The Annals of Statistics
1997, Vol. 25, No. 1, 169–185

INFORMATION AND THE CLONE MAPPING OF CHROMOSOMES

BY BIN YU[1] AND T. P. SPEED[2]

University of California, Berkeley

A clone map of part or all of a chromosome is the result of organizing order and overlap information concerning collections of DNA fragments called clone libraries. In this paper the expected amount of information (entropy) needed to create such a map is discussed. A number of different formalizations of the notion of a clone map are considered, and exact or approximate expressions or bounds for the associated entropy are calculated for each formalization. Based on these bounds, comparisons are made for four species of the entropies associated with the mapping of their respective cosmid clone libraries. All the entropies have the same first-order term $N \log_2 N$ (when the clone library size $N \to \infty$) as that obtained by Lehrach et al.

1. Introduction. The primary goal of the Human Genome Project is to sequence the entire human genome, which consists of about 3×10^9 base pairs (bp) of DNA. Current technology only permits sequencing of fragments of the order of a few hundred to a thousand base pairs of DNA in a single reaction. Consequently, much effort is devoted to fragmenting large DNA molecules, such as chromosomes, in such a way that the sequenced fragments can be readily assembled. Clone maps, which are one form of physical mapping, play a key role in this process, as well as providing a resource permitting the detailed study of chromosomal regions of biological interest.

A clone map of part or all of a chromosome is the result of organizing order and overlap information concerning collections of DNA fragments called clone libraries. Such libraries consist of many, typically thousands or tens of thousands, of DNA fragments from a chromosome or region of interest. Each fragment exists as an insert in an autonomously replicating DNA sequence, which resides within, and replicates with its host cells. In this manner it is possible to generate many copies of the fragment of interest, and the name clone is thus used as an abbreviation for the longer and more accurate name: cloned DNA fragment.

A large clone library might consist of 5000 cloned fragments of average length 100,000 base pairs, from a chromosome of length 100,000,000 base pairs. Assuming that the cloned fragments are randomly located along the

Received February 1995; revised April 1996.

[1] Research partially supported by ARO Grant DAAH04-94-G-0232 and NSF Grant DMS-93-22817.

[2] Research partially supported by NSF Grants DMS-91-13527 and DMS-94-04267.

AMS 1991 *subject classifications.* Primary 60K99; secondary 94A99.

Key words and phrases. Configuration variable, contig, entropy, random clone library.

chromosome, this would mean that any particular spot on the chromosome should be represented on an average of five cloned fragments, giving rise to the term fivefold coverage, or a five-hit library. We note that a library of fragments of this size is still not suitable for DNA sequencing. Typically, one or two further stages of subcloning are needed prior to sequencing, and there may be additional mapping at these stages as well. In such cases both the libraries and the fragments will be smaller, but the principles of mapping remain much the same. For details on clone mapping from the perspective of applied probability, see Lander and Waterman (1988). Nelson and Speed (1994) have a more statistical perspective, and give further references to these aspects of the topic.

In their paper comparing the relative merits of fingerprinting cloned fragments of DNA by hybridization of oligonucleotide probes and by digestion into restriction fragments, Lehrach et al. (1990) raised two interesting questions concerning the creation of clone maps of a chromosome: (1) how much information is needed? and (2) how much information is gained by the hybridization and restriction digestion methods, respectively? The answer to the first question offered by these authors was $\log_2(\frac{1}{2}N!)$ for a library of N clones. This figure corresponds to the average amount of information (the entropy, see the following discussion) required to identify the true ordering of N objects labeled $1, 2, \dots, N$ when it is not possible to distinguish between the ordering $(i_1, i_2, \dots, i_N)$ and its reverse $(i_N, \dots, i_2, i_1)$, but otherwise all orderings are equally likely. However, it is not entirely clear why the ordering of objects in this way corresponds to any formal notion of a physical map, and even if there is such a correspondence, why all possible configurations should be equally likely.

To illustrate these points, let us briefly consider the cases of $N = 2$ and $N = 3$ clones, regarded mathematically as having identical length L bp and being randomly located along a chromosome of length G bp [cf. Lander and Waterman (1988)]. For two such clones we have two configurations, overlap or not, with quite unequal probabilities 2β and $1 - 2\beta$, respectively, where $\beta = L/G$. For three clones there are ten distinguishable configurations: one with no overlaps, three with exactly two clones overlapping, three with two different clone pairs overlapping, but no triple overlap and three distinguishable configurations involving a triple overlap. Again these can be seen to be far from equally probable. In practice, N will be in the hundreds or thousands.

In order to answer question (1) exactly, we would need to enumerate the set $\mathscr{X}$ of distinguishable configurations, calculate their probabilities $\{p(x): x \in \mathscr{X}\}$ and then go on to calculate the entropy $H(\mathbf{X}) = -\sum_{x \in \mathscr{X}} p(x) \log_2 p(x)$ of a random configuration $\mathbf{X}$. The first part of this program has been completed [see Newberg (1993)], but to our knowledge no one has carried the calculation of the probabilities beyond $N = 3$, although this is, in principle, possible. We do not know how to obtain the entropy $H(\mathbf{X})$ exactly, but in the following discussion we will find bounds on entropies of various configuration variables which are relevant to clone mapping.

The reason that the entropy $H(\mathbf{X})$ is the appropriate measure of information is explained in texts on information theory [see, e.g., Craig et al. (1990) and Rényi (1984)]. We content ourselves here with a brief informal explanation, applicable when the elements of $\mathscr{X}$ *are* equally likely, each having probability $1/|\mathscr{X}|$, in which case $H(\mathscr{X})$ achieves its upper bound $\log_2|\mathscr{X}|$. The argument goes like this: to identify any particular element $x \in \mathscr{X}$, we consider successive subdivisions of $\mathscr{X}$ into halves, quarters, eighths, and so on, and if we were told at each stage which half, quarter, eighth, and so forth contained the particular element, we would gain one bit of information each time. Clearly this process cannot finish in less than k steps, where $2^k \le |\mathscr{X}| < 2^{k+1}$, and this k is thus a lower bound to the number of such questions, equivalently bits of information, necessary to identify the particular element in question. More refined procedures can limit the amount of information necessary to $\log_2|\mathscr{X}| + \varepsilon$, where $\varepsilon > 0$ is as small as we wish [see, e.g., Rényi (1984)]. A similar but more complicated argument applies when the elements of $\mathscr{X}$ are not equiprobable [see the discussion of the noiseless coding theorem in Cover and Thomas (1991)].

In this paper we study the entropy $H(\mathbf{X})$ of a random configuration $\mathbf{X}$ most appropriate to the clone mapping problem. The study is done through seven other random structures, $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{Y}$ and Z, each of which can be regarded as embodying a greater or lesser amount of the structure implicit in $\mathbf{X}$, but whose entropies are more accessible. We derive a variety of exact and approximate expressions and lower and upper bounds for the entropies of these quantities. We compute these bounds for clone libraries of interest and the bounds are reasonable for all configuration variables considered and very tight for some. Based on these computations, comparisons are made for four "model" species in terms of information needed for the mapping of their respective cosmid clone libraries. It is somewhat surprising that all the entropies have the same first-order term $N \log_2 N$ when $N \to \infty$, as that obtained in Lehrach et al. (1990). We end the paper with some remarks concerning the more difficult question 2.

In closing this brief introduction we note that in the analysis which follows we essentially ignore the role of distances, although we do consider the placement variable $\mathbf{W}$ in units of thousands of base pairs. Many physical mapping methods produce some information concerning distances as well as clone order, and such information can be very useful in practice, even when (as is often the case) there are large error bounds attached. In particular, it would be misleading to compare the hybridization and restriction digest methods mentioned previously, solely on the basis of the information they produce concerning clone order. The restriction digest method produces fairly precise information about distances, whereas the hybridization method does not. An analysis, which incorporates distance as well as order and overlap information, is beyond us at this time.

2. What is a clone map? We now introduce several different but related abstractions of the notion of a clone map of a chromosome, this being informally an ordering of a library of cloned fragments of the chromosome in question. As noted previously, we adopt the mathematical model for a clone library used in Lander and Waterman (1988), namely, that the N cloned fragments can be identified with N randomly located subintervals of equal length L of a genome of length G. More formally, the left-hand endpoints (say) of the N intervals corresponding to the cloned fragments are independently located uniformly along $[0, G - L]$. It will be convenient at points in the argument to take an alternative, effectively equivalent view of the left-hand endpoints as being the points on $[0, G - L]$ of a homogeneous Poisson process with rate $\lambda = N/G$ per base pair.

2.1. *Fully ordered configurations.* Following the terminology of Alizadeh, Karp, Newberg and Weisser (1993), we use the term *placement* to describe a configuration of positions of the clones along the chromosome, that is, a specification $\mathbf{W} = (W_1, W_2, \dots, W_N)$, where $W_i \in [0, G - L]$ is the location of the left-hand endpoint of the ith cloned fragment, $i = 1, 2, \dots, N$. The units here are base pairs (bp) or kilobase pairs (kb); see the following discussion. Experimental procedures exist which could precisely determine these locations for a clone library, but most clone mappings have more modest aims, seeking to single out a less completely specified configuration from among a class of a priori equivalent alternatives. Before we turn to a discussion of such "coarser" configurations, we make a connection with the work of Lehrach et al. (1990), which stimulated this research. By the *linear ordering* of a clone library, we mean the sequence $\mathbf{V} = (V_1, V_2, \dots, V_N)$ of labels of the ordered left-hand endpoints of the clones; equivalently, the vector of *ranks* of $\mathbf{W} = (W_1, W_2, \dots, W_N)$ listed in reverse order. This variable seems to be the one considered in Lehrach et al. (1990).

2.2. *Island configurations.* We turn now to a second class of clone configurations, those based on the notion of an *island*, which is either a single clone, not overlapping with any other clone in the library, or a set of clones, each pair of which is connected by a chain of overlapping pairs of clones. Islands of two or more clones are usually called *contigs*, and many clone mapping projects have as their initial objective the determination of all contigs in their library and the ordering, up to inversion, of clones within contigs. This is usually the objective of *fingerprint-based* clone mapping, which attempts to infer clone order and overlap from information concerning each of the clones in the library, such as the list of fragment lengths following digestion by restriction enzymes, or the pattern of hits and misses following hybridization with a panel of probes. Fingerprint-based clone mapping projects usually turn to quite different techniques such as radiation hybrid or fluorescence in situ hybridization (FISH) mapping [see, e.g., Cox, Burmeister, Price and Myers (1990) and Trask (1991)].

The most basic island configuration variable is Z, the *number* of islands. More informative is the variable $\mathbf{U} = (U_1, U_2, \dots, U_N)$ of *island sizes*, which is a *partition of the integer* N, that is, $\Sigma_1^N U_i = Z$, $\Sigma_1^N i U_i = N$; or, equivalently,

U_i is the number of islands containing i clones. The components of $\mathbf{U}$ are the multiplicities of the *block sizes* of the *partition* $\mathbf{Q}$ *of the set* $\{1, 2, \dots, N\}$ of clone labels into islands. Here $\mathbf{Q}$ is the unordered list of disjoint subsets of $\{1, 2, \dots, N\}$, usually called blocks or equivalence classes, but called islands in this context, whose union is $\{1, 2, \dots, N\}$.

More informative again than $\mathbf{Q}$ is the configuration variable we term the *distinguishable orderings* of the clones and denote by $\mathbf{Y}$, namely, the variable which refines $\mathbf{Q}$ by including information on the ordering of clones within contigs, up to inversion. Thus $\mathbf{Y}$ tells us which clones are together in a contig and, up to a flip, the order in which they appear, but it contains no information on the relative positions of distinct islands along the genome.

There is one last refinement which we mention, namely, the configuration variable discussed in Newberg (1993), which includes information on the depth of coverage within contigs. We denote this configuration variable by $\mathbf{X}$, and note that it may be regarded as refining $\mathbf{Y}$ by containing not just information on the labels of the left-hand endpoints of the clones within each contig, up to inversion, but the labels of the interleaved sequence of the left-hand and right-hand endpoints of the clones, again up to inversion. Newberg (1990) calls two configurations of clones *topologically similar* if one can be transformed into the other by permuting the islands and/or reflecting some of the islands. An adjustment of the amount by which any pair of clones overlap leaves one with a topologically similar clone ordering, if no endpoint of a clone is moved past an endpoint of another clone. With this definition, $\mathbf{X}$ is the set of equivalence classes of topologically distinct configurations, called *interleavings* in Newberg (1993) and Alizadeh, Karp, Newberg and Weisser (1993).

2.3. *Pairwise overlaps*. Many fingerprint-based clone mapping projects take as their starting points the determination of pairwise overlaps among the clones in their library [see, e.g., Branscomb et al. (1990), Craig et al. (1990) and Fu, Timberlake and Arnold (1992)]. For this reason we define the *pairwise overlap* variable $\mathbf{P} = (P_{ij}\colon 1 \le i < j \le N)$, where $P_{ij} = 1$ if clones i and j overlap, and $P_{ij} = 0$ otherwise. It is clear that $\mathbf{P}$ can be obtained from $\mathbf{X}$ but not from $\mathbf{Y}$. In seeking to estimate $H(\mathbf{P})$ we do not mean to imply that pairwise comparisons are the best, or even an effective way to ascertain pairwise overlap information. Indeed, many of the most common clone mapping methods, such as STS-content mapping [Green and Green (1991)] and restriction mapping [Olson et al. (1986)], do not attempt to determine pairwise overlaps at all. Nevertheless, it seems to us of interest to ask just how large $H(\mathbf{P})$ is in relation to the entropies of other, more refined configuration variables.

This concludes our discussion of the different abstractions of the notion of a clone map of a chromosome based on a library of cloned DNA fragments from that chromosome. As with all mathematical idealizations, our variables all fail to account for many features of real clone mapping projects. Our hope is that the features we do retain are the important ones, and that our results are at least qualitatively correct and useful. We now illustrate the different variables just introduced in a simple case.

Example. Suppose that $G = 150$, $L = 20$ and $N = 8$. We list the set of configuration variables refining $\mathbf{W} = (120, 50, 10, 45, 105, 55, 20, 76)$. The vector of *ranks* of these values, viewed as observations on $[0, 150]$, is $(1, 5, 8, 6, 2, 4, 7, 3)$, and so $\mathbf{V} = (3, 7, 4, 2, 6, 8, 5, 1)$. Using the values in $\mathbf{W}$, it is easy to ascertain that

$$\mathbf{X} = \{(373'7')^*, (4264'2'6')^*, (88')^*, (515'1')^*\},$$

where 3 (resp. $3'$) denotes the left-hand (resp. right-hand) end of clone 3 or vice versa, and * indicates the fact that the ordering is only unique up to reversal. In a similar notation we have

$$\mathbf{Y} = \{(37)^*, (426)^*, (8), (51)^*\},$$

while $\mathbf{Q} = 15|246|37|8$, $\mathbf{U} = (1^1, 2^2, 3^1)$ and $Z = 4$.

3. Results. In this section we present our approximations to the entropy of the configurations just described. All proofs are collected in the appendices.

We have sought close nonasymptotic upper and lower bounds to the entropy expressions of interest, and have been quite successful in this regard with $H(\mathbf{Q})$ and $H(\mathbf{Y})$, and somewhat less so with $H(\mathbf{X})$ and $H(\mathbf{P})$. Exact calculations of $H(\mathbf{W})$ and $H(\mathbf{V})$ are straightforward. It is also of interest to consider our results asymptotically as $N \to \infty$. In so doing, we could keep L/G fixed and let $c = NL/G$ increase, or we could keep c fixed and let L/G decrease. A value of c in the range 3–10 is typical, with $c = 5$ being quite common, although values in the range 40–50 have been used. Our figures and tables have c fixed at 5.

The easiest entropy to evaluate is $H(\mathbf{W})$ which is just $N \log_2(G - L) \approx N \log_2 G$. This last expression can be rewritten as

$$H(\mathbf{W}) = N \log_2 N + N \log_2(L/c)$$

by making the substitution $c = NL/G$. It is clear that the leading term is $N \log_2 N$, and also that the second term depends on the units in which L is measured. The most reasonable choice would seem to be kilobase pairs (kb), in which the values $G = 100{,}000$ kb, $L = 40$ kb (corresponding to a cosmid library) and $c = 5$ give $N = 12{,}500$ and $H(\mathbf{W}) = 2.1 \times 10^5$, compared with $N \log_2 N = 1.7 \times 10^5$.

As pointed out in Lehrach et al. (1990), we may use Stirling's formula to get

$$\begin{aligned} H(\mathbf{V}) &= \log_2(\tfrac{1}{2}N!) \\ &\approx N \log_2 N + \tfrac{1}{2}\log_2 N - (\log_2 e)N - \log_2(\sqrt{2\pi}) - 1. \end{aligned}$$

Now let us define

$$\overline{L}(\mathbf{U}) = \mathbb{E}\{Z\}[\log_2 N - \log_2 e] + \tfrac{1}{2}\log_2 \mathbb{E}\{Z\} + \log_2\left(\sqrt{2\pi}e^{1/12}\right),$$

$$\overline{L}(\mathbf{U}) = \mathbb{E}\{Z\}\left(\log_2 \frac{Np^2}{q(1-q^N)} + (\log_2 q)c_N\right),$$

$$\underline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N,$$

$$\overline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N + \left(\log_2\left(\sqrt{2\pi}e^{1/12}\right)\right)\mathbb{E}\{Z\},$$

where $a_N = \mathbb{E}\{F^N \log_2 F^N\}$, $b_N = \frac{1}{2}\mathbb{E}\{\log_2 F^N\}$ and $c_N = \mathbb{E}\{F^N\}$, and F^N is a truncated geometric random variable with $p = e^{-c}$ and truncation at N. That is, for $q = 1 - p$, $P(F^N = j) = pq^{j-1}/(1 - q^N)$, $j = 1, 2, \dots, N$. We have the following bounds on the entropies.

Result A (Finite-sample entropy bounds). Let us introduce the following abbreviations:

$$\underline{H}(\mathbf{Y}) = \log_2 N! - \overline{L}(\mathbf{U}) - Np(1-p),$$

$$\overline{H}(\mathbf{Y}) = \log_2 N! - \underline{L}(\mathbf{U})^+ - Np(1-p) + \mathbb{E}\{Z\}H(F^N) + \log_2 N,$$

$$\underline{H}(\mathbf{Q}) = \log_2 N! - \overline{L}(\mathbf{U}) - \overline{M}(\mathbf{U}),$$

$$\overline{H}(\mathbf{Q}) = \log_2 N! - \underline{L}(\mathbf{U})^+ - \underline{M}(\mathbf{U}) + \mathbb{E}\{Z\}H(F^N) + \log_2 N,$$

$$\overline{H}(\mathbf{X}) = N\log_2 N + N\log_2(4/e) - \log_2 N,$$

$$\underline{H}(\mathbf{X}) = \underline{H}(\mathbf{Y}),$$

$$\overline{H}(\mathbf{P}) = \overline{H}(\mathbf{X}),$$

$$\underline{H}(\mathbf{P}) = \underline{H}(\mathbf{Q}),$$

where

$$H(F^N) = \sum_{j=1}^{N} -P(F^N = j)\log_2 P(F^N = j).$$

Then our main bounds may be expressed as

$$\underline{H}(S) \le H(S) \le \overline{H}(S),$$

where S may be $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Q}$ or $\mathbf{P}$.

Result B (Asymptotic expansions for entropies). The following expressions are valid as $N \to \infty$:

(i) $$H(\mathbf{W})/N\log_2 N = 1 + o(1),$$

(ii) $$H(\mathbf{V})/N\log_2 N = 1 + o(1),$$

(iii) $$(1 - e^{-c}) + o(1) \le H(\mathbf{X})/N\log_2 N \le 1 + o(1),$$

$$\text{(iv)} \qquad H(\mathbf{Y})/N\log_2 N = (1 - e^{-c}) + o(1),$$

$$\text{(v)} \qquad H(\mathbf{Q})/N\log_2 N = (1 - e^{-c}) + o(1),$$

$$\text{(vi)} \qquad H(\mathbf{P})/N\log_2 N = (1 - e^{-c}) + o(1).$$

The finite-sample bounds in Result A are really only useful when they are not very far apart. Fortunately, they are reasonably close for all four configuration variables considered here and very close for **Y** and **Q**. Figure 1 is the log-log plot of the entropy bounds for $c = 5$ and $N = 100, \dots, 20{,}000$, and it is clear that the bounds are very tight for $H(\mathbf{Y})$ and $H(\mathbf{Q})$, tight for $H(\mathbf{X})$, but not so close for $H(\mathbf{P})$. It is also comforting to see that **W**, **X** and **Y**, which are all reasonable definitions of a clone map, turn out to have very similar entropies. The other interesting and useful observation is that $H(\mathbf{V})$ is numerically very close to $H(\mathbf{Y})$ for the range of N that we considered and for $c = 5$. Therefore, the simple Stirling expansion for $H(\mathbf{V})$ can be used as a valid short-hand formula for $H(\mathbf{Y})$ when $c = 5$. This shows that Lehrach et al.'s intuition works well here since the coverage is high enough that most of the randomness in the configuration variable **Y** comes from the permutation which is captured in **V**.

It is perhaps remarkable that the entropies of **W**, **V**, **X**, **Y**, **Q** and **P** all turn out to have the first-order term $N\log_2 N$, asymptotically, as obtained in Lehrach et al. (1990) (cf. Result B). Moreover, the constant for the first-order

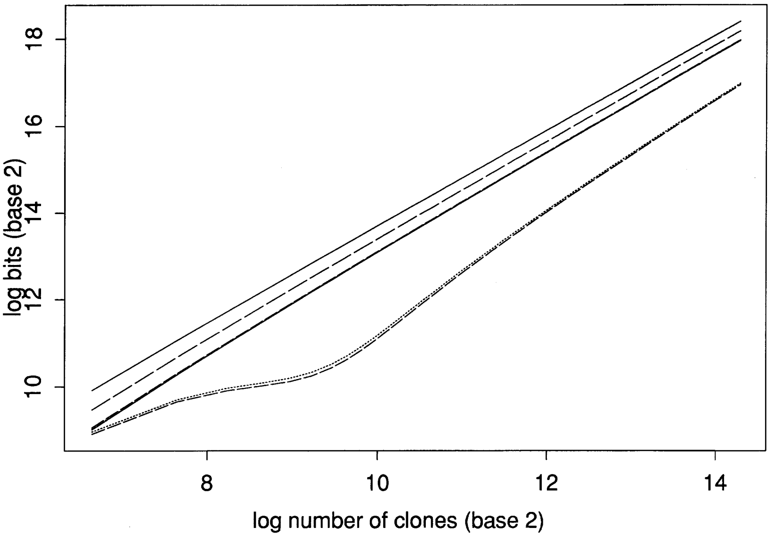

FIG. 1. $\log_2 H(\mathbf{W})$ (*top line*); $\log_2 \overline{H}(\mathbf{X})$ (*second line from top*); $\log_2 \overline{H}(\mathbf{Y})$, $\log_2 H(\mathbf{V})$ *and* $\log_2 \underline{H}(\mathbf{Y})$ *in the third line* (*cluster*) *and in that order from top*; $\log_2 \overline{H}(\mathbf{Q})$ *and* $\log_2 \underline{H}(\mathbf{Q})$ *in the bottom line* (*cluster*) *and in that order from top. Here the basic unit for* **W** *is* kb, $L = 40$ kb *and* $c = 5$. $\log_2 \underline{H}(\mathbf{Y})$ *and* $\log_2 \overline{H}(\mathbf{X})$ *serve as lower and upper bounds for* $\log_2 H(\mathbf{X})$ *and* $\log_2 \underline{H}(\mathbf{Q})$ *and* $\log_2 \overline{H}(\mathbf{X})$ *serve as lower and upper bounds for* $\log_2 H(\mathbf{P})$.

terms of $H(\mathbf{Y})$, $H(\mathbf{Q})$ and $H(\mathbf{P})$ is the same, namely, $1 - e^{-c}$. Unfortunately, this asymptotic result is not so useful for the values of N which are relevant here, because the term which makes the difference between $H(\mathbf{Y})$ and $H(\mathbf{Q})$ (cf. Figure 1) is $M(\mathbf{U})$, which is $O(N)$. The problem is that $\log_2 N$ is asymptotically larger than any constant term, but in this case it is much smaller than the corresponding constant (~ 260) in the $O(N)$ term.

An interesting fact which follows from the entropy bounds is that $H(\mathbf{P})/H(\mathbf{X}) \geq 0.20$ for $c = 5$ and $N = 100, 200, \ldots, 20{,}000$ (cf. Figure 2). (Note that the turns on the ratios for small N are probably artifacts of our bounds, not indicative of the true ratios of the entropies.) This implies that the pairwise variable $\mathbf{P}$ contains a substantial proportion of the information in the interleaving variable $\mathbf{X}$. However, although the pairwise mapping approach is definitely a good starting point for any clone mapping effort, recovering the pairwise variable $\mathbf{P}$ efficiently may well be improved by using multiple comparisons.

Table 1 lists the entropy bounds for specific cosmid ($L = 40$ kb) clone libraries corresponding to the G for a bacterium *E. coli*, yeast *S. cerevisiae*, roundworm *C. elegans* and humans. Here we observe behavior similar to that found in the figures. Table 2 gives the bounds on $H(\mathbf{W})$, $H(\mathbf{X})$ and $H(\mathbf{Y})$ for the last three species in relation to those of the bacterium *E. coli*. The ratios are seen to be species specific rather than specific to the configuration variables. We conclude that it makes sense to say, for example, that cosmid clone mapping for the roundworm requires about 40 times as much information as that for the bacterium *E. coli*, and that such mapping for humans

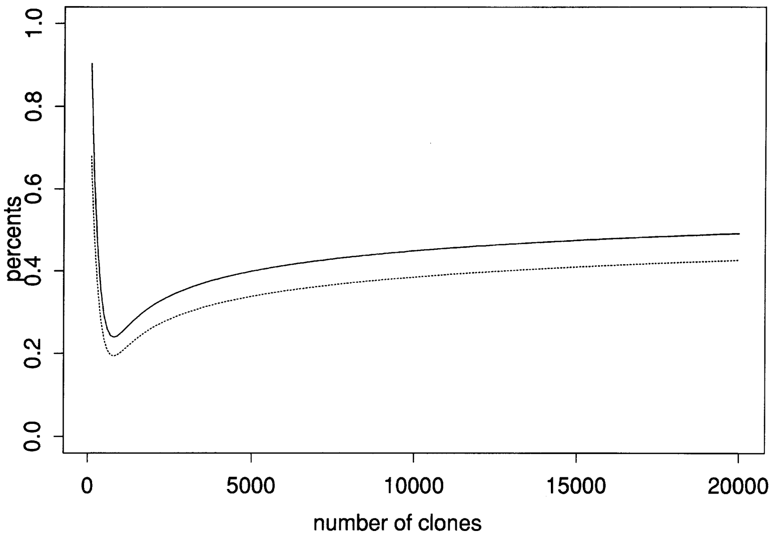

FIG. 2. *Lower bounds on* $H(\mathbf{P})/H(\mathbf{Y})$ (*upper line*) *and* $H(\mathbf{P})/H(\mathbf{X})$ (*lower line*), $c = 5$.

TABLE 1
Entropies and ratios for fivefold cosmid clone libraries of four species (ratios based on unrounded figures)

	Bacterium	**Yeast**	**Roundworm**	**Human**
	$\boldsymbol{N = 500}$	$\boldsymbol{N = 1{,}875}$	$\boldsymbol{N = 12{,}500}$	$\boldsymbol{N = 375{,}000}$
$H(\mathbf{W})$	6.0×10^3	2.6×10^4	2.1×10^5	8.1×10^6
$H(\mathbf{V})$	3.8×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\underline{H}(\mathbf{X})$	3.7×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\overline{H}(\mathbf{X})$	4.8×10^3	2.1×10^4	1.8×10^5	7.2×10^6
$\underline{H}(\mathbf{X})/\overline{H}(\mathbf{X})$	0.79	0.82	0.85	0.89
$\underline{H}(\mathbf{Y})$	3.7×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\overline{H}(\mathbf{Y})$	3.8×10^3	1.8×10^4	1.5×10^5	6.4×10^6
$\underline{H}(\mathbf{Y})/\overline{H}(\mathbf{Y})$	0.99	0.99	0.99	0.99
$\underline{H}(\mathbf{Q})$	1.1×10^3	5.5×10^3	7.0×10^4	3.9×10^6
$\overline{H}(\mathbf{Q})$	1.2×10^3	5.7×10^3	7.2×10^4	4.0×10^6
$\underline{H}(\mathbf{Q})/\overline{H}(\mathbf{Q})$	0.95	0.96	0.98	0.99
$\underline{H}(\mathbf{P})$	1.1×10^3	0.6×10^4	0.7×10^5	3.9×10^6
$\overline{H}(\mathbf{P})$	4.8×10^3	2.1×10^4	1.8×10^5	7.2×10^6
$\underline{H}(\mathbf{P})/\overline{H}(\mathbf{P})$	0.23	0.26	0.40	0.55

requires about 1500 times as much information as that for the bacterium *E. coli*.

4. Final comments. We close our discussion with some brief remarks on the important question (2) raised in Section 1: how much information is gained by the hybridization and restriction digestion methods, respectively? It is not our intention to offer a thorough discussion of this topic here, as we hope to present something more complete in a future paper. Rather, our aim here is simply to point out that the situation is not quite as simple as the discussion in Lehrach et al. (1990), page 45, suggests.

Suppose that we collect data $D_1, D_2, \dots, D_n$ on our clone library, for example, D_n could be the pattern of responses of our clones ($+$ or $-$) to the nth in a sequence of hybridization with short oligonucleotides. Each such

TABLE 2
Entropies of $\mathbf{W}, \mathbf{X}$ *and* $\mathbf{Y}$ *relative to E. coli*, $c = 5$

	Yeast	**Roundworm**	**Human**
$H(\mathbf{W})$	4.3	35	1350
$\overline{H}(\mathbf{X})$	4.5	37	1505
$\underline{H}(\mathbf{X})$	4.7	40	1700
$\overline{H}(\mathbf{Y})$	4.7	40	1690
$\underline{H}(\mathbf{Y})$	4.7	40	1700

data item has an entropy $H(D_n)$, indeed the full collection has an entropy $H(D_1, D_2, \dots, D_n)$, but if our aim is constructing a clone map using these data, the relevant entropy is $H(\mathbf{X}|D_1, D_2, \dots, D_n)$, the conditional entropy of the library configuration $\mathbf{X}$ given the data $D_1, D_2, \dots, D_n$. The computation of this quantity is not at all straightforward, even if the data items $D_1, D_2, \dots, D_n$ are mutually independent and identically distributed, given $\mathbf{X}$, as might be the case with a sequence of hybridizations involving short oligonucleotides of the same length. In such a case $H(D_1, D_2, \dots, D_n) = nH(D_1)$, but no such simplification occurs for $H(\mathbf{X}|D_1, \dots, D_n)$, although it should be possible to determine the asymptotic behavior of this quantity as $n \to \infty$. In a future paper we hope to discuss this issue more fully.

APPENDIX A

Upper and lower bounds for H(Q) and H(Y). Let $\mathbf{u} = (1^{u_1}, 2^{u_2} \dots)$ be a partition of the number N, and suppose that $\sum_1^N u_i = z$. We will use the notation $\mathbf{U}(\cdot)$ to denote the partition of N associated with the configuration in parentheses.

LEMMA A.1. *The number of configurations* $\mathbf{Q}$ *for which* $\mathbf{U}(\mathbf{Q}) = \mathbf{u}$ *is*

$$\frac{N!}{\Pi_{i=1}^{N}(i!)^{u_i} u_i!}. \tag{A.1}$$

PROOF. This is well known [see, e.g., Aigner (1979)].

LEMMA A.2. *The number of configurations* $\mathbf{Y}$ *for which* $\mathbf{U}(\mathbf{Y}) = \mathbf{u}$ *is*

$$\frac{N!}{\Pi_{i=1}^{N} u_i!} \frac{1}{2^{z-u_1}}. \tag{A.2}$$

PROOF. It is clear that the number we seek in this lemma is the number (A.1) multiplied by the number of directionless permutations of clones within islands. However, the latter is just

$$\prod_{i=2}^{N} \left(\tfrac{1}{2} i!\right)^{u_i}$$

and the result follows once we note that $\sum_{i=2}^{N} u_i = z - u_1$. □

LEMMA A.3. *The configurations* $\mathbf{Y}$ *with* $\mathbf{U}(\mathbf{Y}) = \mathbf{u}$ *are equally likely.*

PROOF. By symmetry.

EXAMPLE. It is easy to see that the configurations $\mathbf{y}_1 = \{(37)^*, (426)^*, (8), (51)^*\}$ and $\mathbf{y}_2 = \{(32)^*, (785)^*, (4), (16)^*\}$, for example, are equiprobable.

Corollary A.1.

$$\text{(i)} \qquad H(\mathbf{Q} \mid \mathbf{U}) = \log_2 N! - L(\mathbf{U}) - M(\mathbf{U}),$$

$$\text{(ii)} \qquad H(\mathbf{Y} \mid \mathbf{U}) = \log_2 N! - L(\mathbf{U}) - Np(1-p),$$

where

$$L(\mathbf{U}) = \mathbb{E}\left\{\log_2 \prod_{i=1}^{N} U_i!\right\} \tag{A.3}$$

and

$$M(\mathbf{U}) = \mathbb{E}\left\{\log_2 \prod_{i=1}^{N} (i!)^{U_i}\right\}. \tag{A.4}$$

Proof. These relations are consequences of Lemmas A.1 and A.2 and the equiprobable assertion of Lemma A.3.

We turn now to obtaining upper and lower bounds $\overline{L}(\mathbf{U})$, $\overline{M}(\mathbf{U})$ and $\underline{L}(\mathbf{U})$, $\underline{M}(\mathbf{U})$ of $L(\mathbf{U})$ and $M(\mathbf{U})$. In the calculations that follow, we use upper and lower bounds for factorials easily obtained from Stirling's formula [see, e.g., Feller (1968), page 52]

$$n^{n+1/2}e^{-n} \le n! \le n^{n+1/2}e^{-n}\sqrt{2\pi}e^{1/12}. \tag{A.5}$$

We also make use of the readily proved fact that the distribution of the sizes of islands is a truncated geometric with probability $p = e^{-c}$, where $c = NL/G$. More fully, the (ordered) sequence $F_1^N, F_2^N, \dots$ of island sizes consists of identically distributed random variables with common distribution $\mathrm{pr}(F^N = i) = pq^{i-1}/(1-q^N)$, $i = 1, 2, \dots, N$. Lander and Waterman (1988) give the proof for N large in which case F^N is approximated by a geometric. Taking the truncation into account gives more accurate results in our bounds when N is in the hundreds. It follows that $\mathbb{E}(Z - U_1) = Ne^{-c}(1 - e^{-c})$, since, for $i = 1, 2, \dots, N$,

$$\mathbb{E}U_i = \mathbb{E}\sum_{j=1}^{Z} I_{\{F_j^N = i\}} = \mathbb{E}\{Z\}P(F^N = i)$$
$$= np^2q^{i-1}/(1-q^N).$$

[More precisely, $\mathbb{E}U_i \approx np^2q^{i-1}/(1-q^N)$, since Z is very weakly related to the sequence $\{I_{\{F_j^N = i\}}\}$ $j = 1, 2, \dots$. Equality holds if Z is independent of this sequence.] We note that the preceding approximations are not expected to work for very small N's, but we believe they do work when N is in the hundreds, say larger than 500.

Lemma A.4.

$$\underline{L}(\mathbf{U})^+ \le L(\mathbf{U}) \le \overline{L}(\mathbf{U}),$$

where $x^+ = \max\{x, 0\}$,

$$\overline{L}(\mathbf{U}) = \mathbb{E}\{Z\}[\log_2 N - \log_2 e] + \tfrac{1}{2}\log_2 \mathbb{E}\{Z\} + \log_2\left(\sqrt{2\pi}e^{1/12}\right)$$

and

$$\underline{L}(\mathbf{U}) = \mathbb{E}\{Z\}\left[\log_2 N - (2c+1)\log_2 e + (e^c - 1)\log_2(1 - e^{-c})\right].$$

PROOF. Since $\Sigma_1^N U_i = Z$, we must have

$$\binom{Z}{U_1 \quad U_2 \quad \cdots} \geq 1,$$

in which case

$$\begin{aligned}\mathbb{E}\{\log_2 \Pi_i U_i!\} &\leq \mathbb{E}\{\log_2 Z!\}\\ &\leq \mathbb{E}\left\{\left(Z + \tfrac{1}{2}\right)\log_2 Z - (\log_2 e)Z + \log_2\left(\sqrt{2\pi}e^{1/12}\right)\right\}.\end{aligned}$$

Now $Z \leq N$, and so the right-hand side of the preceding formula is

$$\leq \mathbb{E}\{Z\}\log_2 N + \tfrac{1}{2}\mathbb{E}\{\log_2 Z\} - (\log_2 e)\mathbb{E}\{Z\} + \log_2\left(\sqrt{2\pi}e^{1/12}\right),$$

which is just the expression $\overline{L}(\mathbf{U})$.

For the lower bound $\underline{L}(\mathbf{U})$ we argue as follows:

$$\begin{aligned}L(\mathbf{U}) &= \sum_{i=1}^{N} \mathbb{E}\{\log_2 U_i!\}\\ &\geq \sum_{i=1}^{N} \mathbb{E}\{U_i \log_2 U_i - (\log_2 e)U_i\}\\ &= \sum_{i=1}^{N} \mathbb{E}\{U_i \log_2 U_i\} - (\log_2 e)\mathbb{E}\{Z\} \quad \text{since } \sum U_i = Z\\ &\geq \sum_{i=1}^{N} \mathbb{E}\{U_i\}\log_2 \mathbb{E}\{U_i\} - (\log_2 e)\mathbb{E}\{Z\} \quad \text{since } x\log_2 x \text{ is convex.}\end{aligned}$$

Now $\mathbb{E}\{U_i\} = Np^2q^{i-1}/(1-q^N)$ where $p = e^{-c}$ and $q = 1 - p$ and so, continuing the preceding sequence of inequalities,

$$\begin{aligned}L(\mathbf{U}) &\geq \sum_{i=1}^{N} Np^2q^{i-1}\left[\log_2 \frac{Np^2}{1-q^N} + (i-1)\log_2 q\right] - (\log_2 e)\mathbb{E}\{Z\}\\ &= Np\log_2 \frac{Np^2}{q(1-q^N)} + Np(\log_2 q)c_N - (\log_2 e)\mathbb{E}\{Z\}\\ &= \mathbb{E}\{Z\}\log_2 \frac{Np^2}{q(1-q^N)} + (\log_2 q)c_N - (\log_2 e)\mathbb{E}\{Z\},\end{aligned}$$

which is seen to be $\underline{L}(\mathbf{U})$ once we recall that $\mathbb{E}\{Z\} = Ne^{-c}$ and $c_N = \mathbb{E}F^N$. This completes the proof of Lemma A.4. Note that the leading term in each case is $e^{-c}N\log_2 N$. Obviously, $U \geq 0$. Hence $L(\mathbf{U}) \geq L^+(\mathbf{U})$. □

In the following lemma a_N and b_N are moments $\mathbb{E}\{F^N \log_2 F^N\}$ and $\tfrac{1}{2}\mathbb{E}\{\log_2 F^N\}$, where F^N has a truncated geometric distribution with parameter $p = e^{-c}$, $c = NL/G$, and truncation at N.

Lemma A.5.

$$\underline{M}(\mathbf{U}) \le M(\mathbf{U}) \le \overline{M}(\mathbf{U}),$$

where

$$\underline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N$$

and

$$\overline{M}(\mathbf{U}) = Ne^{-c}(a_N + b_N) - (\log_2 e)N + \log_2\big(\sqrt{2\pi}e^{1/12}\big) \times \mathbb{E}\{Z\}.$$

Proof. By definition,

$$M(\mathbf{U}) = \mathbb{E}\left\{\log_2 \prod_{i=1}^{N} (i!)^{U_i}\right\} = \mathbb{E}\left\{\sum_i U_i \log_2 i!\right\}.$$

We first use the lower bound of (A.5), obtaining

$$M(\mathbf{U}) \ge \mathbb{E}\left\{\sum_{i=1}^{N} U_i\big(i + \tfrac{1}{2}\big)\log_2 i - (\log_2 e) i\right\}$$
$$= \sum_{i=1}^{N} \big(i \log_2 i + \tfrac{1}{2}\log_2 i\big)\mathbb{E}\{U_i\} - (\log_2 e)N \quad \text{since } \sum_{i=1}^{N} iU_i = N.$$

Now $\mathbb{E}(U_i) = Np^2q^{i-1}/(1-q^N)$ as before.

To complete this, we need to recall that

$$\mathbb{E}\big\{F^N \log_2 F^N\big\} = \sum_{i=1}^{N} i(\log_2 i)\frac{pq^{i-1}}{1-q^N}$$

and

$$\frac{1}{2}\mathbb{E}\{\log_2 F\} = \frac{1}{2}\sum_{i=1}^{N} (\log_2 i)\frac{pq^{i-1}}{1-q^N}.$$

As mentioned in the statement of the lemma, these will be denoted by a_N and b_N, respectively, giving

$$M(\mathbf{U}) \ge Ne^{-c}(a_N + b_N) - (\log_2 e)N = \underline{M}(\mathbf{U}).$$

Turning now to the upper bound, the same reasoning leads to

$$M(\mathbf{U}) \le Ne^{-c}(a_N + b_N) - (\log_2 e)N + \big(\log_2\big(\sqrt{2\pi}e^{1/12}\big)\big)\mathbb{E}\{Z\},$$

where we have used the fact that $\sum_i U_i = Z$. However, the right-hand side of the preceding formula is just $\overline{M}(\mathbf{U})$ and we are finished. □

Lemma A.6.

$$0 \le H(\mathbf{U}|Z) \le \mathbb{E}\{Z\}H(F^N).$$

Proof.

$$H(\mathbf{U}|Z) = \sum_k \operatorname{pr}(Z = k)H(\mathbf{U}|Z = k)$$

and

$$H(\mathbf{U}|Z = k) \le H\big(F_1^N, \dots, F_k^N\big) \le kH(F^N),$$

since **U** is a function of $Z = k$ identically distributed random variables with the same truncated geometric distribution as F^N and its (conditional) entropy is bounded from above by the entropy of $F_1^N, F_2^N, \dots, F_k^N$ when they are independent. The lemma now follows by substituting this second equation in the previous one. □

Corollary A.2.

$$0 \le H(\mathbf{U}) \le Ne^{-c}H(F^N) + \log_2 N.$$

Proof. The relation is an immediate consequence of the lemma, once we recall that $Z \le N$ and $\mathbb{E}Z = Ne^{-c}$. □

APPENDIX B

An upper bound for H(X). In his thesis Newberg (1993) obtained recurrence relations and asymptotic expressions for the total number $C(N)$ of interleavings involving any number of islands which can be formed from N equal-sized randomly located cloned fragments. His asymptotic expression is given in the following result.

Proposition B.1.

$$C(N) \sim \frac{e^{3/8}\sqrt{2}}{8N}\left(\frac{4N}{e}\right)^N \quad as\ N \to \infty.$$

Corollary B.1.

$$\begin{aligned} H(X) &\le \log_2 C(N) \\ &= N\log_2 N + N\log_2\left(\frac{4}{e}\right) - \log_2 N \\ &\quad + \frac{3}{8}\log_2(e) - \frac{5}{2} + o(1) \quad as\ N \to \infty. \end{aligned}$$

APPENDIX C

Proofs of Results A and B.

Proof of Result A. Note that, for $S = \mathbf{Y}$ or $\mathbf{Q}$,

$$H(S) = H(S \mid \mathbf{U}) + H(\mathbf{U}).$$

The bounds for $S = \mathbf{Y}$ follow from Corollaries A.1(ii) and A.2 and Lemma A.4. The bounds for $S = \mathbf{Q}$ follow from Corollaries A.1(i) and A.2 and Lemma A.5. The bounds for $S = \mathbf{X}$ follow from Corollary B.1 and the fact that

X is a function of **Y** and the lower bound on $H(\mathbf{Y})$. We dropped the constant term in the upper bound for **X** in Corollary B.1 since it makes only negligible difference. Finally, the bounds for $S = \mathbf{P}$ follow from the facts that

$$H(\mathbf{P}) \geq H(\mathbf{Q}) \geq \underline{H}(\mathbf{Q})$$

and

$$H(\mathbf{P}) \leq H(\mathbf{X}) \leq \overline{H}(\mathbf{X})$$

(because **Q** is a function of **P** and **P** is a function of **X**). □

Proof of Result B. (i) and (ii) follow directly from the finite-sample bounds on $H(\mathbf{X})$, $H(\mathbf{Y})$ and $H(\mathbf{P})$, and the exact expressions for $H(\mathbf{W})$ and $H(\mathbf{V})$, and so does

$$H(\mathbf{P}) \geq (1 - e^{-c}) N \log_2 N (1 + o(1)).$$

Because **Q** is a function of **P**,

$$H(\mathbf{P}) = H(\mathbf{Q}) + H(\mathbf{P} \mid \mathbf{Q}).$$

For any given configuration **Q**, let $\mathbf{U} = \mathbf{U}(\mathbf{Q})$. Then, for any island of i clones, **P** can only take $2^{i(i+1)/2}$ possible values. It follows that

$$\begin{aligned} H(\mathbf{P} \mid \mathbf{Q}) &\leq \mathbb{E} \log_2 \Big(\prod_i 2^{U_i \times i(i+1)/2} \Big) \\ &\leq \sum_i \mathbb{E} U_i (i^2 + i)/2 \\ &= \sum_i N p^2 q^{i-1} (i^2 + i)/2 \\ &= N e^{-c} \sum_i p q^{i-1} (i^2 + i)/2 \\ &= N e^{-c} \big(\mathbb{E}\{F_N^2\} + \mathbb{E}\{F_N\} \big)/2(1 - q^N) \\ &= O(N) \quad \text{as } N \to \infty. \end{aligned}$$

Hence

$$H(\mathbf{P}) \leq H(\mathbf{Q}) + O(N) = (1 - e^{-c}) N \log_2 N (1 + o(1)). \qquad \square$$

Acknowledgment. The authors would like to thank Dr. Phil Green for his very valuable comments and insights, which led to a greatly improved revision of their initial paper.

REFERENCES

Aigner, M. (1979). *Combinatorial Theory*. Springer, New York.

Alizadeh, F., Karp, R. M., Newberg, L. A. and Weisser, D. C. (1993). Physical mapping of chromosomes: a combinatorial problem in molecular biology. In *Proceedings of the Fourth Annual ACM–SIAM Symposium on Discrete Algorithms*, Austin, TX. ACM, New York.

Branscomb, E., Slezak, T., Pae, R., Galas, D., Carrano, A. V. and Waterman, M. (1990). Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics* **8** 351–366.

COVER, T. and THOMAS, J. (1991). *Elements of Information Theory*. Wiley, New York.

COX, D. R., BURMEISTER, M., PRICE, E. R. and MYERS, R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. *Science* **250** 245–250.

CRAIG, A. G., NIZETIC, D., HOHEISEL, J. C., ZEHETNER, G. and LEHRACH, H. (1990). Ordering of cosmic clones covering the Herpes simplex virus type-I (HSV-I) genome—a test case for fingerprinting by hybridization. *Nucleic Acids Res*. **218** 2653–2660.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

FU, Y.-X., TIMBERLAKE, W. E. and ARNOLD, J. (1992). On the design of genome mapping experiments using short synthetic oligonucleotides. *Biometrics* **48** 337–359.

GREEN, E. D. and GREEN, P. (1991). Sequence-tagged sites (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications* **1** 77–90.

LANDER, E. S. and WATERMAN, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2** 231–239.

LEHRACH, H., DRMANAC, R., HOHEISEL, J., LARIN, Z., LENNON, G., MONACO, A. P., NIZETIC, D., ZEHETNER, G. and POUSTKA, A. (1990). Hybridization fingerprinting in genome mapping and sequencing. In *Genome Analysis*. *Genetic and Physical Mapping* (K. E. Davies and S. M. Tilghman, eds.) **1** 39–81. Cold Spring Harbor Laboratory Press.

NELSON, D. O. and SPEED, T. P. (1994). Statistical issues in constructing high resolution physical maps. *Statist. Sci*. **9** 334–354.

NEWBERG, L. A. (1993). Finding, evaluating and counting DNA physical maps. Ph.D. dissertation, Dept. Electrical Engineering and Computer Science, Univ. California, Berkeley.

OLSON, M. V., DUTCHIK, J. E., GRAHAM, M. Y., BROUDEUR, G. M., HELMS, C., FRANK, M., MACCOLLIN, M., SHEINMAN, R. and FRANK, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Nat. Acad. Sci. U.S.A.* **83** 7826–7830.

RÉNYI, A. (1984). *A Diary on Information Theory*. Wiley, New York.

TRASK, B. J. (1991). Fluorescence in situ hybridization—applications in cytogenetics and gene mapping. *Trends in Genetics* **7** 149–154.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
367 EVANS HALL #3860
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: binyu@stat.berkeley.edu