PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Design and Analysis of Chromosome Physical Mapping Experiments

David J. Balding

Phil. Trans. R. Soc. Lond. B 1994 **344**, 329-335
doi: 10.1098/rstb.1994.0071

Design and analysis of chromosome physical mapping experiments

DAVID J. BALDING

School of Mathematical Sciences, Queen Mary & Westfield College, University of London, Mile End Road, London E1 4NS, U.K.

SUMMARY

Mathematical and statistical aspects of constructing ordered-clone physical maps of chromosomes are reviewed. Three broad problems are addressed: analysis of fingerprint data to identify configurations of overlapping clones, prediction of the rate of progress of a mapping strategy and optimal design of pooling schemes for screening large clone libraries.

1. INTRODUCTION

A physical map of a genome is a set of markers whose locations have been accurately determined. The term *physical* indicates that the distances are measured in base pairs and distinguishes such maps from, for example, *genetic* maps which are based on recombination probabilities. A physical map can take several forms, but usually consists of an ordered collection of overlapping clones, DNA fragments which can be accurately reproduced in the laboratory. The task of map construction in this case is to take a library of randomly generated clones and to determine the order and pattern of overlaps of the clones. Chromosome-specific libraries are often available in which case mapping the genome of an organism reduces to mapping each chromosome.

Physical map construction is fundamentally the development of infrastructure to facilitate several avenues of research in molecular genetics (Daniels & Blattner 1987). A physical map allows the co-ordination of information from various sources. For example, physical maps can be integrated with genetic maps to enhance efforts to locate specific genes. Further, an ordered-clone physical map provides direct access to the DNA from regions of interest in the genome. In addition, the construction of the map gives information on the large-scale structure of the genome, such as the locations of restriction sites and interspersed repeats (Stallings *et al.* 1991).

The task of clone map construction can be likened to a one-dimensional jigsaw puzzle problem of considerable complexity. A clone map of a chromosome can consist of several thousand clones, in which case the number of possible orderings is extremely large. In certain aspects of map construction, mathematical (including statistical) analyses can be fruitful. Mathematical structures for representing data can enhance its usefulness whereas statistical analyses can allow optimal extraction of information from experimental data. Predictions of the rate of progress of a mapping strategy can help in the selection of relevant parameter values as well as assisting in choosing between alternative strategies. Careful experimental design can help to devise efficient experimental protocols. The cost in time and resources required to undertake such mathematical analyses can be substantial, particularly in view of the rapid advances in technology which lead to frequent changes of the questions requiring answers. In many cases the investment can, however, reap worthwhile rewards.

Several such applications of mathematical methods are discussed here. Candidates for inclusion are numerous and attention is naturally focused on those with which the author has been involved, largely as a visitor to the Theoretical Group at the Center for Human Genome Studies, Los Alamos National Laboratory, U.S.A. The paper consists of three main sections which give an overview of three distinct problems in map construction.

Section 2 discusses some aspects of map construction based on detecting overlapping clones using 'fingerprint' data. The discussion is influenced by the chromosome 16 cosmid map constructed at Los Alamos (Stallings *et al.* 1990), but the principles are relevant to other mapping protocols, such as the whole human genome YAC (yeast artificial chromosome) mapping project of Bellané-Chantelot *et al.* (1992), see also Lacroix & Codani (1991).

In § 3 the problem of predicting the rate of progress of a mapping strategy is considered. Several particular strategies have been analysed in the literature. Here, we outline general techniques based on the theory of alternating renewal processes, which can be adapted to physical mapping coverage problems.

Finally, § 4 addresses the problem of efficient pooling designs for large library screening projects, an important aspect of mapping based on single-copy

Phil. Trans. R. Soc. Lond. B (1994) **344,** 329–335
Printed in Great Britain

landmarks such as STSs (sequence-tagged sites). A well-studied class of combinatorial designs has been established to provide optimal solutions to a particular formalization of the library-screening problem. Approximate construction of these designs leads to a flexible approach to pooling which has important advantages over currently-employed methods: increased probability of a 'one-pass' solution, fewer pools required and the possibility of incorporating error detection.

2. CONSTRUCTING CLONE MAPS

The techniques of constructing ordered-clone genome maps were pioneered by several groups including Coulson *et al.* (1986) and Olson *et al.* (1986). Numerous subsequent mapping projects have developed a wide range of experimental techniques and strategies (see, for example, Davies & Tilghman 1992).

The core task of map construction is to identify configurations of overlapping clones. This is usually achieved by obtaining information about (or 'fingerprinting') each clone and then inferring overlaps from positive correlations in the fingerprint data. Several types of fingerprint data are available and consequently the details of the analyses vary. Most often, pairwise overlaps are first identified, and then higher level integration of this information is attempted. For some types of fingerprint data, simultaneous assessment of the data for many clones is possible, thus avoiding the intermediate step of detecting pairwise overlaps (Lehrach *et al.* 1990).

(*a*) *Fingerprint data*

An important source of fingerprint information is the restriction digest. Here, an enzyme cuts the clone at 'restriction sites' which are determined by the locations of a specific, short sequence. The digest may be 'complete', when the clone is cut at every restriction site, or 'partial' when only some of the possible cuts are made. The lengths of the resulting fragments are measured using gel electrophoresis (based on the observed distances travelled through a gel under an electric field). The restriction digest data is subject to several sources of statistical 'noise', of which the principal two are measurement error and missing fragments. Length measurement error is often approximately Gaussian with standard deviation proportional to length. Missing fragments may be due to lengths too short or too long for the resolution of the gel, or to distinct fragments of similar lengths not being distinguished.

Several partial digests, or two complete digests with distinct enzymes followed by a double digest with both, may give enough information to infer the locations of the restriction sites on each clone. Overlap of two clones is then inferred by matching the restriction site locations in the region of overlap (Kohara *et al.* 1987). This 'restriction map' provides a highly informative fingerprint, but is difficult to obtain (Lander 1991), particularly in view of the length measurement error. A simpler approach is based on the unordered fragment lengths from one or more digests. In this case all possible orderings must be considered in assessing the possible overlap of two clones, which poses challenging computational problems (Whittaker *et al.* 1993). Some theoretical aspects of map construction based on detecting pairwise overlaps using restriction fragment data are discussed by Branscomb *et al.* (1990). These authors simplify the analysis of the measurement error by discretizing and introducing a trinomial model for matching gel bands.

Another source of fingerprint data is the hybridization outcomes for one or more probes. The probe, labelled with a tracer, binds to a specific sequence, or one of a set of sequences, and hence the occurrence of one or more copies of the sequence(s) can be determined. It is possible to probe the individual fragments from a restriction digest and hence increase the information available from a shared fragment (Stallings *et al.* 1990). Statistical methods for detecting overlaps based on such compound restriction fragment–hybridization data are developed by Balding & Torney (1991). Alternatively, the entire clone can be probed with a large number of probes (Lehrach *et al.* 1990). Because hybridization outcomes are simply 0 or 1, the joint likelihood of large possible configurations of clones can be evaluated, a task which is computationally infeasible for more complex fingerprints. Fu *et al.* (1992) consider the number of probes required to detect pairwise overlaps.

Alternative sources of fingerprint data include end-specific probe hybridization of clones (Evans & Lewis 1989) and terminal sequencing of restriction fragments (Brenner & Livak 1989).

Another approach to map construction, which has recently become predominant, is based on single-copy landmarks, such as STSs detected using the PCR (polymerase chain reaction). STS maps have important advantages over other fingerprint-based methods (Olson *et al.* 1989). An STS is a global reference point, in contrast with reference points which are relative to one or more clones, a feature which assists in communication between laboratories. Further, because an STS is unique, the sharing of an STS by two or more clones provides firm evidence of their mutual overlap. This obviates the need for overlap analyses, but creates new requirements. The uniqueness of STSs means that several thousand of them are typically required for a chromosome mapping project and each must be tested against a library of several thousand clones. Experimental designs to simplify this task are the subject of § 4.

(*b*) *Detecting pairwise overlaps*

In early mapping projects (for example, Coulson *et al.* 1986; Vissing *et al.* 1987) the overlap of two clones was inferred by specific criteria such as a minimum number of restriction fragments apparently shared. This approach has a number of disadvantages: uncertainty about possible shared fragments is not incorporated and overlaps which are strongly supported by the data are not distinguished from those

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

which are merely plausible. Consequently, a more careful analysis can lead to more efficient use of the experimental data. In addition, a framework in which one simply decides whether or not two clones overlap is unsatisfactory. In dealing with large, complex maps subject to consistency criteria and the possibility of additional information, it is preferable to have a concise, readily interpreted summary of the strength of the evidence for overlap. One can then simultaneously investigate many plausible maps, rather than the unique, probably incorrect, map given by the overlap/no overlap outcomes.

There are important theoretical reasons for the view that the best summary of the strength of support given by data to a hypothesis such as 'these two clones overlap' is the likelihood ratio (Berger & Wolpert 1988). The likelihood ratio is simply the ratio of the probability of the fingerprint data given overlap to this probability given no overlap. Further, the likelihood ratio is needed in order to compute the probability that two clones overlap based on the fingerprint data and knowledge of the cloning techniques. These posterior probabilities allow efficient use of the experimental data and provide a convenient summary of the current state of information in a partially constructed map.

The use of posterior overlap probabilities calculated via Bayes Rule has been advocated by, for example, Michiels *et al.* (1987), Branscomb *et al.* (1990) and Balding & Torney (1991). In some other applications of statistics, the use of Bayes Rule is controversial because of difficulties in specifying prior distributions. Here, however, a prior probability of overlap based on independent and uniformly random clone locations is useful for most purposes and there is enough information about other parameter values to render unimportant the details of the prior distribution. Computational problems arise in the evaluation of exact likelihood ratios for complex fingerprints, such as those involving restriction digests, but approximations have been developed in some cases (Branscomb *et al.* 1990; Whittaker *et al.* 1993).

(*c*) *Identifying candidate maps*

The essential features of a candidate map may be conveniently represented as a graph in which nodes are identified with clones and the overlap of two clones is indicated by an undirected arc between the corresponding nodes. Similarly, the current state of a mapping project may be summarized by a graph in which arcs are labelled with the posterior overlap probability. By generating arcs according to these probabilities, highly likely overlap configurations can be generated and investigated.

The transition from pairwise overlap probabilities to map poses further problems. For large maps and typical fingerprint data, the number of possible maps with near maximum probability, based only on the pairwise information, is very large. This problem can in principle be reduced by calculating joint overlap probabilities for possible configurations of three or more clones. The computational aspects are, however, difficult for restriction fingerprints. The number of possible maps can be greatly reduced by exploiting consistency constraints imposed by the linearity of chromosomes. For example, for any four clones A, B, C, and D, it cannot be that A overlaps B, B overlaps C, C overlaps D, and D overlaps A unless either A overlaps C or B overlaps D. One possible way to exploit these constraints is to update the posterior probabilities according to whether or not a specific overlap leads frequently to 'illegal' configurations (Whittaker *et al.* 1993).

As mentioned above, in the case of hybridization-only fingerprint data, it may be possible to search for likely maps directly from the hybridization data, without first identifying pairwise overlaps. Alizadeh *et al.*(1992) adapted techniques developed for the travelling salesman problem to search for optimal maps based on hybridization data in the case of no experimental error. Mott *et al.* (1993) devised algorithms which allow for noisy data, based on simulated annealing and minimum-spanning subsets of probes.

3. PREDICTING THE RATE OF PROGRESS

For any given map construction strategy, it is of interest to predict the rate of progress for a given number of clones and amount of fingerprint information. For example, one may wish to predict the number of 'islands' of overlapping clones, the total length of the chromosome included in an island or the probability that a particular region is completely mapped. Such problems are referred to as 'coverage' problems. Theoretical analysis of coverage can lead to predictions of the feasibility of the project, comparison of alternative strategies and to optimal choices for the parameters of a given strategy, such as the frequency of hybridization sites or number of STSs. In addition, the optimal point at which to change strategies can be investigated.

Once again, it is usually assumed that clones are chosen independently and uniformly along the chromosome. Although there may in practice be regions which are difficult to clone, this assumption seems to be adequate for many purposes. Further, it is convenient to treat the chromosome as an infinite, continuous axis so that the centres of clones form a Poisson process of rate λ, say. If a clone occupies the interval (a,b) we will speak of it 'beginning' at a and 'terminating' at b. We also assume that clone lengths are independent and each clone has probability $G(x)$ that its length does not exceed $x \geqslant 0$.

Lander & Waterman (1988) make the above assumptions and assume further that there exists θ such that the overlap of two clones is detected whenever the overlap length forms a proportion at least θ of the mean clone length. They derive a number of results including the expected number of apparent islands and the expected length of an island. Arratia *et al.* (1991), Barillot *et al.* (1991) Ewens *et al.* (1991) and Torney (1991) all consider coverage problems for STS mapping in which overlap is detected if and only if an STS site occurs in the overall region. Palazzolo *et al.* (1991) and Zhang &

Marr (1993) consider the case that the STS probes are made complementary to the two ends of a particular clone, so that all overlaps of the chosen clone are detected. Fu *et al.* (1992) discuss coverage problems for a circular chromosome in the case of constant clone size and all overlaps detected.

A clone map can be viewed as a doubly-infinite sequence of independent variables, $\ldots, Y_{-1}, X_0, Y_0, X_1, Y_1, X_2, \ldots$, where the X_i are identically distributed and Y_i are identically-distributed. The X_i are interpreted as the island lengths and the Y_i as the 'ocean' lengths. This model is a particular case of an alternating renewal process and many existing coverage results, and important extensions of them, can be obtained in the general setting of the theory of such processes. For some mapping strategies, island length may not be independent of the lengths of neighbouring oceans. However, the relative order of the islands is unimportant for most coverage problems and any dependence will, in any case, be very weak.

The alternating renewal process model described above for clone mapping also describes several other well-studied problems. For instance, a type-II counter is a device which has a 'dead' period following the occurrence of each 'event'. The events are analogous to the start of a clone, and the distribution of dead periods is thus analogous to the distribution of islands. In addition, covered regions in a clone map are analogous to the busy periods of an $M/G/\infty$ queueing process.

Results from renewal process theory relevant to these applications are given by Hall (1988, Chapter 2). Let γ be the Laplace-Stieltjes (L-S) transform of the X_i. Then

$$\gamma(s) = \frac{1}{\delta(s)} - \frac{1}{\mu(s)}, \tag{1}$$

where δ is the L-S transform of the Y_i and μ is the L-S transform of the expected number of islands which overlap the interval $(0, t)$, given that an island terminates at 0. Assuming initially that all overlaps are detected, the ocean lengths have the exponential distribution with mean $1/\lambda$ and hence $\delta(s) = \lambda/(\lambda + s)$. Further, the probability $p_0(t)$ that the point $t > 0$ is not included in an island, given that an island terminates at the origin, is

$$p_0(t) = \exp\left[-\lambda \int_0^t \{1 - G(x)\}\mathrm{d}x\right]\mathrm{d}t, \tag{2}$$

and hence

$$\mu(s) = \lambda \int_0^\infty \exp\left[-st - \lambda \int_0^t \{1 - G(x)\}\mathrm{d}x\right]\mathrm{d}t. \tag{3}$$

The L-S transform of island length then follows from (1). Although explicit inversion of γ may often not be feasible, the case of fixed clone length can be solved (Hall 1988, p. 88) and the moments of island length can be evaluated in general. In addition, the Laplace transform π of the probability that an interval of a given length is completely covered satisfies

$$\pi(s) = \lambda s^{-2} \mathrm{e}^{-\alpha\lambda}(\gamma(s) - 1) + s^{-1}(1 - \mathrm{e}^{-\alpha\lambda}), \tag{4}$$

in which α denotes the mean clone length. An explicit solution is available in the case of fixed clone lengths (Hall 1988, p. 102).

The above formulae implicitly include in the definition of coverage islands consisting of only one clone, whereas one is often interested only in multi-clone islands, or 'contigs'. Corresponding results for contigs can be obtained by noting that the number of clones in an island has the Geometric distribution with parameter

$$\lambda \int_0^\infty \mathrm{e}^{-\lambda x} G(x)\mathrm{d}x, \tag{5}$$

and, conditional on an island having only one clone, its length distribution is given by G.

The assumption that all overlaps are detected is unrealistic for many mapping strategies: in practice, the overlap of two clones may not be identified if the length of the shared region is small. This may be approximately modelled by interpreting G in terms of an 'effective' clone length, shorter than the true length and such that any overlap of the effective region can be detected. For example, consider the case that clones have fixed length L and the probability that the overlap of two clones is detected, given overlap of length t, is $2t/L$ for $0 \leqslant t \leqslant L/2$ and 1 for $L/2 \leqslant t \leqslant L$. Coverage probabilities may be approximately obtained from (1) and (4) by assuming that all overlaps are detected and

$$G(x) = \begin{cases} 0 & x \leqslant L/2 \\ 2x/L - 1 & L/2 \leqslant x \leqslant L \\ 1 & x \geqslant L \end{cases} \tag{6}$$

In the case of random STS mapping, assuming that STSs occur at the points of a Poisson process of rate σ, independently of the clones, the probability that an overlap of length t is detected is $1 - \mathrm{e}^{-\sigma t}$. One can thus proceed approximately as above, assuming in effect that a clone terminates at its final STS. Alternatively, one can proceed exactly from (1) by deriving a formula analogous to (2) for the probability that the point t is not covered given that an island terminates at the origin.

4. POOLING

In this section we consider the design of large-scale library screening projects for rare or single-copy markers, in particular STSs. The task is to determine efficiently which clones in the library contain a given STS and hence overlap. For any collection of clones, a PCR assay establishes whether or not at least one of the clones contains the STS (Green & Olson, 1990).

Substantial gains in the efficiency of library screening experiments can be obtained by implementing a 'pooling' strategy. The essential feature of pooling is that, rather than assaying each clone individually, many clones are combined, or 'pooled', and a single PCR for a given STS is applied to the pool. If none of the clones in the pool is 'positive' (i.e. none contains the STS of interest) then they are all eliminated at the cost of only one assay. If at least one clone is positive, further assays are required to identify the positive clones. Whenever the proportion of clones which are

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

positive is small, pooling can substantially reduce the number of assays required to identify them.

As an illustration, consider a simple 'row and column' pooling design in which the clones are arbitrarily divided into lots of size 96, corresponding to the 96 wells of a standard microtiter plate (Evans & Lewis 1989). From each plate 20 pools are constructed by pooling together the samples from each of the 8 rows and the 12 columns of the microtiter plate. Thus each clone occurs in two pools, one corresponding to its row and one corresponding to its column. Now the 20 pooled samples are each screened for the STS marker. Assuming no experimental error, if all the clones are negative then this fact is immediately established. If there is precisely one positive clone among the 96 then two pools will give a positive result. These identify the row and column of the positive clone and hence uniquely identify the clone. A case of two positive clones is illustrated in figure 1. Here, the clones in row 3, column 3 and row 5, column 8 are positive and hence there are four positive pools, those corresponding to rows 3 and 5 and columns 3 and 8. However, it is not possible to uniquely identify the positive clones from the pool outcomes: the same outcomes would result if the clones in row 3, column 8 and row 5, column 3 were positive. Thus, at the cost of 20 assays, 92 clones are established to be negative, leaving only four unresolved clones which require further assays. In comparison, the simple 'one-clone-one-assay' design always requires 96 assays.

Barillot *et al.* (1991) generalize row and column designs to 'hypercube' designs of arbitrary dimension. They also propose the construction of multiple hypercubes which are 'mutually orthogonal' in some sense, to reduce the number of unresolved clones (that is, clones whose STS status cannot be determined from the pool outcomes). They illustrate these ideas in the case of a 72 000 clone library for which they devise two (three-dimensional) cubes, with a total of 258 pools. Chumakov *et al.* (1992) implement a pooling scheme for this library which incorporates some elements of the hypercube design. Their scheme is based on partitioning the library into 94 lots each of which is screened separately. A total of 2726 pools are required overall, although for any one STS only about 350 PCRs are actually required.

	1	2	3	4	5	6	7	8	pool
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	-	-	+	-	-	-	-	-	+
4	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	+	+
6	-	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-
pool	-	-	+	-	-	-	-	+	

Figure 1. A row and column design with 96 clones and 20 pools. Two clones are positive (row 3, column 3 and row 5, column 8) and hence four pools are positive.

One advantage of hypercube designs is that they are well-suited to manual construction. However, for large-scale projects which use a programmable robot this advantage is less important and other design criteria become more relevant. One such criterion is the amount of intervention by the experimenter. There are substantial advantages in having a fixed set of pools such that the status of all the clones can usually be resolved from these pool outcomes, without supplementary assays. In particular, such 'non-adaptive' designs facilitate automation using robots. In addition, since substantial effort is required to create a pool for PCR screening, it is desirable to minimize the number of pools required. Non-adaptive designs use the same set of pools in repeated screenings and hence minimize the total number of pools utilized. Completely non-adaptive designs are not feasible in practice. However, one can demand that a pooling design is nearly non-adaptive in the sense that a 'one-pass' solution occurs with high probability. The probability of a one-pass solution is thus an important performance criterion for pooling designs. Finally, efficient designs can be sensitive to experimental error. For example, in figure 1, any one false negative pool outcome will result in both a positive clone being missed and a negative clone being wrongly classified as positive. Therefore another important criterion is that the level of robustness to experimental error should be subject to control.

In general, hypercube designs are not optimal in terms of the two criteria discussed above. An alternative class of pooling designs is based on combinatorial structures known as maximum-size (v, k, t)-packings (henceforth 'packings'). Any collection of n subsets of $\{1, 2, \ldots, v\}$ can be interpreted as a pooling design on v pools and n clones: each subset in the collection corresponds to a clone and specifies the pools which contain that clone. A packing is a maximum-size collection subject to the requirements: (i) each clone occurs in precisely k pools; and (ii) any two clones occur together in at most t pools.

Requirement (ii) is crucial: it is not desirable for two clones to co-occur in a large number of pools since if one of the pair is positive it may then be difficult to determine the status of the other. Special cases of packings, known as t-designs and Steiner systems (Beth *et al.* 1986), have regularity properties which make them particularly suitable for pooling designs.

The number of clones n which can be accommodated by a packing is bounded above by

$$n \leqslant \binom{v}{t+1}\binom{k}{t+1}^{-1}. \qquad (7)$$

The bound (7) is achieved in the case of Steiner systems. The construction of t-designs and Steiner systems is difficult in general. However, due to their importance in many other applications, many explicit constructions have been documented in the literature

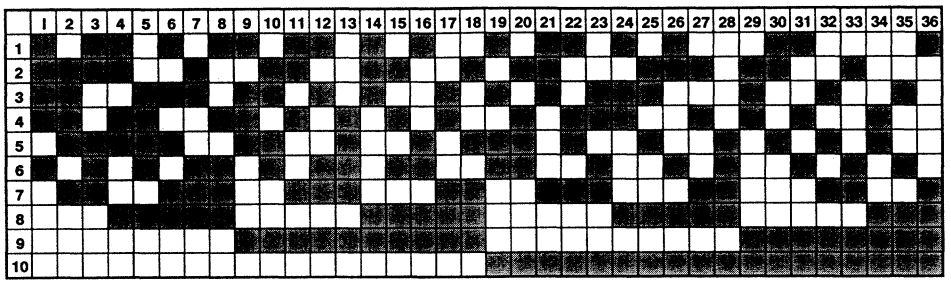

Figure 2. A t-design with $n = 36$, $v = 10$, $k = 5$ and $t = 3$. A shaded square in row i and column j indicates that the jth clone occurs in the ith pool.

(Beth *et al.* 1986). Efficient computational methods for constructing near maximum-size packings have also been developed (M. Goldberg & E. Knill, unpublished results).

Consider a packing design with $k = pt + 1$. If there are at most p positive clones then every negative clone must occur in a negative pool and hence the status of every clone can be inferred. If the number of positive clones exceeds p then this fact can be established from the pool outcomes. Therefore this design affords a one-pass solution whenever the number of positive clones does not exceed p. By appropriate choice of p, the probability of a one-pass solution can thus be made arbitrarily high. Further, if $k = pt + q + 1$ then, in addition, up to q errors can be detected. Balding & Torney (1994) establish that, for $p = 1, 2$ and any $q \geqslant 0$, the size of any nonadaptive p-positive, q-error solution on v pools satisfies the bound (7) with $k = pt + q + 1$. Therefore if the corresponding Steiner system exists it is best possible and, in general the packing design will either be optimal or near-optimal.

As a small example, consider the t-design with $n = 36$ clones and $v = 10$ pools illustrated in figure 2. In this design, every clone occurs in $k = 5$ pools and any two clones coincide in at most $t = 3$ pools. If there is one positive among the 36 clones, it can be uniquely identified from any nine of the ten pool outcomes. In addition, if more than one clone is positive then at least seven pools will be positive and hence this fact can be determined from any nine of the ten outcomes.

The design of figure 2 has a one-error-detection property: with one erroneous pool outcome, it is not possible to observe five positive pools. Further, if we know that there is no more than one positive clone then the design is also one-error correcting: if at most one pool outcome is in error then the one positive clone can always be uniquely identified. The design is not one-error-correcting in general, however, as if six positive pools are observed then it is not possible to decide whether this results from one positive clone and one false positive pool, or two positive clones and one false negative pool. If only false negatives (or only false positives) are possible, then the design is fully one-error correcting.

This example is included for illustrative purposes: most useful pooling designs will be substantially larger. However, even here the efficiency of t-designs is apparent. The best 'row and column' design on ten pools can accommodate only 25 clones in five rows and five columns, without any guaranteed error detection.

For a larger example, suppose that the library contains 2000 clones, among which the number of positives is binomial with mean 2. In this case a cube design can be constructed on 39 pools. The design would provide a one-pass solution only when there are 0 or 1 positive clones, which occurs with probability 0.41. The expected number of unresolved clones for this design is 18. A design based on a (39,9,4)-packing has probability 0.68 of being a one-pass solution and has expectation 11 for the number of unresolved clones. Thus the packing design is superior to the hypercube design with the same number of pools. In addition, the packing design incorporates two error-detection in the presence of one positive. At the cost of a few additional pools, one could guarantee one-error detection in the presence of two positives.

Financial support and hospitality from the Center for Human Genome Studies, Los Alamos National Laboratory, U.S.A., are gratefully acknowledged. In particular, I thank Dr David Torney for introducing me to the field and for continuing discussions and collaborations. Work supported in part by the U.K. Science and Engineering Research Council under grants GR/F 98727 and GR/J 05880 and by the Nuffield Foundation.

REFERENCES

Alizadeh, F., Karp, R.M., Newberg, L.A. & Weisser, D.K. 1992 *Physical mapping of chromosomes: a combinatorial problem in molecular biology*. Berkeley: International Computer Science Institute Technical Report TR-92-066.

Arratia, R., Lander, E.S., Tavaré, S. & Waterman, M.S. 1991 Genomic mapping by anchoring random clones: a mathematical analysis. *Genomics* **11**, 806–827.

Balding, D.J. & Torney, D.C. 1991 Statistical analysis of DNA fingerprint data for ordered clone physical mapping of human chromosomes. *Bull. math. Biol.* **53**, 853–879.

Balding, D.J. & Torney, D.C. 1995 Optimal pooling designs with error detection. *J. Comb. Th. A.* (In the press.)

Barillot, E., Dausset, J. & Cohen, D. 1991 Theoretical analysis of a physical mapping strategy using random single-copy landmarks. *Proc. natn. Acad. Sci. U.S.A.* **88**, 3917–3921.

Barillot, E., Lacroix, B. & Cohen, D. 1991 Theoretical analysis of library screening using a N-dimensional pooling strategy. *Nucl. Acids Res.* **19**, 6241–6247.

Bellanné-Chantelot, C., Lacroix, B., Ougen, P. *et al.* 1992 Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* **70**, 1059–1068.

Berger, J.O. & Wolpert, R.L. 1988 *The likelihood principle*. Hayward: Inst. Math. Statist.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Beth, T., Jungnickel, D. & Lenz, H. 1986 *Design theory*. Cambridge University Press.

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

Brenner, S. & Livak, K.J. 1989 DNA fingerprinting by sampled sequencing. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8902–8906.

Chumakov, I., Rigault, P., Guillou, S. *et al.* 1992 Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature, Lond.* **359**, 380–387.

Coulson, A., Sulston, J., Brenner, S. & Karn, J. 1986 Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. natn. Acad. Sci. U.S.A.* **83**, 7821–7825.

Daniels, D.L. & Blattner, F.R. 1987 Mapping using gene encyclopaedias. *Nature, Lond.* **325**, 831–832.

Davies, K.E. & Tilghman, S.M. (eds) 1992 *Strategies for physical mapping*. Cold Spring Harbor Laboratory Press.

Evans, G.A. & Lewis, K.A. 1989 Physical mapping of complex genomes by cosmid multiplex analysis. *Proc. natn. Acad. Sci. U.S.A.* **86**, 5030–5034.

Ewens, W.J., Bell, C.J., Donnelly, P.J., Dunn, P., Matallana, E. & Ecker, J.R. 1991 Genome mapping with anchored clones: theoretical aspects. *Genomics* **11**, 799–805.

Fu, Y.X., Timberlake, W.E. & Arnold, J. 1992 On the design of genome mapping experiments using short synthetic oligonucleotides. *Biometrics* **48**, 337–359.

Green, E.D. & Olson, M.V. 1990 Systematic screening of yeast artificial chromosome libraries by the use of the polymerase chain reaction. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1213–1217.

Hall, P. 1988 *An introduction to the theory of coverage processes*. New York: Wiley.

Kohara, Y., Akiyama, K. & Katsumi, I. 1987 The physical map of the whole E. coli chromosome: Application of a new strategy for rapid analysis & sorting of a large genomic library. *Cell* **50**, 495–508.

Lacroix, B. & Codani, J.J. 1991 *Computational aspects of human genome physical mapping*. INRIA Technical report 1560.

Lander, E. & Waterman, M.S. 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239.

Lander, E. 1991 Analysis with restriction enzymes. In *Mathematical methods for DNA sequences* (ed. M. S. Waterman), pp. 35–51. Boca Raton: CRC Press.

Lehrach, H., Drmanac, R., Hoheisel, J. *et al.* 1990 Hybridization fingerprinting in genome mapping and sequencing. In *Genetic and physical mapping* (ed. K. Davies & S. Tilghman), pp. 39–82. Cold Spring Harbor Laboratory Press.

Michiels, F., Craig, A.G., Zehetner, G., Smith, G.P. & Lehrach, H. 1987 Molecular approaches to genome analysis: a strategy for the construction of ordered overlapping clone libraries. *CABIOS* **3**, 203–210.

Mott, R., Grigoriev, A., Maier, E., Hoheisel, J. & Lehrach, H. 1993 Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. *Nucl. Acids Res.* **21**, 1965–1974.

Olson, M.V., Dutchik, J.E., Graham, M.V. *et al.* 1986 Random-clone strategy for genomic restriction mapping in yeast. *Proc. natn. Acad. Sci. U.S.A.* **83**, 7826–7830.

Olson, M., Hood, L., Cantor, C. & Botstein, D. 1989 A common language for physical mapping of the human genome. *Science, Wash.* **245**, 1434–1435.

Palazzolo, M.J., Sawyer, S.A., Martin, C.H., Smoller, D.A. & Hartl, D.L. 1991 *Proc. natn. Acad. Sci. U.S.A.* **88**, 8034–8038.

Stallings, R.L., Torney, D.C., Hildebrand, C.E. *et al.* 1990 Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. natn. Acad. Sci. U.S.A.* **87**, 6218–6222.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E. & Moyzis, R.K. 1991 Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. *Genomics* **10**, 807–815.

Torney, D.C. 1991 Mapping using unique sequences. *J. molec. Biol.* **217**, 259–264.

Vissing, H., Grosveld, F., Solomon, E. *et al.* 1987 Progress towards the construction of a total restriction fragment map of a human chromosome. *Nucl. Acids Res.* **15**, 1363–1375.

Whittaker, C.C., Mundt, M.O., Faber, V. *et al.* 1993 Computations for mapping genomes with clones. *Int. J. Genome Res.* **1**, 195–226.

Zhang, M.Q. & Marr, T.G. 1993 Genome mapping by nonrandom anchoring: a discrete theoretical analysis. *Proc. natn. Acad. Sci. U.S.A.* **90**, 600–604.

Note added in proof (3 May 1994): The following review paper, which focuses on the chromosome 19 cosmid map constructed at Lawrence Livermore National Laboratory, U.S.A., will appear in the August 1994 issue of *Statistical Science*: Nelson, D.O. & Speed, T.P. 'Statistical issues in constructing high resolution physical maps'.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES