*Proc. Natl. Acad. Sci. USA*
Vol. 88, pp. 3917–3921, May 1991
Genetics

# Theoretical analysis of a physical mapping strategy using random single-copy landmarks

(DNA/contigs/sequence-tagged sites/pools)

EMMANUEL BARILLOT*, JEAN DAUSSET, AND DANIEL COHEN

Centre d'Étude du Polymorphisme Humain, 27 rue Juliette Dodu, F-75010 Paris, France

*Contributed by Jean Dausset, December 27, 1990*

**ABSTRACT** **An approach to physical mapping is analyzed here. This procedure consists of fingerprinting random clones with single-copy landmarks extracted randomly from a region of interest. Results are presented in terms of number of contigs (sets of overlapping clones), number of isolated clones, average length of a contig, and coverage of the genome by contigs larger than a given size. (*i*) The expected results of an ideal project are presented. (*ii*) Certain problems that could influence progress of the map (variability in clone insert length, double inserts, etc.) are considered. (*iii*) An optimal project, which consists of a 7-fold representative library fingerprinted with an average of five sequence-tagged sites per clone, is analyzed. (*iv*) We present strategical considerations for the proposed approach, and a strategy that minimizes the number of laboratory tests without significant information loss is proposed: clones are arranged on a matrix and pooled according to rows and columns. A fingerprint is determined for each pool, and analysis allows retrieval of the positive clones. This method reduces the number of laboratory tests done by a factor of 160.**

Recent laboratory and analytical techniques support the feasibility of constructing physical maps of complex genomes. The first step is to obtain a library of recombinant clones covering the region of interest with sufficient redundancy. Then information (a fingerprint) is extracted from each clone. Fingerprints of each pair of clones are compared, and overlap is declared when there is significant similarity. Finally, this repeated process leads to the building of sets of overlapping clones called contigs. The fingerprint can be a simple restriction fragment pattern (1, 2), a restriction fragment pattern containing repetitive sequences (3), a restriction map (4, 5), or a hybridization pattern obtained with probes for the ends of clones (6) or with random oligonucleotides probes (7). For fingerprinting clones, single-copy probes (SCPs) have the advantage of supplying powerful information. Indeed, two clones positive for the same SCP can be declared to be overlapping without ambiguity. In fact, this valuable information is available not only from hybridizations with SCPs but also from PCRs with sequence-tagged sites (STSs). The generic expression single-copy landmark (SCL) will be used to designate any single-copy fingerprinting entity—e.g., SCPs or STSs.

In relation to other fingerprint types, SCLs have the fundamental advantage of providing "a common language between all types of mapping" (as mentioned for STSs by Olson *et al.*, ref. 8). SCLs permit unification of results from very different approaches. Another advantage of STSs is that they need not be stored as biological material. The information resides in oligodeoxynucleotide sequences and protocols, which can be stored in an available data base (8).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Our purpose is to present a physical mapping method based on randomly selected SCLs (RSCL method) to evaluate the optimal number of clones and SCLs required for construction of a useful map by calculating the expected results and to elaborate a strategy that minimizes laboratory work.

## Principles of a RSCL Mapping Project

One way of obtaining SCLs is to extract them randomly from a region of interest (for example, a chromosome). Then the SCLs have to be tested on recombinant clones from a library covering at least the studied region. In this paper, the only case examined is that of randomly selected clones. A clone fingerprint consists of the SCLs it comprises. Two clones sharing a SCL are overlapping without any doubt because by definition a SCL appears only once in the genome. Thus, it is easy to assemble clones within contigs and to order the clones. Concomitantly, the assembly of SCLs within contigs (i.e., sets of two or more adjacent SCLs) can be achieved due to the close relation between the contigs of clones and the contigs of SCLs.

To plan a RSCL mapping project, the number of clones and SCLs necessary for obtaining the desired map must be evaluated. Several definable variables characterize the degree of completeness of the resulting map. According to Staden (9) and Lander and Waterman (10), any set of two or more overlapping clones will be called a contig. The term island will designate any resulting set of clones after analysis, even a contig or an isolated clone. The next section presents the formulae that describe the expected number of islands, the expected length of an island, and the expected number of isolated clones. [Lander and Waterman (10) have already calculated these formulae on the assumption that the fingerprint scheme can detect overlap between two clones whenever they share a minimum fraction of their length. We did not assume this; thus, our model should be more robust, but, on the other hand, the model can be applied only to the RSCL method.] Furthermore we introduce another variable that characterizes degree of completeness of the map: the "coverage." This last entity represents the percentage of the genome covered with contigs of a length greater than a given minimum size. For example, the coverage at the threshold of 3 megabases (Mb) is the proportion of the fingerprinted genome covered with contigs >3 Mb. The *Discussion* describes the basis for using SCLs in a physical mapping project. All mathematical proofs are deferred until the final section.

Abbreviations: SCP, single-copy probe; STS, sequence-tagged site; SCL, single-copy landmark; RSCL, randomly selected single-copy landmark; YAC, yeast artificial chromosome; Mb, megabase pair; CEPH, Centre d'Étude du Polymorphisme Humain.

*To whom reprint requests should be addressed.

### Mapping Results Expected With a SCL Strategy

Theoretical calculation of the expected results requires elaboration of a model that fits as closely as possible with the phenomenon being described. This fit is not easy, and some simplifying assumptions are necessary (the effects of these assumptions will be discussed later): (*i*) All clones are selected randomly from the genome. This assumption is probably not strictly true because of possible cloning bias. (*ii*) All SCLs are selected randomly from the genome. Again, this assumption is not strictly true due to cloning bias. For this case SCLs and clones are assumed to be independently distributed. (*iii*) All inserts are of equal size. This assumption depends on the nature of the vector: it is almost true with respect to cosmids or bacteriophages but not for yeast artificial chromosomes (YACs). (*iv*) All inserts are unique within any clone (i.e., a given clone cannot contain two inserts from different origin). Practically all libraries have a certain rate of double inserts (and even double YACs for YAC libraries).

We define the following symbols: $G$, haploid genome length in base pairs (bp); $L$, length of clone insert in bp; $N$, number of clones from the region of interest; $S$, total number of SCLs used; $s = SL/G$ (average number of SCLs per clone length); $C = NL/G$ (redundancy of coverage of the library); $a = N/G$ (probability per bp of starting a clone); $b = S/G$ (probability per bp of starting a SCL). The following equations describe the expected results of a physical mapping project based on the RSCL method.

**Expected Results with Clones of Constant Length.** PROPOSITION 1. *If the four assumptions, presented above, are true, then*:

(i) *The expected number of islands of clones is*

$$N(ce^{-s} - se^{-c})/(c - s).$$

(ii) *The expected number of isolated clones is*

$$N(e^{-s} + se^{-(c+s)} + (s/(c - s))^2e^{-c}((c - s - 1)e^{-s} + e^{-c})).$$

(iii) *The expected length of an island of clones is*

$$L[(1/c) + (s(2c - s)e^{-c} + c^2(c - s - 1)e^{-s})/(c(c - s)^2)]/[(ce^{-s} - se^{-c})/(c - s)].$$

(iv) *The expected number of islands of SCLs is*

$$S(ce^{-s} - se^{-c})/(c - s).$$

(v) *The expected number of isolated SCLs is*

$$S(e^{-c} + ce^{-(c+s)} + (c/(c - s))^2e^{-s}((s - c - 1)e^{-c} + e^{-s})).$$

(vi) *The expected length of an island of SCLs is*

$$L[(1/s) - (c(c + cs - 2s - s^2)e^{-s} + s^2e^{-c})/(s(c - s)^2)]/[(ce^{-s} - se^{-c})/(c - s)].$$

The formulae concern characteristics of the islands, but it is easy to deduce characteristics of the contigs: for example, substracting *ii* from *i* gives the expected number of contigs of clones.

Fig. 1 shows the expected number of contigs of clones as a function of the library redundancy for different values of $s$, the average number of SCLs per clone. First clones analyzed in the project increase the number of contigs because most of them create additional contigs; then clones begin to fall into contigs or to close gaps between contigs and reduce the number of contigs. When redundancy exceeds a particular value, the number of contigs does not decrease significantly because almost all junctions between contigs are impossible due to the lack of SCLs. At this point, it is not profitable to introduce additional clones. Fig. 2 shows the average length of a contig of clones as a function of the redundancy for different values of $s$. Because this number can be approximated by the ratio of the genome size and the number of contigs, the curves in Fig. 2 present inverse variations with regard to Fig. 1.

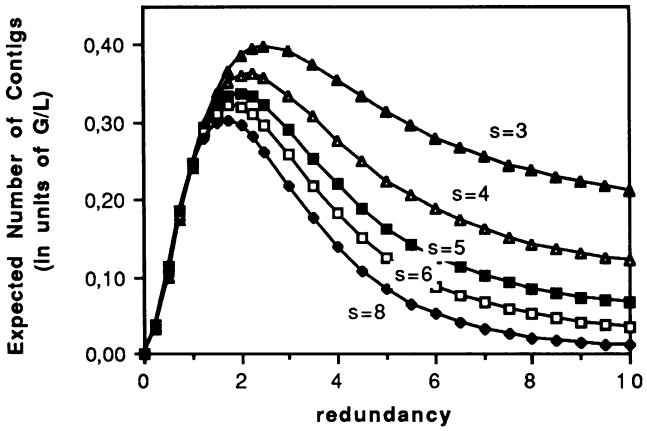


FIG. 1. Expected number of contigs as a function of redundancy of the library for different $s$ values, the average number of SCLs per clone. Inserts are assumed to have a constant length $L$. Because the vertical axis is graduated in $G/L$ units, the ratio of genome size to insert length, the value read on the graph has to be multiplied by $G/L$ to obtain the desired number in units of contigs. For example, for a 150-Mb human chromosome covered with 430-kb YACs, $G/L = 349$. When $c$ equals 7 [Centre d'Étude de Polymorphisme Humain (CEPH) YAC library] and $s$ equals 5, graph value is 0.10, and the effective $349 \times 0.10 = 35$ contigs.

We now consider the expected results when the assumptions are not true.

**Expected Results with a Nonuniform Clone and/or SCL Distribution.** If the first and/or second assumptions are not true, some parts of the genome are underrepresented, and progress of the project is slower in these regions. Completion of the map will then imply approaches other than a random strategy.

**Expected Results with Clones of Variable Length.** If the third assumption is not true (if the clone length is variable), formulae in *Proposition 1* will have to be corrected. The main formulae are as follows:

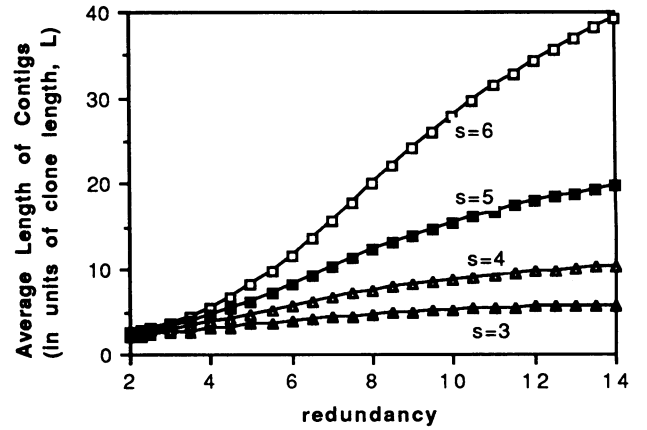


FIG. 2. Expected length of a contig as a function of redundancy of the library for different $s$ values, the average number of STSs per clone. Inserts are assumed to have a constant length $L$. The vertical axis is graduated in $L$ units. When $c$ equals 7 and $s$ equals 5, the average length of contigs will be 10.3 units of $L$—i.e., 4.4 Mb for the CEPH YAC library [$L$ = 430 kilobases (kb)].

PROPOSITION 2. *If clones are of variable length, with average length* L *and length density function* f(l), *then:*

(i) *The expected number* $I_c$ *of islands of clones is*

$$N\int_0^\infty f(l)\left[exp\left(-\frac{sl}{L}\right)exp\left(-\frac{c}{L}\int_1^\infty f(l'')(l''-l)dl''\right)+\int_0^l\left(\frac{s}{L}exp\left(-\frac{sl'}{L}\right)exp\left(-\frac{c}{L}\int_{l'}^\infty f(l'')(l''-l')dl''\right)\right)dl'\right]dl.$$

(ii) *The expected length of an island of clones is*

$$\left[G+N\int_0^\infty f(l)\left[exp\left(-\frac{sl}{L}\right)exp\left(-\frac{c}{L}\int_l^\infty f(l'')(l''-l)dl''\right)D(l)+\int_0^l\left(\frac{s}{L}exp\left(-\frac{sl'}{L}\right)exp\left(-\frac{c}{L}\int_{l'}^\infty f(l'')(l''-l')dl''\right)D(l')\right)dl'\right]dl\right]/I_c,$$

where

$$D(l)=\int_0^l\frac{cl'}{L}\left(\int_{l'}^\infty f(l'')dl''\right)\times exp\left[-\frac{c}{L}\left(\int_{l'}^l f(l'')(l''-l')dl''+(l-l')\int_l^\infty f(l'')dl''\right)\right]dl'-(L/c)exp\left[-\frac{c}{L}\left(\int_0^l f(l'')l''dl''+l\int_l^\infty f(l'')dl''\right)\right].$$

(iii) *The expected number* $I_s$ *of islands of SCLs is*

$$S\int_0^\infty\left[\frac{s}{L}exp\left(-\frac{sl}{L}\right)exp\left(-\frac{c}{L}\int_l^\infty f(l')(l'-l)dl'\right)\right]dl.$$

(iv) *The expected length of an island of SCLs is*

$$\left[G-S\int_0^\infty\left[\frac{sl}{L}exp\left(-\frac{sl}{L}\right)exp\left(-\frac{c}{L}\int_l^\infty f(l')(l'-l)dl'\right)\right]dl\right]/I_s.$$

**Expected Results with Libraries Containing Double Inserts.** If the fourth assumption is not true (i.e., if the library being fingerprinted contains double inserts), the proportion *p* of the double inserts has to be known to evaluate the effect on the expected map (Figs. 3 and 4).

PROPOSITION 3. *If 100*p% *of the clones contain double inserts, the formulae of* Proposition 2 *can be used after replacement of the measured values of the number of clones* N, *the average length* L, *the average number of SCLs per clone* s, *and the clone length density function* f(l) *with new values* N′, L′, s′, f′ (1):

$$N' = N(1+p),$$

$$L' = L/(1+p),$$

$$s' = s/(1+p), \text{ and}$$

$$f'(l) = (1-p)f(l) + \int_l^\infty pf(l')P(l', l)dl',$$

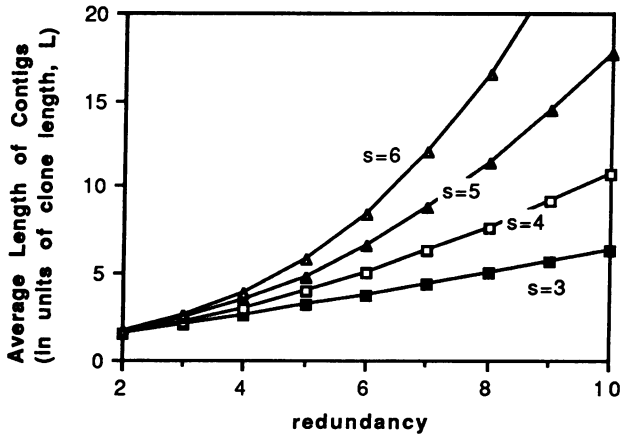


FIG. 3. Expected length of a contig as a function of redundancy of the library for different values of *s*, the average number of STSs per clone. Inserts are supposed to have a variable length with average value *L*, and a rate of 30% of double inserts is assumed in the library. The length density function was determined from the CEPH YAC library. The vertical axis is graduated in *L* units. Now the average length of contigs when *c* equals 7 and *s* equals 5 is 8.8 units of *L*—i.e., 3.8 Mb for the CEPH YAC library (*L* = 430 kb).

*where* P(l′, l) *is the probability for a double insert of length* l′ *of containing an insert of length* l.

## Discussion

Some general remarks can be made concerning the analysis of the different formulae and figures: As with any fingerprinting technique, the amount of work is inversely proportional to the average insert length. A 5-fold representative library is clearly minimal to construct a useful map (Figs. 1–4). On the other hand, too much redundancy is not useful: above a given redundancy (depending on the average number of SCLs per clone, the insert length density function, and the percentage of double inserts), it is not interesting to introduce new clones because they either fall into a contig or remain isolated but rarely merge two islands (almost all possible links have already been achieved). Note that the average number of SCLs per clone has a very sensitive effect on the progress of

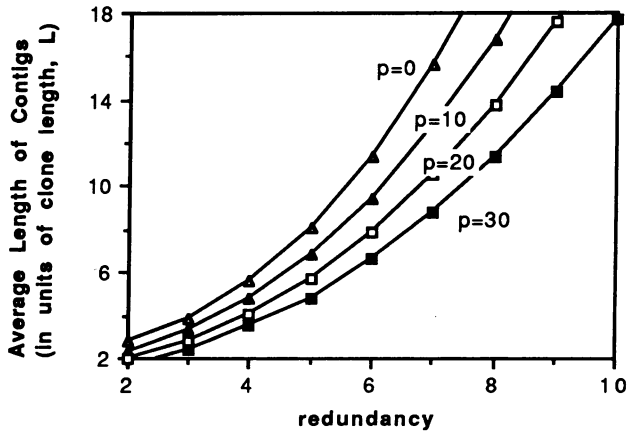


FIG. 4. Expected length of a contig as a function of redundancy of the library for different values of the percentage *p* of double inserts in the library. Average number of SCLs per clone, *s*, was 5. The vertical axis is graduated in *L* units, the average insert length. For a redundancy of 7, contigs are on average almost twice as large in a library free of double inserts as in a library containing 30% of double inserts. The length density function used for calculations is that of the CEPH YAC library, but curves for any library are available upon request.

the map and at least four SCLs per clone are required (Figs. 1–3). Variability of insert length in the library has an important impact on results: Due to the large inserts, the average contig length is greater with inserts of variable length than with inserts of constant length, but, on the other hand, the number of isolated clones is much higher because of small clones (data not shown). Double inserts significantly slow down the progress of map construction (Fig. 4). For example, when redundancy is 7, it is equivalent to build a map with a library free of double inserts and four SCLs per clone on average or with a library containing 30% of double inserts and five and one-half SCLs per clone—i.e., with an additional 38% of SCLs. In addition, double inserts are the main source of incorrect merges (a double insert can merge the respective contigs of its two inserts), even if some of them can be detected when they create inconsistency in the map. Thus, double inserts must also be accounted for: their quantity has to be estimated before starting the project.

All formulae are linear in $N$, the number of clones analyzed, and it would seem indifferent to map a whole genome directly or to execute the project chromosome by chromosome. In fact, the two approaches are not equivalent; Lander and Waterman (10) have already advanced several good reasons in favor of subdivision. The main effect is the reduction of the rate of false positives. Indeed, for a chromosome-specific project, the SCLs will fingerprint the inserts coming from the studied chromosome only, and, therefore, all double inserts from two different chromosomes will behave as single inserts; then the percentage of problematic double inserts is notably reduced.

When using a physical map to explore the region neighboring a landmark, knowing the probability that a given locus will be included within a contig larger than a given size is useful. An equivalent way to evaluate this probability is to find the proportion of the genome covered with contigs larger than a given size. Fig. 5 shows this coverage of the genome in function of this given size. The results confirm that four or five SCLs per clone is a minimal value.

To summarize, the choice of the values of $c$ and $s$ depends on the desired results but is also a compromise: values of $c$ that are too low would require a large number of SCLs. For example, mapping a 150-Mb chromosome with the present CEPH YAC library (ref. 11; 60,000 clones, average length 430 kb, redundancy 7) and 5 SCLs per clone (a total of 1750 SCLs) would give 33 contigs of mean length 3.8 Mb (Fig. 4, assuming a rate of 30% for double inserts). Fifty-nine percent of the chromosome would be covered by contigs >3 Mb, 11% by contigs from 2 to 3 Mb, and 10% by contigs from 1 to 2 Mb, leaving 16% of the chromosome unmapped (Fig. 5). Although the map would not be complete, it can be used for many purposes, and it will be more effective to consider nonrandom approaches to fill in persistent gaps.

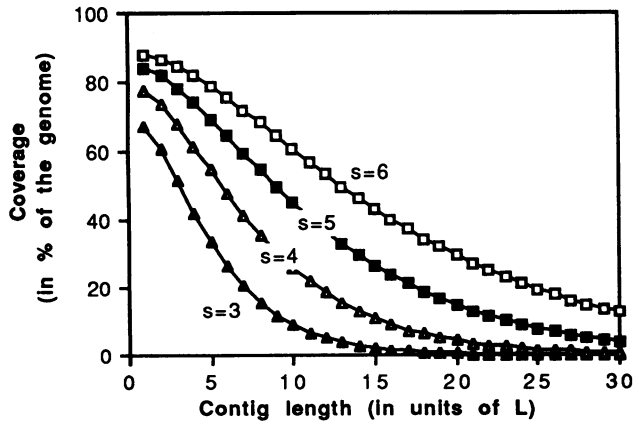


FIG. 5. Percentage of genome covered with contigs larger than a given size, as a function of this given size, for different values of $s$, the average number of STSs per clone, assuming a library containing 7 genome equivalents and 30% of double inserts. The horizontal axis is graduated in units of $L$, the average insert length. For example, when $s$ equals 5, 69% of the genome is covered with contigs >5 $L$ units and 45% with contigs >10 $L$ units.

## Mathematical Proofs

This section contains all the mathematical proofs of the propositions presented in *Results* and uses the same notations. Those readers essentially interested in applications may prefer to omit these demonstrations. For all calculations, the SCL length (typically 200–500 bp for STSs) was neglected with regard to an insert length (≈40 kb for cosmids, 100 kb to >1 Mb for YACs). Moreover, the end effects were not taken into account; this is justified when total fingerprinted length is large compared with insert length.

Let us select an arbitrary direction on the genome—for example, from left to right. A clone will begin at its left end and finish at its right end. Note that the first assumption of *Mapping Results Expected with a SCL Strategy* considers that the number of clones beginning within a length of $l$ bp follows an exponential law with a parameter of $cl/L$ and the second assumption states that the number of SCLs in a length of $l$ bp follows an exponential law with a parameter of $sl/L$.

*Proof of Proposition 1:* To prove $i$, we must calculate the probability for one clone of being at the right end of an island—i.e., of being the last right clone of an island. This situation occurs when a clone contains no SCL or when no other clone begins after it and before its last SCL. The first event has the probability $\exp(-s)$ (remember the exponential law), and the second event is obtained by integrating $[b \exp(-bl) \exp(-a(L - l))]$ from 0 to $L$ (the first term of the integrand expresses the probability for the last SCL of being $l$ bp from the end of the clone, the second one the probability that no clone begins on the $L - l$ bp before the last SCL). The two events are incompatible, so summing them gives the desired probability. Multiplying that probability by the number of clones gives the number of islands and demonstrates $i$.

Now let us calculate the probability for one clone of remaining isolated. This situation can happen in the absence of SCL on the clone [probability $\exp(-s)$] or when only one SCL is present on the clone and there is no other clone on this SCL [probability, $s \exp(-s) \exp(-c)$], or when several SCLs are present on the clone, but there is no other clone on these SCLs {note that the absence of another clone on the first and last SCL is necessary and sufficient, so when $l$ is the distance between the last SCL and the end of the clone and $l'$ is the distance between the first SCL and the beginning of the clone, the probability of the event is obtained by integrating $b \exp(-bl)\,[b \exp(-bl') \exp(-a(2L - l - l'))]$ from 0 to $L$ for $l$ and from 0 to $L - l$ for $l'$}. Multiplying the sum of these three probabilities by number of clones gives the number of isolated clones and shows $ii$.

To evaluate the expected length of an island of clones, substract from the total length of the genome $G$ the gaps between adjacent islands or add the overlap when the two islands present an undetected common part. An equivalent result is obtained by substracting the distance $l$ from the beginning of the last clone of an island to the beginning of the first clone of the following island, minus one clone length $L$. To calculate this distance, integrate all the possible distances $l$ (i.e., from 0 to ∞), ponderated by their occurring probability $P(l/\text{island end})$. This probability is evaluated with Bayes formula: $P(l/\text{island end}) = P(\text{island end}/l)P(l)/P(\text{island end})$. $P(\text{island end})$ has already been calculated (formula 1), $P(l) = a \exp(-al)$ (no clone on a length $l$, then a clone) and $P(\text{island end}/l) = \exp(-b(L - l))$ if $l < L$ (no SCLs on the overlap

between the two clones), $l$ otherwise (if $l > L$, there is no overlap, so therefore the end of the island is certain). If we multiply this distance, minus $L$, by the number of contigs and substract that product from $G$, we obtain the expected sum of all island lengths. Dividing by the expected number of islands approximates the expected length of an island. Computer simulations (we generated, randomly, clones and SCLs on the genome and analyzed the resulting maps) have shown that the error introduced by this approximation was <2%.

Reasoning similar to the above, applied to SCLs, shows *iv*, *v*, and *vi*. Note that any clone free of a SCL and included within another clone was not considered an island.

*Proof of Proposition 2:* Formulae of *Proposition 2* are only the generalization of those of *Proposition 1*. Reasoning is essentially similar, and detailed information is available upon request. Note simply that any clone free of SCLs and included within another clone was not considered an island.

*Proof of Proposition 3:* The following reasoning allows use of the formulae of *Proposition 2:* it is considered that each double insert is, in fact, two different single-insert clones. Thus, it is trivial to calculate the new values of the number of clones, the average length, the average number of SCLs per clone, and the length density function. To evaluate $P(l', l)$ all base pairs of the double inserts were assumed to have the same probability of being the one in which ligation had occurred. The objection can be made that the observed number of islands will be smaller than the calculated number because double inserts create junctions between the respective islands of their two inserts. But (*i*) it would be dishonest to consider two illegitimately merged islands as a single island and (*ii*) these false junctions would be detected and eliminated when they created inconsistancy in the map.

*Calculation of the coverage.* The coverage was calculated according to the method given by Michiels *et al.* (12), and we confirmed the results with computer simulations: the difference between the two methods is ≈5%.

## Conclusion and Perspectives

Finally, compare the performances of the RSCL strategy to the other approaches to physical mapping. Methods based on restriction fragment patterns provide poor information, and so they require many clones (with a minimal detectable overlap of 60%, a 12-fold representative library is needed to obtain the same results as the project described above, consisting of a library with a redundancy of 7, 30% of double inserts, and five SCLs per clone). Furthermore, with those methods, false positives are a greater problem because (*i*) it is more difficult to evaluate the rate of false positives and (*ii*) a clone can inaccurately merge two contigs, even when a clone is not a double insert, which is not possible with SCLs. The other methods presented in the Introduction seem interesting, too, but the impressive arguments of the "common language" and the "easy communicability" ensure a theoretical advantage with SCLs. Furthermore, the physical map of a precise region can be built with the RSCL strategy (needing only selection of SCLs from the studied part of the genome) but not with the other mapping strategies that assume clone selection.

The question of technical feasibility of the SCL physical mapping approach remains. The objection that complex fingerprints are unusable in large-scale mapping projects because these fingerprints would require too much work has been presented. Indeed, mapping a 150-Mb chromosome with the CEPH YAC library, for example (60,000 clones; average insert length 430 kb; redundancy 7), and five SCLs per clone would require 105 million tests (1750 SCLs tested on 60,000 clones). With this number, a direct approach is clearly impossible.

To bypass this problem, YACs can be grouped together in "pools" on which each SCL will be tested. This clone-pooling strategy is analogous to the method described by Evans and Lewis (6): imagine that all clones (the $N$ clones from the library) are arranged in a two-dimensional matrix and pool the clones according to rows and columns of the matrix (two copies of the library are needed). Then test a given SCL on each pool (now there are $2N^{1/2}$ tests instead of $N$, a noticeable reduction). A positive clone will render its row and column positive, so all positive clones will be located at the intersection of a positive row and a positive column. But the reverse is not true: an intersection of a row containing a positive clone with a column containing a different positive clone is not necessarily a positive clone. In fact, if $P$ clones are positive, then generally $P$ rows and $P$ columns are positive (this is not always true because two positive clones can be in the same row or column), and $P^2$ intersections are potentially positive (i.e., candidates) but only $P$ intersections are effectively positive. To recover the actually positive clones, one or more additional matrices are needed. The new matrices must be configured as differently as possible from the previous one, so that a given illegitimate candidate from a given matrix will probably not be retrieved as an illegitimate candidate in another matrix and, thus, can be detected. Obviously, the higher the number of positive clones, the higher the number of matrices needed to eliminate all illegitimate candidates.

In fact, the clones can even be pooled in a three-dimensional space or in any dimension. In a paper in preparation, we describe how to choose the best pooling strategy to obtain the desired map with a low error rate. The results are encouraging: consider the above example (mapping a 150-Mb chromosome with the CEPH YAC library and using 1750 STSs) where 105 million tests were needed. When the 60,000 clones are pooled in four-dimensional space and 384 pools are used, 672,000 tests are needed, and only one false positive would be expected in the map. With regard to the 105 million tests of the nonpooling approach, the amount of work is divided by almost 160! In addition, this strategy can be used for general library screening: only 192 pools (YACs being arranged in a four-dimensional space) are needed to screen the 60,000 clones of the CEPH YACs library. Each pool would contain ≈4000 YACs, and analysis of the 192 results would yield a mean of seven false positives for seven true positives.

This work was supported in part by the Ministère de la Recherche et de la Technologie and the Association Française contre les Myopathies.

1. Olson, M. V., Dutchik, 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.
2. Coulson, A., Sulston, J., Brenner, S. & Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7821–7825.
3. Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A. & Moysis, R. K. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 6218–6222.
4. Kohara, Y., Akiyama, K. & Isono, K. (1987) *Cell* **50**, 495–508.
5. Kuspa, A., Vollrath, D., Cheng, Y. & Kaiser, D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8917–8921.
6. Evans, G. A. & Lewis, K. A. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5030–5034.
7. Craig, A. G., Nizetic, D., Hoheisel, J. D., Zehetner, G. & Lehrach, H. (1990) *Nucleic Acids Res.* **18**, 2653–2660.
8. Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245**, 1334–1335.
9. Staden, R. (1980) *Nucleic Acids Res.* **8**, 3673–3694.
10. Lander, E. S. & Waterman, M. S. (1988) *Genomics* **2**, 231–239.
11. Albertsen, H. M., Abderrahim, H., Cann, H. M., Dausset, J., Le Paslier, D. & Cohen, D. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4256–4260.
12. Michiels, F., Craig, A. G., Zehetner, G., Smith, G. P. & Lehrach, H. (1987) *CABIOS* **3**, 203–210.