Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 8034–8038, September 1991
Genetics

Optimized strategies for sequence-tagged-site selection in genome mapping

MICHAEL J. PALAZZOLO*†‡, STANLEY A. SAWYER§¶, CHRISTOPHER H. MARTIN*‡, DAVID A. SMOLLER*, AND DANIEL L. HARTL*

*Department of Genetics, Box 8232, Washington University School of Medicine, St. Louis, MO 63110; and §Department of Mathematics, Washington University, St. Louis, MO 63110

Communicated by Gerald M. Rubin, May 31, 1991

ABSTRACT **The physical mapping of complex genomes is based on the construction of a genomic library and the determination of the overlaps between the inserts of the mapping clones in order to generate an ordered, cloned representation of nearly all the sequences present in the target genome. Evaluation of the relative efficiency of experimental procedures used to accomplish this goal must minimally include a comparison of the fraction of the genome covered by the ordered arrays (or "contigs"), the average size of the contigs, and the cost, in terms of time and resources, required to generate the map. Sequence-tagged-site (STS) content mapping is one strategy that has been proposed and is being utilized for this type of experiment. This paper describes three STS selection schemes and presents computer simulations of contig-building experiments based on these procedures. The results of these simulations suggest that a nonrandom STS strategy that uses paired probes requires one-third to one-fourth as many STS assays as are required in random and nonpaired approaches, and also results in a map that has both greater genome coverage and a larger average contig size. This strategy promises to reduce the time and cost required to build a high-quality physical map.**

One of the great advances of molecular genetics has been the development of methodologies in which large genomes are fragmented into easily manipulable pieces, which can be inserted into either bacterial, phage, or yeast vectors and then introduced into host cells that can be isolated and maintained as distinct clones (1–4). By analyzing a large number of genomic clones it is possible to begin to determine the physical relationships between the genomic inserts and thus generate a physical map (5–7). Such a map can facilitate a variety of experiments by providing a high-resolution framework for analysis and storage of genetic and molecular information. In addition, the clones upon which the map is based can serve as a source of DNA for the analysis of particular genetic regions. The application of this approach to the human genome promises to revolutionize the ability to isolate genes that are identified by mutations that result in human diseases. Furthermore, physical map construction should facilitate the production of substrates for large-scale genomic sequencing.

A few key concepts need to be considered in order to compare mapping strategies (8). The first concept is that of a "contig." This term refers to a set of clones that have been shown, by any of a variety of experimental techniques, to overlap. One criterion for evaluating mapping progress is the average size of the contigs, which is the average length of the genomic regions covered by the inserts in the contigs. Second, progress towards completion of the map can be measured by the fraction of the genome that is covered by the contigs.

The average size of the contigs and the fractional coverage of the genome in mapping experiments are limited by three factors. The first is that some regions of the genome are not represented in the library. One cause of this is the relative difficulty in cloning certain genomic regions. Sequences that are difficult to clone lead to more gaps, smaller contigs, and poorer coverage. Another reason for unrepresented regions is that some sequences are not cloned by chance. In a three-hit library almost 5% of the genomic sequences would be uncloned, in a four-hit about 2%, and in a five-hit less than 1%. The second limitation is the size of the mapping clones. The larger the cloned insert, the fewer the number of clones that have to be analyzed in order to achieve comparable coverage and continuity. This fact explains the increasing popularity of cloning vectors that can accommodate large inserts, such as yeast artificial chromosomes (YACs) (2) and P1 pacmids (3, 4). The third limiting factor comes from the economic and practical limitations of the strategy used to identify physical overlaps between mapping clones.

Several approaches have been utilized to identify overlaps, which include restriction mapping (5, 7), fingerprinting (6), oligonucleotide probe hybridization (9), and end-directed chromosome walking (10). These procedures have been used primarily to assemble contigs from libraries that were constructed in phage λ or in cosmids; they tend to be much more difficult to apply to vectors that contain larger inserts. The development of the polymerase chain reaction (PCR), in turn, led to the proposal of a strategy known as sequence-tagged site (STS) content mapping (11), which has been successfully applied to contig building in YAC libraries (12). STS content mapping is based on the premise that two clones that share the same single-copy genomic sequence must overlap. Therefore, any unique sequence present in the genome can be used as an STS probe to identify a set of overlapping clones that represents the same region of the genome. STS mapping procedures are particularly attractive because, in addition to providing a mechanism for contig building, they allow the introduction of mechanisms for community accessibility, flexibility, and biological content into the map as it is being constructed (11).

One of the major difficulties of using STS content mapping in contig-building experiments is the requirement for assaying tens or hundreds of thousands of mapping clones with thousands of STS markers. The labor-intensive nature of this highly repetitive work results in significant costs in terms of time and resources. Many current STS content mapping strategies are based on the relatively random selection of STS probes. With random STS selection, a significant effort is spent mapping STS markers that provide redundant infor-

Abbreviations: STS, sequence-tagged site; YAC, yeast artificial chromosome.

†To whom reprint requests should be addressed.

‡Present address: Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720.

¶To whom requests for the computer programs should be addressed.

mation. This limitation is functionally analogous to those seen in random, shotgun DNA sequence analysis (13) or in collecting large numbers of non-cross-hybridizing cDNA clones from normalized cDNA libraries. Methods have been developed for sorting subtracted cDNA libraries in which cDNA clones that are isolated are then used as probes to identify those cDNA clones that are not yet collected (14). An extension of this concept should allow an ordered, nonrandom procedure for the identification of a much more efficient set of STS markers for contig-building experiments.

A second limitation in current STS selection schemes is the relatively small size of a contig identified by a single STS. This is due to the fact that a single STS acts as a point and can identify only the clones that contain it. It is attractive to speculate that linked STSs might generate larger contigs. A pair of STSs, selected from the ends of a single mapping clone, should act as an STS with the width of a typical cloned insert and generate a larger contig. The computer simulations described below suggest that significant economy can be derived when the STS markers used in physical mapping experiments are selected in a paired, nonrandom fashion.

RESULTS AND DISCUSSION

A Single-End Clone-Limited Strategy for the Selection of STSs. One of the key elements in our proposed approach is the use of data accumulated by prior STS mapping to generate new STSs that are much less likely to provide redundant information than are STSs chosen by a random selection scheme. One such strategy is the following. First, a multihit library is generated that, on the average, covers the genome five times. This library is then arranged in a gridded array so that each mapping clone can be maintained as a distinct entity. A single clone from the library is selected and a single-copy genomic sequence is obtained from one end of the cloned genomic insert. This initial STS marker is used as a probe to identify other mapping clones that contain this sequence; this yields the first contig. A second mapping clone is chosen from among those clones that have not yet been identified as members of a contig by STS mapping. An STS probe is then derived from one end of this clone and the library is rescreened to build the second contig. This procedure is then repeated in an iterative fashion until every mapping clone has either been used to provide an STS marker or shown to contain a mapped STS. In contrast to random strategies, there exist a relatively small number of potential STSs in our approach, since the number of STS markers is limited by the number of mapping clones that constitute the library. With a single-end clone-limited selection scheme, there can never be more STSs than there are clones. However, in screening a multihit library, most mapping clones will be assigned to contigs long before they would be chosen to provide STSs. After all clones have been assigned, the research group is free to switch to a new strategy to complete the map.

A Double-End Clone-Limited Strategy. In the Introduction, we speculated that assays using paired STSs might yield larger contigs than can be produced using single markers. With a single-end clone-limited strategy (described in the preceding paragraph) with a five-hit library, a typical early contig will contain 6 clones, the selected clone plus 5 others. If the STS is not derived from one of the mapping clones, then an early contig is likely to consist of 5 mapping clones. An intuitive reason for the difference is that selecting a random clone biases the choice of genomic position to regions where there is a higher density of clones. A calculation based on the Poisson distribution shows that the average size of one of these early contigs is about 1.6 times the size of the cloned inserts for the random STS strategy, and about 1.8 times for the single-end strategy. If the mapping clones are 100 kilobase pairs (kb), this is about 160 kb for the random strategy and 180 kb for the single-end approach. In contrast, a double-end clone-limited STS selection algorithm uses two linked STSs. The strategy to select these markers is similar to the single-end clone-limited approach except that both ends of each selected clone are used as linked STS probes. A contig generated by this scheme would, on average, contain 11 clones, the mapping clone and two sets of 5 different clones that separately overlap with the STSs from each end. The average size of this type of early contig will typically be about 2.6 times the size of the clone inserts, or about 260 kb in the example where the mapping clones contain 100-kb inserts.

Computer Simulations. Computer simulations were done to study the relative effect of three strategies for selecting STS markers: (*i*) random selection of probes, (*ii*) single-ended clone-limited methods, and (*iii*) the double-ended clone-limited selection schemes described above. The initial simulations were based on considerations of a physical map for a 100-megabase-pair (Mb) genome from a library of 100-kb mapping clones that provide a 5-fold average coverage of the genome. Both the description of and the actual C program that implements these algorithms and graphs the results are available upon request from one of the authors (S.A.S.).

Computer-Simulated Results That Assess Genome Coverage. The results of the first set of simulations indicate that, compared with random STS methods, the clone-limited strategies provide nearly complete genome coverage with far fewer STS mapping assays. The results of these simulations are presented graphically in Fig. 1*A*. The *x* axis corresponds to the number of STS assays performed, the *y* axis corresponds to the percentage of the genome covered by contigs, and the slope of the curve represents the rate of progress toward completion of the project. No more coverage can be obtained from a five-hit library after each of the mapping clones is assigned to a contig. This is the point at which 99% of the genome is likely to be covered, and is represented in the graph as the point at which each curve stops.

These data suggest that both clone-limited strategies are 3- to 4-fold more efficient than the random strategy in providing nearly complete coverage. Specifically, fewer than 1500 STS assays are likely to be required to position 5000 mapping clones into contigs in both clone-limited strategies, whereas about 5000 assays are necessary in the random approach (Fig. 1*A*). The relative rates of progress in the early stages of the project are approximately the same in all three selection schemes. The clone-limited procedures become relatively more efficient after about 60–80% of the genome has been mapped. In contrast, the random selection strategies become increasingly more inefficient as the map nears completion, with completion being defined in this instance as the assignment of all the mapping clones to contigs. Calculations based on the Poisson distribution suggest that, in a random selection strategy, about two-thirds of the effort will be required to finish the final 20% of the clone assignments.

Computer Simulations To Estimate Contig Number, Average Contig Size, and Contig Size Distribution. Further simulations were performed to assess the number and size of the contigs; they strongly indicate that the double-end (paired STS) clone-limited strategy is far superior in ordering the genome into larger contigs than either the random or the single-end (nonpaired) approach. Fig. 1*B* relates the number of STSs used to the number of contigs that are generated. All three curves rise steeply in the initial part of the mapping project because most of the STSs each identify a cluster of clones, typically about 5–6 in the random and single-end schemes and 11 in the double-end approach. The curves begin to flatten and then descend at the point at which the STS mapping strategy begins associating these small clusters into larger contigs. The curves depicting the progress of the

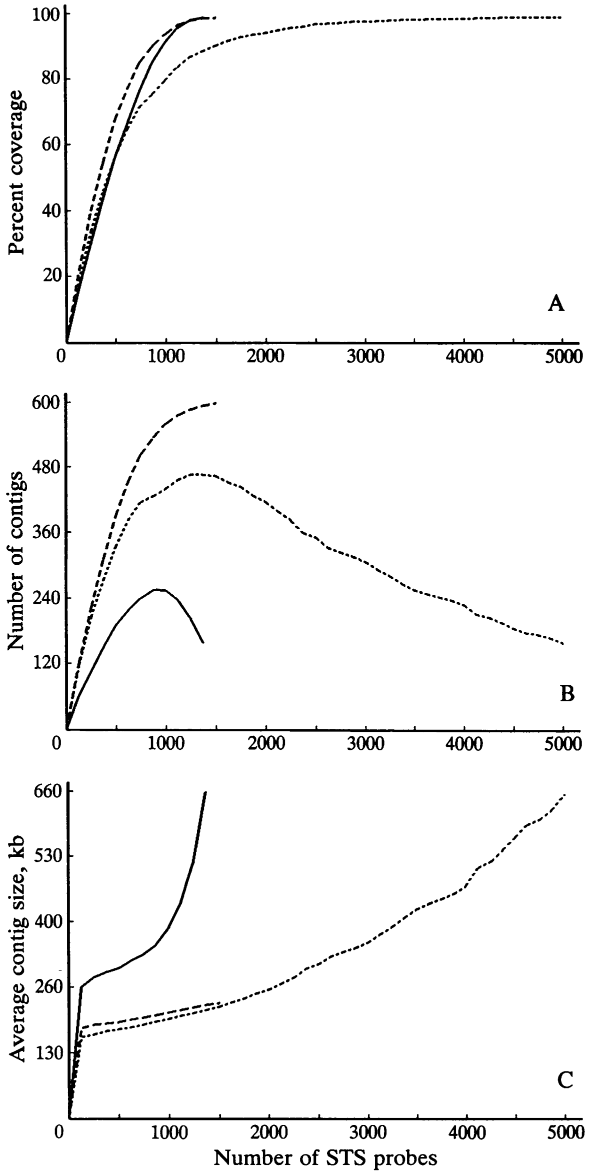

FIG. 1. The relationships between the number of STS assays performed and the percent genome coverage (*A*), the number of contigs (*B*), and the average contig size (*C*). These graphs describe the results of simulations that were performed to compare the efficiency of three different STS selection strategies: (*i*) random (short-dashed curve in each part of the figure), (*ii*) single-end clone-limited (long-dashed curve), and (*iii*) double-end clone-limited (solid curve). The genome size is set at 100 Mb and the mapping clone insert size at 100 kb, and the library contains 5000 clones.

clone-limited approaches terminate when all clones have been assigned to contigs. Fig. 1*C* describes the relationship of the number of STS assays to the average size of the contigs.

Relative Efficiencies for Building Large Contigs. Perhaps the most significant finding of the simulation is that the double-end clone-limited strategy is clearly the most successful at efficiently building large contigs. At the point where the genome has been nearly completely covered, the average contig size for the double-end clone-directed strategy was 677 ± 41 kb (mean ± SD over 10 runs). In contrast, a strategy that uses an equivalent number of randomly selected STSs leads to coverage of only about 85% of the genome (Fig. 1*A*) and the average contig size is less than 200 kb. Even after 3–4 times as many STS assays, the average contig size generated by random STS selection is still smaller than that provided by the double-end method. The single-end clone-limited scheme is about as effective as the linked approach in providing nearly complete coverage (Fig. 1*A*), but the average contig size is no better than that provided by the random strategy (Fig. 1*C*). To present a picture of the size distribution of identified contigs at the point in which 1500 STS assays have been performed, the data were recalculated and presented as a series of bar graphs (Fig. 2).

Another benefit of the clone-limited strategy is that the results of the experiments dictate the point at which the initial strategy is exhausted. This occurs when all the mapping clones have been assigned to contigs and nearly complete coverage of the genome is obtained and can be seen in Fig. 1 *B* and *C* as the point at which the curves abruptly terminate. This notion has practical experimental significance. Most previous mapping experiments were based on fingerprinting or restriction mapping. The initial clones to be mapped were randomly chosen. In each of these efforts there was a point at which a choice was made to stop the initial strategy and move to a more directed approach, in an effort to build larger

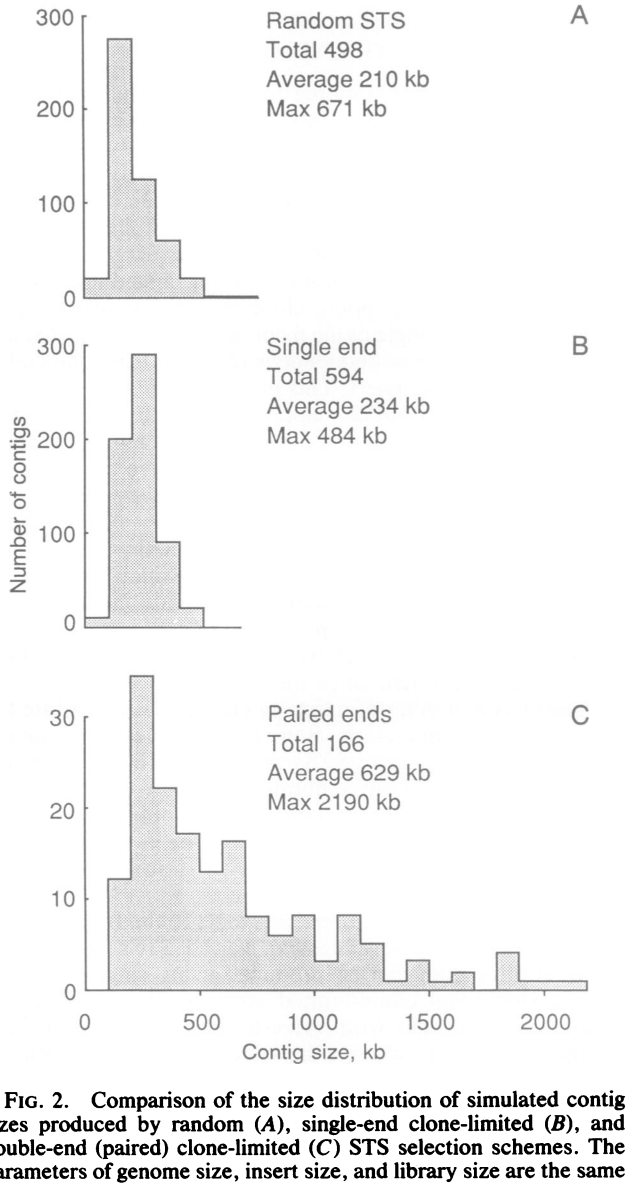

FIG. 2. Comparison of the size distribution of simulated contig sizes produced by random (*A*), single-end clone-limited (*B*), and double-end (paired) clone-limited (*C*) STS selection schemes. The parameters of genome size, insert size, and library size are the same as in the simulations described by Fig. 1. Each size distribution was generated by analyzing the data when 1500 STS assays had been performed, which corresponds to the point in the simulation when most (or all) of the clones had been assigned to contigs in the clone-limited strategies.

contigs. In the clone-limited scheme this occurs quite early and there is an explicit experimental result that identifies that this point has been reached. In our simulations, this point occurred when 1350–1400 STS assays had been performed (approximate range over 10 runs) for the double-end strategy and 1500–1550 STS assays for the single-end scheme.

Computer Simulations Based on Cloning Vectors That Can Accommodate Larger Cloned Genomic Inserts. A second set of simulations suggests that increased efficiency can be obtained in these types of mapping experiments when larger cloned inserts are used. In this set of modeling experiments two different analyses were performed. In one set the sizes of the mapping inserts were set at 200 kb and in the other the cloned inserts were 500 kb. In these models it is possible to compare the results obtained when the clones were selected from libraries comparable to ones that could be constructed in a YAC cloning vector. Like the first set of experiments, the data from these simulations were analyzed to determine the fractional coverage, the number of contigs, and the average size of each of the contigs in relation to the number of STS assays performed. The results (data not shown) suggest that the relative efficiency of the double-end strategy remains the same as compared to the other two approaches, in terms of number of STSs required to provide complete coverage and the average size of the contigs generated at this point in the project. Furthermore, in these experiments, the number of STS assays, the number of clones to provide five-hit coverage, and the number of contigs tend to decrease linearly with the size of the cloned insert, while the average contig size tends to increase linearly with the size of the mapping insert. Specifically, the results of five runs of the simulation using a double-end clone-limited strategy suggest that for a library consisting of clones with 100-kb inserts, a 5000-member library had to be screened with, on average, 1357 ± 14 probes to generate a physical map with about 157 ± 8 contigs with an average contig size of 670 ± 34 kb. For 200-kb cloned inserts, a 2500-member library had to be screened with 692 ± 13 double-ended, nonrandomly selected probes in order to generate a map containing 68 ± 8 contigs with an average contig size of 1.54 ± 0.16 Mb. For 500-kb inserts, a 1000-member library was probed with 270 ± 4 STSs to provide nearly complete genome coverage and to generate 35 ± 3 contigs of average size 3.07 ± 0.22 Mb. In other words, the average contig size (at the point at which all clones have been assigned to a contig) with the double-end clone-limited STS selection strategy tends to be 6–7 times larger than the size of the cloned insert.

A Second-Round STS Selection Strategy To Complete the Physical Map. One of the key concepts in our proposed mapping approach is that the information obtained in the process of constructing the map can also be used to economically guide its progress. It should be possible to extend this notion in a second round of nonrandom STS selection to efficiently complete the physical map. The first round of clone-limited STS selection schemes is exhausted when all of the mapping clones have been assigned to contigs. At this point 99% of the genome is likely to be covered by contigs, based on the few limited assumptions described above. However, it is important to realize that the number of contigs has not, at this point, been reduced to a minimum. In other words, while no more coverage can be obtained from this library, there is still information present in the mapping library that can be used to build significantly larger contigs. Continued STS content mapping is likely to identify overlaps between contigs that have not yet been identified. The most efficient way to approach this goal and complete the project is to obtain STSs from the clones positioned at the ends of the contigs.

The results from the first set of computer simulations can be used as an example of our proposed second round of nonrandom STS screening to complete the physical map. As mentioned in the preceding paragraph, a shift in STS selection strategy during the mapping is necessary in the clone-limited strategies because the unassigned clones that provide the STS markers are rapidly depleted. This occurs in the simulation with the double-end paired STS selection scheme after about 1400 STS assays. At this point 99% of the genome is covered by about 150 contigs of almost 700 kb (average size). There are then two classes of STSs and two classes of mapping clones. The majority of STSs (about 1100 of the 1400) mark clones that are positioned internally in the contigs. The second set of 300 STSs defines the ends of each of the contigs. Probes generated from both ends of the clones at the ends of contigs should efficiently extract the remaining information present in the library. Note that the entire library need not be screened in this second round of map construction, as a contig-end STS marker is likely to be present only in other contig-end clones. Since each contig end should contain about 4 clones, only 1200 clones have to be screened with each of the second-round probes. Clones hybridizing with one end of a contig can be hybridized with one another in order to determine the end clone that extends the farthest beyond the final STS in the contig. The other end clones can then be dropped, thus further reducing the number of contig-joining screens. Once contig ends are joined in this fashion, both sets of end clones (about 8 clones) can be positioned internally in a contig. The sources of potential end-joining STSs will rapidly shrink during the course of map closure.

Other Sources of STS Markers. One of the major benefits of mapping large numbers of STSs is to provide a map with high resolution due to a great density of individual markers. While the nonrandom strategy uses a minimal number of markers, which will be relatively evenly spaced, it does not provide the resolution afforded by mapping a greater number of random markers. However, once the physical map has been constructed, it becomes less expensive to position additional STS markers on the map. For example, one strategy that can be used to identify the positions of STSs on the physical map uses a protocol in which a series of binary pools, each containing multiple mapping clones in successively smaller combinations, are screened in a serial manner that allows unambiguous identification of the mapping clones corresponding to any new STS. Before the library has been organized, each STS would, on average, correspond to different clones in five such pools and each such "hit" would have to be pursued by five parallel screens in each round to identify all the corresponding mapping clones. In contrast, once all the clones have been assigned to contigs, the gridded array of the library can be reorganized in such a fashion so that each pool represents a contig or a set of contigs. When these organized pools are screened, typically only a single pool should contain the STS at each stage and only a single screen would be required in most rounds. In other words, once the clones have been assembled into contigs, it should be almost 5 times easier to position new, random STS markers onto the map. Thus, it may be most efficient to complete the map by using a nonrandom clone-limited strategy and then increase the resolution of the map by using new STS markers from other sources.

At the point when the map has been nearly completed it may be useful to use STS sources that are likely to incorporate important biological information into the map. One of the most useful types of STS markers are unique sequences that have already been positioned on the genetic map—for example, restriction fragment length polymorphisms (RFLPs) (15). The use of RFLPs as STSs would allow the contigs to be assigned to chromosomal regions and the correlation of the physical map with genetic linkage data. Another benefit might be obtained from correlating RFLP markers to the physical map early on: if the potential relationships between contigs can be assessed prior to second-round experiments aimed at map completion, then it might be possible to predict which contig ends are likely to cross-hybridize, and small-scale STS screening might allow more efficient completion of the physical map. Another potential source of STSs that are rich in biological information are cDNA clones that could be used to identify expressed sequences. The feasibility of inexpensively and rapidly assembling large numbers of non-cross-hybridizing cDNA clones has already been demonstrated (14).

Clone-Limited Strategies and Library Bias. In a clone-limited STS strategy the damage from skewed libraries is minimized. Some cloning strategies are inefficient for some sequences. This type of problem is perhaps best represented by the analysis of the progress of the physical map of the nematode genome (6). A greater number of clones had to be fingerprinted than were theoretically expected for the coverage obtained (9). In contrast, the *Escherichia coli* mapping experiments closely matched theoretical expectations (7). This suggests that some regions of the nematode genome tend to be underrepresented in the mapping libraries, while the *E. coli* sequences were more evenly distributed in the cloned inserts. It is not clear whether these results are genome-specific or due to the different types of vectors utilized in each project. With the clone-limited approach, the clones will be sorted into contigs earlier in a skewed library than in a more representative one, and redundant effort will not be spent uncovering a bias. Ultimately, with a clone-limited approach the coverage of the genome is solely dependent on the quality of the library.

Conclusions. The results of the computer simulations presented in this paper suggest that an over-reliance on unpaired random STSs may significantly increase the cost and retard the completion of physical maps of large genomes. In the initial stages of the project, the random STS selection strategy will provide a comparable degree of coverage to that obtained by clone-limited selection schemes. However, the rate of acquiring information from random STS mapping rapidly degrades and the process becomes increasingly inefficient as the map nears completion. The results of our simulations, as well as calculations based on the Poisson distribution, suggest that random strategies require enormous additional expense, as compared with clone-limited approaches, when the goal is to assign all the mapping clones into contigs. A second major result of our simulations is that paired STS markers can be used more efficiently than single STSs to build large contigs. Finally, the double-end clone-limited STS selection strategy generates a physical map that identifies a relatively small set of contig-end probes. These sequences can be selected in a second-round STS screen to identify the remaining unidentified overlaps. In other words, the proposed STS selection strategy (double-end clone-limited followed by a second-round screen using contig-end STSs) promises to elicit most of the information present in the library to obtain a map that represents the genome in as complete and continuous a fashion as the quality of the library allows. The application of these approaches to physical mapping may greatly reduce the cost and effort required, while also improving the quality, and thus the utility, of the genome map that is presented to the research community that the map is ultimately intended to benefit.

Note Added in Proof. We have recently become aware of a mathematical model of the random STS strategy (16) that leads to formulas whose predictions are remarkably close to the random STS curves in Fig. 1. That work does not apply to our clone-limited schemes.

We thank Carol A. Mayeda for a critical reading of the manuscript and for help with the figures. We thank Howard Ochman and Dan Krane for helpful conversations. M.J.P. is a Lucille P. Markey Scholar and his part in this work was supported by a Scholar's Grant from the Lucille P. Markey Charitable Trust. This work was also supported by Grant HG00357 to D.L.H. from the National Center for Human Genome Research and by National Institutes of Health Grant GM44889 and National Science Foundation Grant DMS-8802039 to S.A.S.

1. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
2. Burke, D. T., Carle, G. F. & Olson, M. V. (1987) *Science* **236**, 806–812.
3. Sternberg, N. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 103–107.
4. Sternberg, N., Ruether, J. & deRiel, K. (1990) *New Biol.* **2**, 151–162.
5. Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. & Frank, T. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7826–7830.
6. Coulson, A., Sulston, J., Brenner, S. & Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7821–7825.
7. Kohara, Y., Akiyama, K. & Isono, K. (1987) *Cell* **50**, 495–508.
8. Lander, E. S. & Waterman, M. S. (1988) *Genomics* **2**, 231–239.
9. Craig, A. G., Nizetic, D., Hoheisel, J. D., Zehetner, G. & Lehrach, H. (1990) *Nucleic Acids Res.* **18**, 2653–2660.
10. Evans, G. A. (1991) *BioEssays* **13**, 39–44.
11. Olson, M. V., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245**, 1434–1440.
12. Green, E. D. & Olson, M. V. (1990) *Science* **250**, 94–98.
13. Deininger, P. L. (1983) *Anal. Biochem.* **129**, 216–223.
14. Palazzolo, M. J., Hyde, D. R., VijayRaghavan, K., Mecklenburg, K., Benzer, S. & Meyerowitz, E. (1989) *Neuron* **3**, 527–539.
15. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32**, 314–331.
16. Arratia, R., Lander, E. S., Tavaré, S. & Waterman, M. S. (1991) *Genomics*, in press.