Ordering of cosmid clones covering the Herpes simplex virus type I (HSV-I) genome: a test case for fingerprinting by hybridisation

A.G.Craig+, D.Nizetic, J.D.Hoheisel, G.Zehetner and H.Lehrach
Genome Analysis Laboratory, Imperial Cancer Research Fund Laboratory, PO Box 123, Lincoln’s Inn Fields, London WC2A 3PX, UK

Received January 25, 1990, Revised and Accepted March 27, 1990

ABSTRACT

To allow the efficient construction of ordered clone libraries, we have been investigating the use of 'oligonucleotide fingerprinting' as an approach to identify overlapping clones, and ultimately restore the linear order of the clone set. To test the effectiveness of the procedure, we have constructed a cosmid library from the genome of the human DNA virus HSV-I and used hybridisation to multiple oligonucleotides selected from the nucleotide sequence to reconstruct the order of clones and oligonucleotides on the genome.

INTRODUCTION

With the rapid increase in interest in cloning human genetic disease loci, the production of a high density molecular map based on ordered libraries of overlapping cosmid or yeast artificial chromosome (YAC) clones has become a high priority. The main approaches discussed to construct such ordered libraries fall into two classes, based either on fingerprinting strategies or on the use of hybridisation techniques analogous to chromosome walking.

Fingerprinting techniques, commonly based on the use of gel electrophoresis to generate restriction digest patterns from each clone, and their use to identify clone overlaps, has been applied successfully to produce ordered clone libraries of different organisms (1–3). This type of technique has two main advantages, the small increase in the mapping effort per clone with the size of the project (proportional to the logarithm of the number of clones) and the insensitivity of the protocol to interspersed repetitive sequences, expected to be a special problem in the analysis of mammalian genomes. Since individual clones have to be analysed in protocols involving a number of steps (growth, DNA isolation, enzymic reactions, gel electrophoresis, scoring of bands), rates of clone analysis are expected to be relatively low, though the rate of analysis can be increased by the use of automated systems using multiple fluorescent tags (4).

As an alternative, mapping approaches based on hybridisation are able to analyse large numbers of clones in parallel and therefore offer the theoretical possibility of much higher analysis rates. Approaches based on hybridisation of probes (derived from subgroups of clones) to clone filters have been proposed (5; Miles Brennan, personal communication), and partly used in the in the end stage of the analysis of the E. coli genome (3). Since the number of probes which can be used in each experiment is limited, the analysis of larger genomes requires linear increases in both the number of clones and the number of probe pools. The total effort in the analysis is therefore expected to increase with the square of the genome analysed, making this type of approach less favourable for larger genomes. In addition, hybridisation with pooled probes can be expected to be quite sensitive to repeat sequences in the probes, and especially to low copy number repeats in mammalian genomes.

To be able to combine the advantages of both types of protocols we have proposed a fingerprinting protocol (6–8) based on the use of oligonucleotides as hybridisation probes, which should combine the high data rates achieved by the parallel analysis of many clones in hybridisation experiments with the favourable scaling behaviour (N*log(N)) and repeat insensitivity of the gel fingerprinting approaches. Clones which show a similar pattern of hybridisation to the probes used ('hybridisation fingerprint') can be considered to be overlapping. This information on pairs of overlapping clones can be used to derive the linear order of clones and the oligonucleotide hybridisation sites on the genome.

We describe here an experimental test of the procedure. A cosmid library from HSV-I DNA was picked into 96-well microtitre dishes for storage and reference. Using a robotic device the cosmids were spotted in a high density pattern onto nylon membranes, grown in-situ, and converted into DNA by a colony lysis protocol. Filters were then hybridised to radioactively labelled oligonucleotides, selected from the sequence data (9), to assign oligonucleotide hybridisation signature to all cosmids. Analysis of the pattern of hybridisation of oligonucleotides to the clones allowed the identification of the linear order of the cosmids within the HSV-I genome and identified cosmids originating from each of the four isomeric forms of HSV-I (10).

+ Present address Molecular Parasitology Group, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK
MATERIALS AND METHODS

Library Construction
The library was constructed in the cosmid vector Lawrist4 (a derivative of LoristB (11); Pieter de Jong, pers. comm.). Vector arms were prepared by digestion with ScaI, phosphatase treatment, and digestion with BamHI. HSV-I DNA (the kind gift of Dr. D. McGeoch) was partially digested with MboI under conditions maximising the production of fragments in the size range 30 to 50 kb (12), followed by phosphatase treatment (to prevent ligation of non-contiguous fragments during library construction). Ligation was performed and the library was packaged as described (13). The library was plated on E. coli strain ED8767 (14).

After plating, single colonies were picked into 2 × YT supplemented with Hogness freezing medium and 50 μg/ml kanamycin, contained in 96-well microtitre dishes. Picked colonies were incubated for 15 to 30 hours at 37°C until a majority of the wells contained saturated cultures.

Filter Preparation
Cosmid cultures in the microtitre dishes were spotted onto nylon membranes. This was carried out using a robotic device which has a 96-tip head which can be moved accurately over a working area to spot 96 cosmids in an 8 × 12 array onto a nylon membrane (GeneScreen Plus—Dupont). By interleaving 16 patterns of the 96-well microtitre dishes it was possible to spot 1,536 cosmids over an area of 8 cm × 12 cm. Since this number is in excess of that required to ensure complete coverage of the HSV-I genome of 153 kb, we used 384 cosmids (4 microtitre dishes) for both colony size and cosmid copy number, detectable variations in both colony size and cosmid copy number, detectable.
The analysis was complicated by the fact, that HSV-1 exists as a mixture of four isomeric forms, which can be considered as subsegments of a permuted circular dimer sequence (Fig. 4). Accordingly, probes located on one end of the published sequence are found in association with probes from the other end. The different combinations of positive hybridisations with oligonucleotides flanking the repeat regions are shown in Fig. 5 along with the relevant oligonucleotide hybridisation data.

Figure 6(a) shows a subset of cosmids covering the entire HSV-1 genome. The hybridisation patterns for these cosmids (Fig. 6(b)) can then be used to derive an oligonucleotide map (Fig. 6(c)—only one isomeric form shown). An overall map has been (re-)produced from both the gain(+) and loss(−) of probes. A probe density averaging 1 per 10 kb, in the long unique region, and 1 per 5 kb, in the region around the internal repeats, was sufficient to detect four isomeric forms (Fig. 5). The density of probes used for the analysis of complex genomes would however be likely to be higher than that used here, giving more information and thus helping the resolution of analogous 'difficult' regions.

The predicted order of the positions of twelve cosmids was confirmed by preparing DNA and digesting it with EcoRI and ScaI, and analysing the restriction digest patterns by agarose gel electrophoresis (data not shown). The fragment sizes produced from the DNA digests were in complete agreement with the predictions from the HSV-1 DNA sequence.

While most probes fitted well into the expected pattern, unexpected hybridisation results were observed in a few cases. One such inconsistency was observed with the probe HSV3 which in addition to its expected location was found in combination with other oligonucleotide probes which could not possibly be present.
within the distance limited by the maximal size of a cosmid insert. The sequence recognised by HSV3 lies at position 17784 on the HSV-I genome but the results indicated an alternative association with probes 90 kb away. Inspection of the sequence identified a possible incomplete match at position 96709 of the HSV-I sequence due to an A-A mismatch at the 5' end of the oligonucleotide.

Similarly inconsistent results were observed with probe HSV4 (mismatch at position 56708). Additionally, it did not hybridise to a subset of cosmids expected to span the region containing this oligonucleotide sequence, identified by the flanking probes HSV3 and HSV5. No differences in restriction pattern could be identified, ruling out large scale deletions as an explanation for this phenomenon. DNA sequence analysis of this region (data

---

**Figure 3.** Binary signatures of half of the cosmids in the HSV-I library. After scoring, a binary pattern was generated for each cosmid. Some entries are not listed due to the absence of a cosmid at these positions. The order of the oligo probes in each pattern is HSV1 to HSV22.
Figure 4. The appearance of the HSV-I genome derived from the oligonucleotide hybridisation data (as a result of the presence of four isomers) The approximate positions of a subset of the oligonucleotide probes are displayed to show the orientation of the molecules.

Figure 5. A set of cosmids representing the four isomers of the HSV-I genome. Analysis of the binary signatures allows the differentiation of four classes of cosmid representing the four isomers of the HSV-I genome. One example of each class is shown plus the relevant oligomer hybridisation data which permits the assignment of each cosmid. The labels for the cosmids refer to their positions on the filter.
The comparison of the hybridisation patterns, analogous to other fingerprinting procedures, should lead to the identification of the linear order of the originally unordered clones, as well as the derivation of a linear map of oligonucleotide hybridisation sites in the genome (Fig. 6(c)). Once such oligonucleotide maps of specific chromosomes or genome exist, a small number of hybridisations should allow the localisation of any new clone on the pre-existing map, as well as the rapid and efficient ordering of new clone libraries constructed e.g. in more advanced vector systems. In this the oligonucleotide fingerprinting shares essential features with the recently proposed use of 'sequence tagged sites' (STS, 18), allowing the efficient rededuction of both specific clones and ordered clone libraries from information stored in a data base. Such oligonucleotide maps (essentially a form of very partial sequence information, which in theory can be extended to approach the complete sequence (19)) can be used to compare genomes of closely related organisms, identify the position of mutations caused by deletions or translocations, or follow changes in the genomes of cancer cells during tumorigenesis and tumor progression.

Though for practical reasons the experiments described here have been carried out using dodecanucleotides known to occur in the HSV-I sequence (although sequence variants and additional hybridisation sites were identified), this should represent a fair test of the situation encountered in the analysis of unknown genomes. Different types of probes can be considered. Completely random dodecanucleotides are expected to occur in double stranded DNA approximately once every 8 megabases (16 megabases single stranded DNA), corresponding to an expected frequency of 0.5% in cosmid libraries. If the sequence CpG is avoided in the selection of oligonucleotide sequences, an approximately two fold increase (to 1%) in hybridisation frequency is expected (reduced by the fraction of repeat sequences in the clones). Further improvements should be possible, if higher order Markov predictions are used, or if oligonucleotides hybridising more commonly than average are selected empirically. In addition, using modified hybridisation and washing conditions, a fraction (30 to 50%) of random undecanucleotides has been used successfully as hybridisation probes, leading to a fourth increase in hybridisation frequency. Also, a large number of di, tri, and tetranucleotide repeats (e.g. GTGTGTGTGTGT) have been found to have hybridisation frequencies of up to 10% in cosmid libraries from mammalian genomes (unpublished data).

Though less than the optimal 30% to 50% (8), we expect even hybridisation frequencies of a few percent, easily achievable by the use of (random) eleven or twelve-mers, to be sufficient to allow an efficient fingerprinting of libraries covering mammalian chromosomes.

Probes hybridising to 2% of the clones (assuming binary scoring, and neglecting the effect of errors or polymorphisms) will on the average give 0.14 bits of information per clone and hybridisation cycle (approximately 5.6 bits for each hybridising colony, 0.003 bit for each colony, which does not hybridise with the probe), a total of 1300 bits for each hybridisation of a filter
containing approximately 10,000 colonies. Since in analogy to the experiments described for the multiplex sequencing approach (20) hybridisation of 80 to 100 filters per hybridisation cycle should be feasible, the conditions tested here should allow data rates of minimally 100,000 bits per hybridisation cycle, far in excess of the at most few thousand bits per gel generated with considerable more effort in each fingerprinting experiment. For probes hybridising to 10% of all clones, approximately 0.5 bits per clone would be generated, resulting in potential data rates of close to 5,000 bits per filter, or approximately 400,000 bits for each hybridisation cycle of 80 filters.

This pilot study has tested many of the experimental variables of the oligonucleotide fingerprinting technique. Errors (or polymorphisms), unavoidable in any system set up to handle large numbers of clones, have been shown to be occurring at a level well within the tolerances of this approach and have not prevented the ordering of cosmids covering the entire HSV-I genome. In addition, the unusual structure of the HSV-I genome proved a stringent test for this mapping approach due to the presence of four isomeric forms of the virus. Using a density of probes less than that anticipated for the analysis of mammalian chromosomes (8), cosmids specific for each of the four isomers could be identified (Fig. 5).

We expect this or similar approaches to allow the efficient construction of overlapping clone libraries from mammalian chromosomes and genomes, simplifying the identification of genes responsible for human mutations, as well as offering a possible analytical tool to identify changes in a genome occurring either in mutations or in cancer formation.

ACKNOWLEDGEMENTS

The authors wish to thank Duncan McGeoch for providing the HSV-I DNA, Iain Goldsmith for preparing the oligonucleotides, and Joe Kitau for modifying the robot device. We would also like to thank Janet Roden and Lisa Gellen for technical assistance, Michele Price for help in preparing the manuscript, and Annemarie Frischau for comments on the manuscript. This work was funded by the Imperial Cancer Research Fund and an EMBO fellowship to Jörg Hoheisel.

REFERENCES