step is straight-

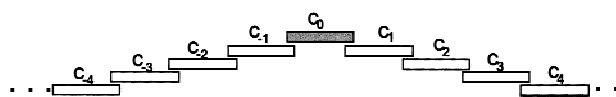


Figure 2 Serial walking of the genome from a single initial clone C_0 .

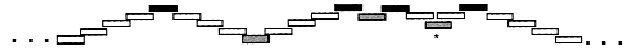


Figure 3 Serial walking of the genome from a collection of seed clones (shown in black). Bidirectional walking steps are shown, with steps that close an ocean shown in grey. The oceans are closed with relatively little overlap in the first two cases but with large overlap in the third case [marked with an asterisk (*)].

forward, the overall elapsed time is currently on the order of 1–2 months.

The obvious solution is to process many BACs in parallel, that is, to simultaneously walk from many seed clones (Fig. 3). Ideally, the seed clones should be dense enough that one can cover the vast majority of the genome in a modest number of walking steps. A reasonable goal would be to cover a mammalian genome in ~1 year.

Parallel walking, however, introduces a problem: The various walks may join with large overlaps, substantially increasing the proportion of redundant sequencing. Figure 3, for example, shows three clones (shown in grey) that join walks. The first two instances involve walks that fortuitously meet with relatively little overlap, but the third instance (marked with an asterisk) involves a walk that meets with large overlap and thereby results in substantial redundant sequencing.

It is important to understand how the expected proportion of redundant sequencing depends on the density of seed clones and the depth of the library. The following section presents the mathematical analysis. The reader interested primarily in applications may wish to proceed directly to Results.

Mathematical Analysis

Our goal is to derive simple formulas providing a good approximation for the proportion of excess sequencing. Toward this end, we will make certain simplifying assumptions. The reasonableness of the approximations will then be demonstrated by their close agreement with simulations, below.

We start with a collection of nonoverlapping seed clones. Following the terminology of Lander and Waterman (1988), each clone is an “island” followed by an “ocean” to be walked. Walking proceeds outward from each clone in a bidirectional fashion, with overlapping clones recognized on the basis of their end sequence. At each round, one identifies instances in which two islands can be joined by a single clone (readily apparent from the two end sequences) and ensures that a walking step is taken from only one of the two islands (to avoid unnecessary excess sequencing).

Although walking proceeds bidirectionally in practice, it simplifies the discussion to suppose that walking proceeds unidirectionally to the right. Each bidirectional walking step is clearly equivalent to two unidirectional walking steps. Figure 4 shows a typical seed clone C_0 followed by consecutive walking steps C_1 ,

C_2, \dots, C_j , where C_j is the first clone that overlaps the seed clone to the right of C_0 (a fact that is readily apparent from its end sequence).

Suppose that the seeds cover proportion π of the genome in clones of length 1. The oceans therefore cover proportion $1 - \pi$, and the mean ocean size must be $\omega = (1 - \pi) / \pi$.

Our first simplifying assumption concerns ocean lengths:

Assumption of exponential oceans (AEO). Nonoverlapping seed clones will be assumed to be chosen such that the resulting ocean lengths follow an exponential distribution (with mean ω).

What is the proportion of redundant sequencing entailed in sequencing the island C_0 together with the ocean on its right? Let the random variable J denote the sum of the lengths of C_1, C_2, \dots, C_j . (Because the clones have unit length, $J = j$.) The amount of total sequencing is $J + 1$, whereas the amount of unique sequence obtained is $X + 1$ (see Fig. 4). The expected proportion of redundant sequencing is thus

$$R = \frac{E(J) + 1}{E(X) + 1} - 1$$

where $E(\cdot)$ denotes the expected value. As noted above, the mean ocean size $E(X)$ is $\omega = (1 - \pi) / \pi$. Therefore, we simply need to calculate $E(J)$.

Calculating $E(J)$ precisely is complicated, but it is possible to make a good estimate by using a second simplifying assumption:

Assumption of constant overlap (ACO). Each clone C_i will be assumed to overlap the previous clone C_{i-1} by exactly the expected amount of overlap, $1 / d$, and thus to extend the sequence by exactly the expected amount, $1 - (1 / d)$.

ACO greatly simplifies the mathematical analysis, owing to an elegant property of exponential distributions: If each island is extended by a constant amount, then the lengths of the remaining oceans (i.e., those that are not closed) follow the same exponential distribution as the initial oceans. This stability property allows the walking process to be analyzed by a simple recursion.

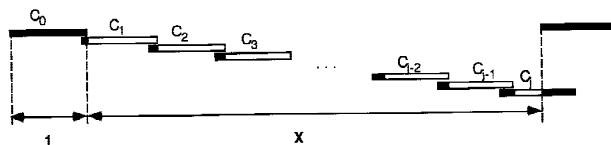


Figure 4 Unidirectional walking from a seed clone C_0 . Each successive clone C_i extends the island, with the redundant sequence shown in grey. The last clone C_j closes the ocean, with redundant sequencing due to both overlap with C_{j-1} and with the next seed clone to the right. Each clone has size 1, and the ocean has size X . The number of clones sequenced is $j + 1$, with each having length 1. The amount of unique sequence obtained (the initial island together with the ocean) is $X + 1$.

Proposition 1

Suppose that the genome is sequenced by seeding with nonoverlapping clones to coverage π and then walking using a library with depth d . Let $\omega = (1 - \pi) / \pi$. Assuming that the ocean sizes are exponentially distributed (AEO) and consecutive walking steps have constant overlap (ACO), we have the following: (i) The expected proportion of redundant sequencing is

$$R(d, \omega) = \frac{E(J) + 1}{\omega + 1} - 1$$

(ii) $E(J) = 1 / p_{d, \omega}$, where $p_{d, \omega} - 1 = e^{-(1 - d^{-1})/\omega}$ is the probability that an ocean is closed with a single walking step. (iii) The proportion of genome not sequenced after k unidirectional walking steps is $(1 - \pi)(1 - p_{d, \omega})^k = (1 - \pi)e^{-k(1 - d^{-1})/\omega}$. The proportion thus decreases geometrically with each walking step.

Proof

Part i was noted above. A formal proof follows from a straightforward application of Wald's equation (see Ross 1970). Part ii uses ACO. The total clone length J required to close an ocean can be calculated by considering a single walking step. With probability $p_{d, \omega}$, the step closes the ocean, resulting in a total clone length of 1. With probability $1 - p_{d, \omega}$, the step fails to close the ocean and leaves a remaining ocean having the same exponential distribution. It follows by recursion that $E(J) = (p_{d, \omega}) 1 + (1 - p_{d, \omega}) [1 + E(J)]$. The desired result follows by solving for $E(J)$. Part iii follows by observing that the lengths of the remaining oceans after k walking steps continue follow the same exponential distribution. The total ocean length remaining after k steps is thus directly proportional to the proportion of oceans that remain unclosed after k steps. The proportion of the remaining oceans that are closed at each walking step is $p_{d, \omega}$, and thus the proportion remaining unclosed after k steps is $(1 - p_{d, \omega})^k$. This completes the proof.

Proposition 1(i) can be generalized to any distribution of initial oceans sizes as follows:

Proposition 2

Suppose that the genome is seeded as in proposition 1 but that the ocean sizes x have probability density $f(x)$. Assuming ACO, the expected proportion of redundant sequencing is

$$R(d, \omega) = \frac{E(J) + 1}{\omega + 1} - 1$$

where

$$E(J) = \int_0^\infty [x / (1 - d^{-1})] f(x) dx$$

and $\lceil x \rceil$ denotes the ceiling function (i.e., the smallest integer $\geq x$).

Proof

Under ACO, each walking step extends by distance $(1 - d^{-1})$. The number J of walking steps needed to close an ocean of size x is thus $\lceil x/(1 - d^{-1}) \rceil$. The result follows simply by taking the expectation of J over the distribution of ocean sizes.

Results

We now apply the results to study the problem of sequencing a genome by walking. Figure 5A shows the proportion R of redundant sequencing, as a function of the initial mean ocean length ω and the BAC library depth d . If the genome is seeded to leave oceans of average length $\omega = 1$ (corresponding to initial coverage $\pi = 50\%$), the proportion of redundant sequencing ranges from 32% to 29% as the BAC library depth ranges from $d = 15$ to $d = \infty$. If the genome is seeded more sparsely so that $\omega = 2$ and $\pi = 33\%$, the redundant sequencing R ranges from 23% to 18% over the same range of BAC library depth. If the oceans are still larger ($\omega = 3$ and $\pi = 25\%$), the redundant sequencing R ranges from 19% to 13%.

It is useful to compare the results with the situa-

tion of serially walking the entire genome with a minimal tiling path (which corresponds to ocean size $\omega \rightarrow \infty$). As noted above, the proportion of redundant sequencing for a minimal tiling path is $1/(d - 1)$. By subtracting this quantity from R , we find the additional redundant sequencing R^* caused by inefficient closure of oceans. Figure 5B shows the corresponding graph for R^* .

The graph shows that the redundant sequencing caused by inefficient closure of oceans depends primarily on the mean ocean size but much less on the library depth d . This makes intuitive sense, because the inefficiency arises primarily from closing small oceans with clones of unit length. The availability of more clones in a deep library thus makes only a modest improvement. The component R^* is decreased by only a few percentage points by going from a library with $d = 10$ to $d = \infty$.

Figure 6 shows the number of unidirectional walking steps required to cover 90% of the genome. The number of unidirectional steps is equal to slightly more than twice the mean ocean length. Because actual walking proceeds bidirectionally, it is thus useful to restate the observation in these terms: *The number of bidirectional walking steps to cover 90% of the genome is approximately equal to the mean ocean length.* (The precise correspondence depends on the library depth d .)

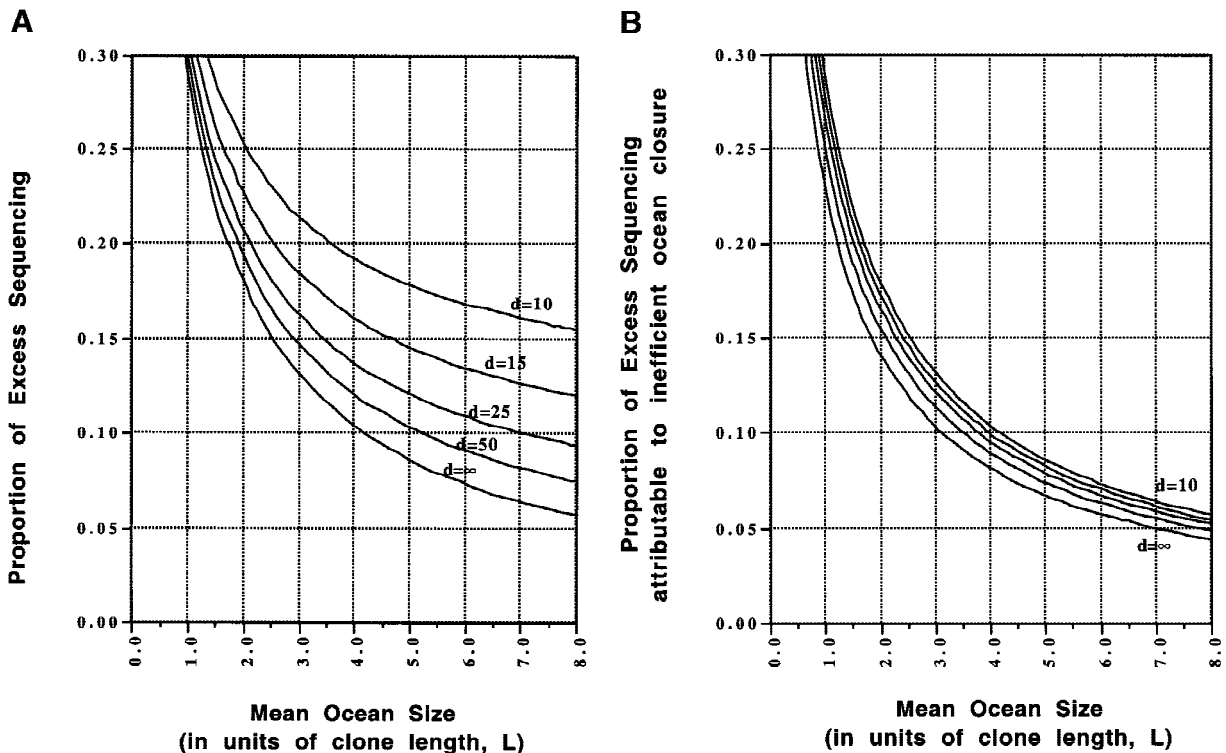


Figure 5 Proportion of excess sequencing. (A) Curves show the proportion of excess sequencing, $R(d, \omega)$ as a function of mean initial ocean size ω for various library depths $d = 10, 15, 25, 50$, and ∞ . (B) Curves show the proportion of excess sequencing attributable to inefficient ocean closure, $R^*(d, \omega)$. This is obtained by subtracting the proportion of excess sequencing for an optimal tiling path in a library of depth d , which is $1/(d - 1)$. That is, $R^*(d, \omega) = R(d, \omega) - 1/(d - 1)$.

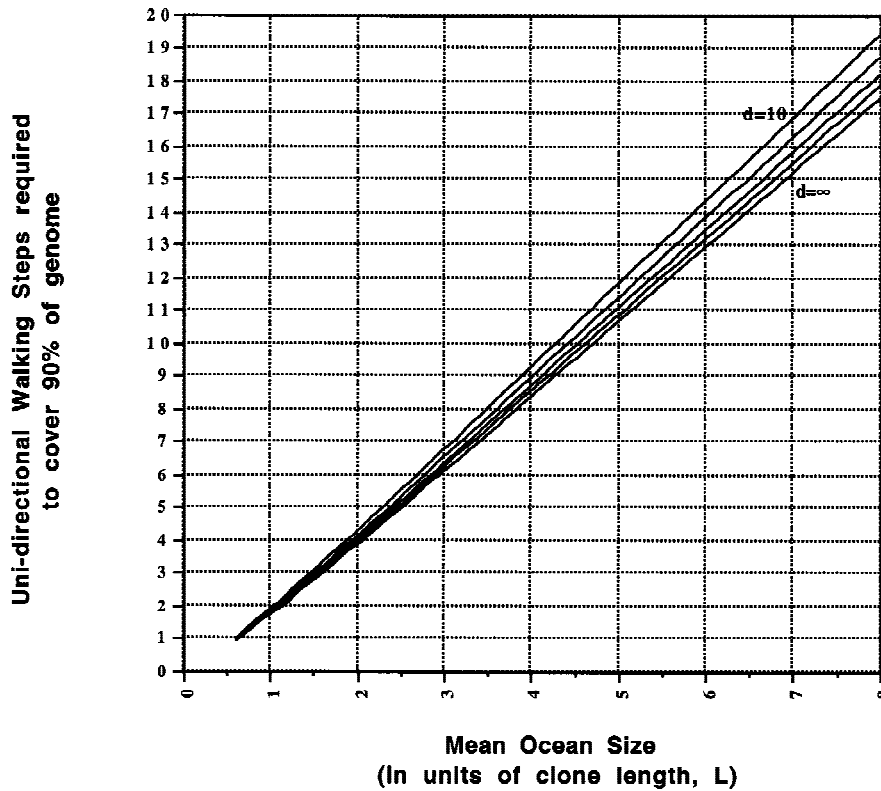


Figure 6 Walking steps. The number of unidirectional walking steps required to cover 90% of the genome, as a function of mean initial ocean size ω for various library depths $d = 10, 15, 25, 50,$ and ∞ . In actual practice, walking occurs in both directions. The number of bidirectional walking steps is thus half as many.

For example, the number of bidirectional walking steps is ~ 1.1 times the mean ocean length when $d = 15$.)

If the number of steps shown in Figure 6 is doubled, the proportion of genome sequenced rises to 98%–99%. [More precisely, the proportion remaining uncovered is $1\% \times (1 - \pi)^{-1}$.] Thus, the number of bidirectional walking steps to cover 98% of the genome is approximately equal to twice the mean ocean length.

From Figures 5 and 6, one can readily assess the increase in redundant sequencing entailed in covering the vast majority of the genome in a given number of steps.

Using Smaller Clones to Close Gaps

The major inefficiency in walking arises from instances in which a clone of length $L = 1$ must be sequenced to close a small ocean; most of the sequencing is redundant. This observation suggests a simple improvement: Use a second BAC library with smaller inserts to close small oceans.

Specifically, suppose that we have two BAC libraries in which the clones have been end-sequenced: our original library B with insert size $L = 1$ and depth d , as well as a second library B' with insert size $L' < 1$ and depth d' . We will continue to assume that walking from each seed clone proceeds unidirectionally to the

right (but see below concerning this point). At each walking step, we would first search our database to see if there is a clone from B' that spans the remaining ocean (based on its end sequences), and if none is found, we would select the minimally overlapping clone from B (which either closes the ocean or extends the walk). In this manner, we would aim to select the smallest clone capable of closing the ocean.

We can adapt the mathematical analysis above to calculate the proportion of redundant sequencing. [To simplify the statement of the result, we will consider only the case in which clones from B typically extend the walk farther than clones from B'; that is, $1 - (1/d) > L' - (1/d')$. This will be true unless L' is close to 1, in which case the second library adds little anyway.]

Proposition 3

Let B and B' denote BAC libraries as above. Suppose that we initially seed the genome with

nonoverlapping clones from B to coverage π and then walk as above, using clones from B' to close oceans whenever possible. Let $\omega = (1 - \pi) / \pi$. Under AEO and ACO, we have the following: (i) Let the random variable J denote the sum of the lengths of the clones used to close an ocean. As before, the expected proportion of redundant sequencing is

$$R(d, \omega) = \frac{E(J) + 1}{(\omega + 1)} - 1.$$

(ii) With $p_{d, \omega}$ defined as in proposition 1, we have

$$E(J) = \frac{1 - p_{d', (\omega/L')}(1 - L')}{p_{d, \omega}}$$

The formula reduces to that in proposition 1 if the smaller insert library B' has no clones ($d' = 0$), because $p_{0, \omega/L'} = 0$.

Proof

We proceed as in the proof of proposition 1. Part i again follows by a simple application of Wald's equation. For part ii, we calculate $E(J)$ by considering a single walking step and applying recursion. With probability $p_{d', (\omega/L')}$, the ocean can be closed with clone from library B', resulting in a total clone length of L' .

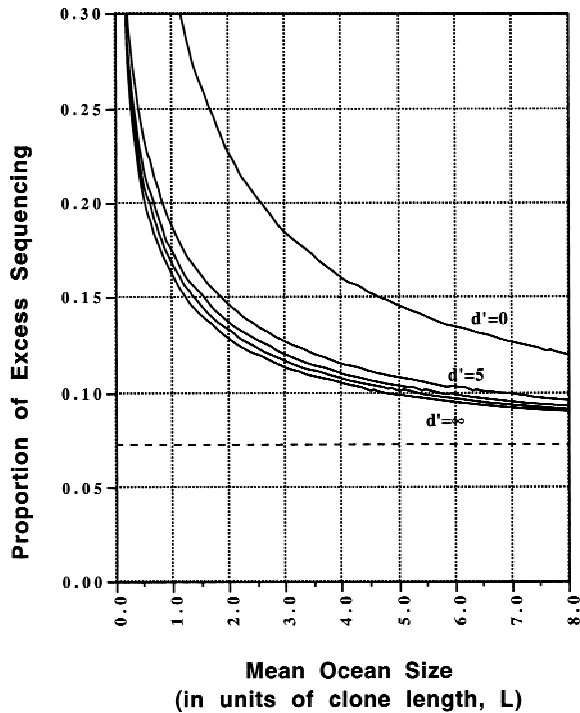


Figure 7 Proportion of excess sequencing, for walking with two BAC libraries. The first library B has clones of insert size $L = 1$ and fixed depth $d = 15$, whereas the second library B' has clones of insert size $L' = 1/2$ and various depths d' . Walking is performed as described in the text. The curves show the proportion of excess sequencing, as a function of mean initial ocean size ω and for various library depths ($d = 10, 15, 25, 50$, and ∞) for the second library B'.

With probability $p_{d,\omega} - p_{d',(\omega/L')}$, the ocean can be closed with a clone from library B but not B', resulting in a total clone length of 1. With probability $1 - p_{d,\omega}$, the walking step cannot close the ocean and remaining ocean having the same exponential distribution. It thus follows by recursion that $E(J) = [p_{d',(\omega/L')}] \cdot L' + [p_{d,\omega} - p_{d',(\omega/L')}] 1 + (1 - p_{d,\omega}) [1 + E(J)]$. The result follows by solving for $E(J)$.

How much efficiency is gained by using a second library B' with smaller insert clones? Figure 7 shows the proportion of redundant sequencing when an initial library B with depth $d = 15$ is supplemented with a second library B' with clones whose inserts are half as long ($L' = 1/2$) at various depths d' .

The savings are substantial, for example, decreasing the proportion of redundant sequencing from ~23%–14% for mean ocean length 2 and from ~18%–12% for mean ocean length 3. The redundant sequencing is considerably closer to the best possible level obtained with a minimal tiling path, ~7.1% for a library with $d = 15$ (indicated by the broken line in Fig. 7). Using a library with half-size clones roughly halves the redundant sequencing R' caused by inefficient closure of oceans.

Notably, the second library B' does not need to

have high depth to have a major impact. A library with depth $d = 5$ is only slightly less effective than a library with infinite depth. This makes intuitive sense, because even relatively low depth assures the existence of clones suitable for covering very small oceans.

The results above concern a second library B' with insert size $L' = 1/2$. This choice of insert size L' is nearly optimal for minimizing the amount of redundant sequencing R . Specifically, one can show by straightforward calculus that the value of L' that minimizes R only changes from 0.55 to 0.50 as the mean ocean length ω goes from 1 to ∞ .

The analysis above assumes that walking from each seed clone proceeds unidirectionally to the right. The fact that walking is actually bidirectional slightly complicates the situation: If the clones for a given round of walking were all sequenced simultaneously, we would sometimes cover oceans inefficiently by walking with large clones at both ends, when using one large and one small clone would suffice (Fig. 8). The additional excess sequencing that would result from such occurrences can be shown to be

$$\rho = (1 - L') \left[\frac{\sum_{i=0}^{\infty} p_{d',\omega/L'} (1 - p_{d,\omega})^{2i+1}}{(\omega + 1)} \right] \\ = (1 - L') \left[\frac{(1 - p_{d,\omega}) p_{d',\omega/L'}}{(2 - p_{d,\omega}) p_{d,\omega}} \right] / (\omega + 1)$$

[Briefly, the first term is the excess contribution to J from sequencing a large clone instead of a short clone; the second term is the proportion of oceans that should properly be closed with one large and one small clone (these are precisely the oceans that would be closed by unidirectional walking with an odd number of large clones followed by a small clone), and the denominator $(\omega + 1)$ occurs as in the calculation of R in 2(i) above.] For typical values of interest (e.g., $d \geq 15$, $d' \geq 5$, $L = 1/2$, and $\omega = 1-5$), the additional excess sequencing would be in the range of 3%.

It is possible to do much better than this by exploiting the fact that the BAC clones will not all be sequenced simultaneously. As shotgun sequence is obtained for each large clone B, one can automatically check the sequence to see whether any subsequent large clone B' can be replaced by an ocean-closing small clone B'' (as in Fig. 8). Provided that B' did not pass through shotgun sequencing simultaneously with B, one can thereby avoid the cost of sequencing a larger



Figure 8 Schematic illustration of the situation of an ocean being closed by two large clones (B and B'), when one large and one small clone (B and B'') would suffice.

clone when a smaller clone will suffice. (One may have incurred the cost of preparing a small-insert library from B', but this is small relative to the cost of sequencing.) The additional excess sequencing is therefore only $\alpha\rho$, where $\alpha < 1$ is the proportion of clones in a given walking step that are processed in parallel through shotgun sequencing. The proportion α depends on the workflow of the sequencing operation—specifically, on the amount of work in process (WIP) in the shotgun sequencing phase. Because the shotgun sequencing phase is relatively rapid (the elapsed time for picking, growing, preparing, and sequencing the small-insert clones is typically on the order of a week), the WIP tends to be relatively small. Even if the proportion of clones simultaneously passing through shotgun sequencing is as high as $\alpha = 10\%$, the additional excess sequencing owing to the occasional failure to use a small clone is only $\alpha\rho = 0.3\%$. In short, the cost is small and does not substantially impact the advantage of using a smaller library. The results derived under the assumption of unidirectional walking are thus not far off.

Optimizing BAC Library Depth

What is the optimal depth of a BAC library to be used for sequencing a genome by walking? One can decrease the cost of redundant sequencing by using deeper BAC libraries, but one incurs the cost of end-sequencing a larger number of BAC clones. Beyond some point, there are diminishing returns to increasing the depth of the BAC library.

The optimal library depth depends on the relative costs of sequencing entire BAC clones versus sequencing BAC ends. With current laboratory procedures and economics, this ratio is in the neighborhood of $\rho = 1000:1$. (Roughly 4000 shotgun sequencing reactions are needed to sequence a 200-kb BAC. Two sequencing reactions are required to sequence its ends, although these reactions are considerably more expensive because sequencing directly from a BAC template requires higher concentrations of reagents.) Given the cost ratio ρ , the optimal library depth is the value d that minimizes the sum of the costs of BAC-end sequencing (proportional to d) and redundant BAC sequence [proportional to $R(d, \omega)$].

Figure 9 shows the optimal library depth when using a single BAC library (curve denoted d^*). The optimal depth increases with the initial mean ocean length ω or equivalently decreases with the initial proportion π of the genome covered. The optimal depth d ranges from 22.5 to 30.8 as ω increases from 1 to 8, approaching an asymptotic limit of 32.8 as $\omega \rightarrow \infty$. (One can show that the optimal library density approaches $\sqrt{\rho + 1}$ as $\omega \rightarrow \infty$, by using straightforward calculus to minimize the asymptotic total cost.) The fact that the optimal d is smaller for small ω makes intuitive sense,

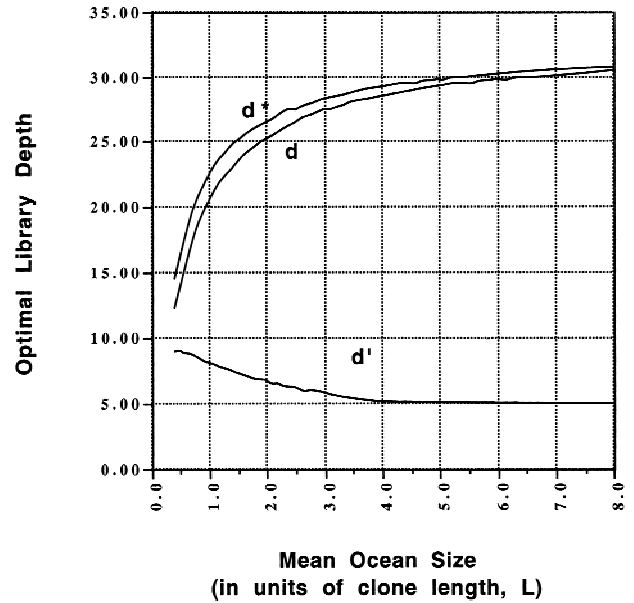


Figure 9 Optimal library depth to minimize total cost, assuming that the ratio of the cost of obtaining the complete sequence of a BAC to the cost of sequencing both ends of a BAC is $\rho = 1000:1$. The top curve (marked d^*) shows the optimal depth when using a single library. The two lower curves show the optimal depths d and d' when using two library B and B', as described in the text. The optimal depths depend on the mean initial ocean size ω . The depth d' of the small insert library was constrained to be at least 5, to ensure that the existence of a properly situated, small-insert clone at the end of the vast majority of islands, consistent with the assumption ACO, described in the text.

because a dense library offers more limited advantage when most oceans are fairly small.

Figure 9 also shows the optimal library depths d and d' when using two BAC libraries with respective insert sizes of 1 and $\frac{1}{2}$. The optimal value of d of the large insert library is only slightly lower than in the previous case, whereas the optimal value of d' is in the neighborhood of 6–7 for relevant values of ω .

The optima are relatively broad. Overall, it seems reasonable to use libraries B and B' with respective depths of $d = 20\text{--}25$ and $d' = 6\text{--}7$.

Seeding the Genome

We next turn to the issue of generating a collection of nonoverlapping seed clones. The analysis above assumes that seed clones are chosen such that initial ocean sizes are exponentially distributed (AEO). How close can one come to this in practice?

The most straightforward experimental approach is to select seeds sequentially from a list of random clones. Each clone is selected and sequenced, subject only to the condition that it does not overlap (as seen from its end sequences) with a previously selected seed clone. In this fashion, one can progressively seed the genome with random, nonoverlapping clones.

This stochastic process is known in the literature as the “parking process” because it is equivalent to cars of constant size (clones) sequentially choosing random parking spots along a long street (genome). Each car is permitted to park in its chosen parking spot provided that it is not blocked by a previously parked car; otherwise, it must drive away.

The resulting “parking distribution” is relevant to many physical processes and has been well studied (Krapivsky 1992). Suppose that one has processed a list of random clones comprising a sublibrary covering the genome to depth t . The expected proportion of genome covered with seed clones will be $\pi(t)$, and the distribution of ocean sizes will be $\omega(x,t)$, where

$$\pi(t) = \int_0^t F(\tau) d\tau$$

$$\Omega(x,t) = \begin{cases} [t^2 F(t) \exp(-(x-1)t] / \pi(t), & x > 1 \\ [2 \int_0^t \tau F(\tau) \exp(-x\tau) d\tau] / \pi(t), & x \leq 1 \end{cases} \text{ and}$$

$$F(t) = \exp \left[-2 \int_0^t \frac{1 - \exp(-\tau)}{\tau} d\tau \right]$$

The parking distribution of ocean sizes is very close to the exponential distribution for low coverage $\pi(t)$ and is still reasonably close for coverage $\pi(t) = 50\%$ (Fig. 10A,B).

The expected proportion of excess sequencing under the parking distribution can be calculated as in proposition 2. The difference Δ between the excess sequencing expected under the exponential distribution and the parking distributions is shown in Figure 11. The parking distribution entails slightly less excess sequencing than the exponential distribution (i.e., $\Delta > 0$), because it has fewer extremely small intervals (as see from Fig. 10A). The difference Δ is quite small over the range of interest, being $\sim 0.6\%$ for $\omega = 1$ and decreasing as ω increases.

It should be noted that the parking process cannot fully cover the genome: As the coverage increases, the remaining spaces become smaller, and the proportion of cars turned away increases. When the remaining spaces are all smaller than a car length, no more cars can park. This happens at the so-called “jamming limit,” which occurs at $\pi = 0.747597$. . . sequential selection of strictly nonoverlapping clones thus cannot provide seed coverage beyond this point.

Fortunately, we are not interested in seeding to $>50\%$ because the cost of excess sequencing begins to skyrocket for $\pi > 50\%$. Seeding the genome to 50% requires starting with a random list of clones covering the genome to depth 1.15 [because $\pi(1.15) \sim 0.5$].

We briefly mention some other approaches to seeding the genome but do not analyze them in detail.

Simultaneous Seeding

One could start with a collection of random clones,

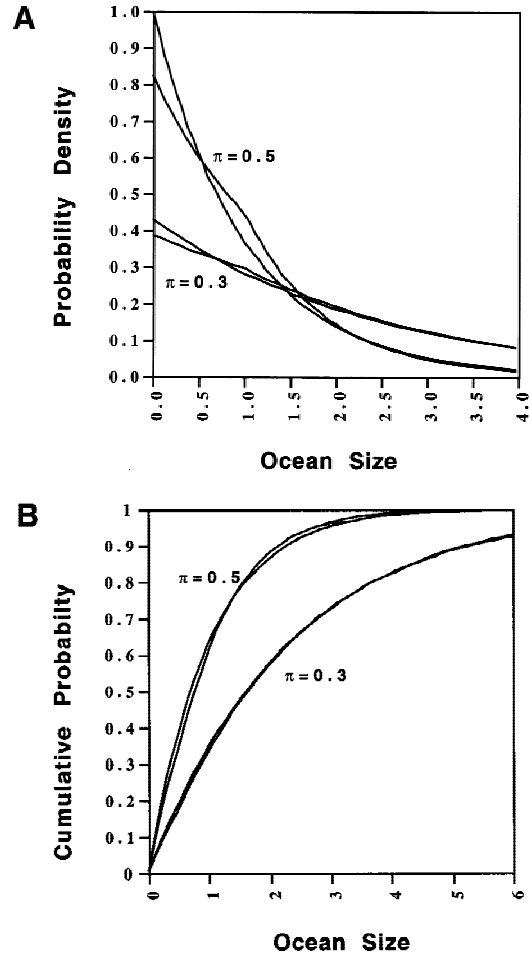


Figure 10 Comparison of parking and exponential distributions. (A) Probability density and (B) cumulative probability distribution, for $\pi = 0.3$ and $\pi = 0.5$. The kinked curves are the parking distribution; the smooth curves are the exponential distribution with the same mean.

prepare small-insert shotgun library from each, and perform a small amount of sequencing from each shotgun library (e.g., covering the clone to an average depth of 0.5-fold). Such “sample sequencing” should be sufficient to detect any significant overlap between clones. One could then select a maximal set of nonoverlapping clones. If the collection of random clones covers the genome to depth t , a maximal nonoverlapping set will cover a proportion $\pi \sim t / (t + 1)$ of the genome, and the oceans will be very nearly exponentially distributed with mean size $1 / t$. The approach allows the genome to be seeded to higher initial coverage, but the sample sequencing is more expensive than end sequencing and it must all be performed in advance. Still, there may be situations in which such an approach may be desirable.

Contig-Based Seeding

One could initially fingerprint the entire BAC library to

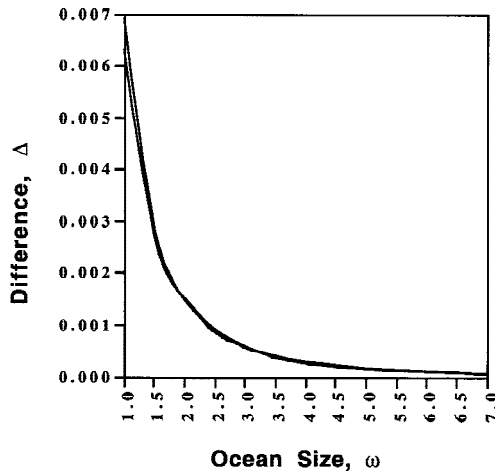


Figure 11 Difference between exponential and parking distributions. Graph shows difference Δ between R_{exp} , the proportion of excess sequencing under exponential distribution, and $R_{parking}$, the proportion of excess sequencing under parking distribution. That is, $\Delta = R_{exp} - R_{parking}$. Two curves are shown, corresponding to $d = 15$ and $d = 50$. The difference Δ is seen to be largely insensitive to d .

construct nonoverlapping contigs, which could be used as the seeds for subsequent walking. Assuming that the contigs were separated by roughly exponentially distributed oceans, the situation could be analyzed in the same manner as above.

Simulations

The simple formulas above are premised on two simplifying assumptions (AEO and ACO). We tested whether the formulas provided reasonable approximations by performing extensive simulations.

The simulations were performed as follows: Libraries were generated by randomly selecting starting points for BAC clones ($L = 2 \times 10^5$ bp, $L' = 1 \times 10^5$ bp) from a mammalian genome (3×10^9 bp) using a uniform distribution to achieve the desired depths d and d' . An initial selection of seed clones was generated as in the parking problem, by considering the clones in a random order and accepting successive nonoverlapping clones until the desired proportion π of the genome was reached. [We confirmed that the simulation-yielded ocean sizes closely fit the parking distribution (not shown).] Walking was then performed by selecting minimally overlapping BACs as described above. For each choice of parameters, a total of 40 simulations were performed, and the results were averaged. Simulations were performed for parameters throughout the range of interest, at the parameters $d \in \{10, 15, 25, 50, \infty\}$ and $\omega \in \{1.0, 1.5, 2.0, 2.5, \dots, 7.0\}$. We calculated $R_{sim}(d, \omega)$, the average proportion of redundant sequencing over the simulations.

We also performed simulations mimicking the ACO assumption, in which the seed clones were se-

lected as above, but walking steps then occurred under the assumption that the minimally overlapping clone overlapped by exactly $1/d$. We calculated $R_{sim,ACO}(d, \omega)$, the average proportion of redundant sequencing over these simulations.

We compared the simulation results to $R_{parking}(d, \omega)$, the excess sequence under the parking distribution predicted by the formula in proposition 2. The difference $\Delta(d, \omega) = R_{parking}(d, \omega) - R_{sim}(d, \omega)$ is shown in Figure 12A. The difference is negligible (on the order of the s.e.m. from the simulations), except for small d and small ω . The discrepancy for small d and ω is readily seen to be due to ACO, by examining the difference

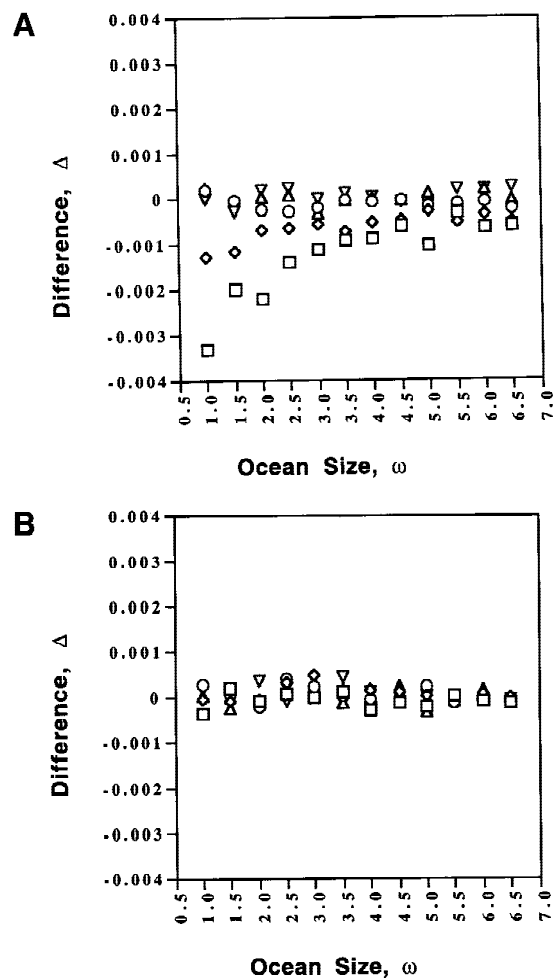


Figure 12 Difference between formulas and simulations. Plots show the differences Δ between the proportion of excess sequencing predicted under the parking distribution ($R_{parking}$) and the proportion of excess sequencing observed in (A) simulations in which the seed clones and walking clones are chosen from a randomly distributed library (R_{sim}) and (B) simulations in which seed clones are chosen from a randomly distributed library, but walking steps are then made under the ACO assumption ($R_{sim,ACO}$). Results are shown for various values of d and ω . (\square) $d = 10$; (\diamond) $d = 15$; (\circ) $d = 25$; (\triangle) $d = 50$; (gray ∇) $d = \infty$.

$\Delta'(d, \omega) = R_{\text{parking}}(d, \omega) - R_{\text{sim,ACO}}(d, \omega)$ shown in Figure 12B. This difference is negligible over the entire range.

We also studied the proportion of genome covered after k steps ($k = 1, 2, \dots, 10$), comparing the prediction under the exponential distribution [proposition 1(iii)] to the results seen in the simulations above. In the range of interest, the predicted proportion of genome covered at each stage was within 0.5% of the results from the simulations (not shown). Finally, we simulated the situation of two libraries, in which d and ω were as above and the smaller library has depth $d' = 5$. Again, the difference Δ was extremely small (not shown).

In summary, the simple formulas derived under the assumption of AEO and ACO are sufficiently close for practical purposes. The differences are quite small and have no impact on the key conclusions.

Conclusion

Sequencing a genome by walking is an inherently serial process. One seeds the genome to an initial coverage π and then engages in sequential rounds of walking.

Completing the task in a reasonable amount of time requires a high degree of parallelism, that is, a high density of seed clones from which walks proceed outward in both directions. The number of such walking steps to cover 90% of the genome is roughly equal to the initial mean ocean size [$\omega = (1 - \pi) / \pi$], whereas the number of steps to cover 98% is roughly twice as large.

There is a clear trade-off between the number of walking steps and the cost of redundant sequencing. As the mean ocean size ω grows from 1 to 2 to 3, the proportion of redundant sequencing decreases from 32% to 23% to 19%. (These values correspond to a library with depth $d = 15$; they are about two percentage points lower for libraries with $d = 25$.)

One solution is to decrease the cycle time required for each walking step, making it feasible to take more serial walking steps within the desired time frame. Because the time required to prepare and validate high-quality shotgun libraries represents a significant component of the cycle time, one might develop efficient and inexpensive methods to prepare shotgun libraries from all clones in the BAC library in advance of sequencing. (The storage requirements are modest, because each shotgun library is stored as a ligation mixture in a single test tube until required.) One could then rapidly initiate each consecutive walking step. Implementing such an approach would require significant streamlining or automation of existing procedures for preparation of shotgun libraries but may be feasible.

A complementary solution is to use a second BAC library with smaller clones, with roughly half the insert

size. This approach dramatically improves efficiency. For the cases involving a 15-fold library above, the proportion of redundant sequencing is 18%, 14%, and 12%. This is much closer to the best possible result obtainable with a 15-fold library of $\sim 7.1\%$ ($= 1/14$). Importantly, the vast majority of the efficiency is obtained using a second BAC library having relatively modest depth, for example, $d' = 5$.

The strategy could, of course, be generalized. One could use multiple BAC libraries each with distinct insert sizes or a single BAC library with extremely variable insert sizes. Mathematical analyses of such strategies would be of interest, although it is notable that the simple two-tiered strategy above already eliminates the majority of the excess redundant sequencing due to inefficient closure of oceans.

We should note that our analysis ignores certain biological and experimental issues:

Variable Insert Size

We have assumed that the BAC libraries have inserts of constant size. A BAC library with variable insert size may be more efficient than one with constant size, because one has the potential to optimize the size of the BACs used to close oceans. Our strategy of using of two BAC libraries with constant inserts of size L and L' is formally equivalent to using a single BAC library with variable insert size equal to either L or L' . In principle, BAC libraries with variable insert size can be extremely efficient: A BAC library with highly variable insert size and infinite depth would allow one to close gaps with essentially no wasted sequencing.

In practice, however, current BAC libraries have a fairly tight size distribution, and thus the effect is rather modest. The human RPCI-11 library, for example, has insert size that is roughly normally distributed with mean 160 kb and coefficient of variation [(CV) defined as ratio of S.D.M.] of somewhat $< 10\%$.

We performed simulations to compare the proportion of redundant sequencing in two situations cases: (1) a single library having mean insert size 1 with the insert size either being constant or normally distributed with $CV = 10\%$, and (2) two libraries having mean insert sizes 1 and 0.5 with the insert sizes either being constant or normally distributed with $CV = 10\%$. As expected, the variable libraries were slightly more efficient than the constant library. For $\omega = 1$, the absolute difference was $\sim 4\%$ in the case of a single library ($d = 15$) and 2% in the case of two libraries ($d = 15$, $d' = 5$). For larger ocean sizes ω , the absolute differences are even smaller. In short, the effects are thus fairly small, and the key conclusions concerning the value of a library containing smaller clones remains valid.

An interesting open question is to find the optimal distribution of insert sizes for a BAC library of a given depth.

Cloning Bias

We have assumed that there is no cloning bias in the genomic library. Cloning bias is known to occur in many cloning systems, but the nature and distribution of cloning bias is poorly understood and therefore difficult to model. Severe cloning bias against some regions would clearly lead to larger initial oceans, as well as lower effective library depth, in these regions. One could conceivably model the genome as composed of large blocks with different cloning bias.

Missing End Sequences

We have assumed that each BAC has been sequenced at both ends. Current BAC-end sequencing projects sometimes fail to generate one of the end sequences owing to technical failures. Such "single-ended" BACs are less inefficient because one might select two such clones to walk from each side of an ocean and be unaware (because of the lack of the opposite sequence) that each clone suffices to close the ocean on its own. The impact of the problem can be assessed through either mathematical analysis or simulation. The problem itself can be overcome simply by sequencing additional BAC clones to achieve the desired depth d in "double-ended" clones.

Repeat Sequences at BAC Ends

We have assumed that each BAC has unique sequence at both ends to permit unambiguous recognition of overlap. However, a clone end may consist entirely of a repeat sequence, which cannot be used for walking. If the repeats are relatively small and fairly uniformly distributed across the genome, the problem can be overcome simply by end-sequencing a larger number of BACs to obtain enough clones with unique sequence at both ends (as in the case of single-ended BACs due to technical failure). If repeats are unevenly distributed across the genome, the problem will lead to underrepresentation of clones in the repeat-rich regions (as in the case of cloning bias). If very long repeats occur, they will prevent the initiation of walks from within the repeat and may necessitate larger overlap. In general, the most practical solution is to increase the depth of the library used for end-sequencing (see also Siegel et al. 1998).

This paper has focused on developing simple models to provide insight into the problem of sequencing a genome by walking. In actual application, one will confront complications such as those above. The best solution is to perform specialized simulations. The results in this paper, however, help define the key issues and trade offs and should be helpful in conceptualizing how to design a clone-based genomic sequencing project.

ACKNOWLEDGMENTS

We thank Bruce Birren and Ken Dewar for helpful comments. S.B. was supported by a Merck/MIT fellowship. This work was supported in part by a grant from the National Institutes of Health (to E.S.L.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**: 212–2018.
- Coulson, A., J. Sulston, S. Brenner, and J. Karn. 1986. Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **83**: 7821–7825.
- Dujon, B. 1996. The yeast genome project: What did we learn? *Trends Genet.* **12**: 263–270.
- Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–511.
- Goebel, S.J., G.P. Johnson, M.E. Perkus, S.W. Davis, J.P. Winslow, and E. Paoletti. 1990. The complete DNA sequence of Vaccinia virus. *Virology* **179**: 247–266.
- Green, P. 1997. Against a whole-genome shotgun. *Genome Res.* **7**: 410–417.
- Krapivsky, P.L. 1992. Kinetics of random sequential parking on a line. *J. Stat. Physics* **69**: 135–150.
- Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**: 231–239.
- Oda, K., Y. Katsuyuki, E. Ohta, Y. Nakamura, M. Takemura, N. Nozato, K. Akashi, T. Kanegae, Y. Ogura, T. Kohchi, and K. Ohya. 1992. Gene organization deduced from the complete sequence of Liverwort *Marchantia polymorpha* mitochondrial DNA. *J. Mol. Biol.* **223**: 1–7.
- Ohyama, K., H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, K. Umesono, Y. Shiki, M. Takeuchi, Z. Chang et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* **322**: 572–574.
- Oliver, S.G. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**: 38–46.
- Olson, M.V., J.E. Dutchik, M.Y. Graham, G.M. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Scheinman, and T. Frank. 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci.* **83**: 7826–7830.
- Roach, J. 1995. Random subcloning. *Genome Res.* **5**: 464–473.
- Ross, S.M. 1970. *Applied probability models with optimization applications*. Holden-Day, San Francisco, CA.
- Sanger, F., A.R. Coulson, B.G. Barrell, A.J.H. Smith, and B.A. Roe. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* **143**: 161–178.
- Sanger, F., A.R. Coulson, G.F. Hong, D.F. Hill, and G.B. Petersen. 1982. Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**: 729–773.
- Siegel, A.F., B. Trask, J. Roach, G.G. Mahairas, L. Hood, and G. van der Engh. 1998. Analysis of sequence-tagged-connector strategies for DNA sequencing. *Genome Res.* **9**: 297–307.
- Venter, J.C., H.O. Smith, and L. Hood. 1996. A new strategy for genome sequencing. *Nature* **381**: 364–366.
- Weber, J.L. and E.W. Myers. 1997. Human whole-genome shotgun sequencing. *Genome Res.* **7**: 401–409.

Received July 22, 1999; accepted in revised form October 29, 1999.