670–674 Nucleic Acids Research, 1995, Vol. 23, No. 4

The construction and analysis of M13 libraries prepared from YAC DNA

Mark Vaudin*, Avtar Roopra, LaDeana Hillier, Ryan Brinkman, John Sulston[1], Richard K. Wilson and Robert H. Waterston

Department of Genetics and Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA and [1]Sanger Center, Hinxton Hall, Hinxton, Cambridgeshire, UK

Received October 5, 1994 Revised and Accepted December 23, 1994

ABSTRACT

Yeast artificial chromosomes (YACs) provide a powerful way to isolate and map large regions of genomic DNA and their use in genome analysis is now extensive. We modified a series of procedures to produce high quality shotgun libraries from small amounts of YAC DNA. Clones from several different libraries have been sequenced and analyzed for distribution, sequence integrity and degree of contamination from yeast DNA. We describe these procedures and analyses and show that sequencing at about 1-fold coverage, followed by database comparison (survey sequencing) offers a relatively quick method to determine the nature of previously uncharacterized cosmid or YAC clones.

INTRODUCTION

Random or 'shotgun' sequencing of M13 libraries made from larger genomic clones such as cosmids is an efficient method to rapidly sequence genomic DNA (1–4). To be effective in large scale sequencing projects (5,6) efficient subclone library construction must be a routine procedure which produces a faithful representation of the original genomic clone. Random libraries typically are prepared by mechanical shearing of DNA and blunt-end ligation (7). However, the routine production of high quality libraries can be difficult to achieve from small amounts of DNA.

As part of the effort to sequence the genome of the nematode *C.elegans* (6,8) we have investigated the feasibility of constructing representative M13 libraries from yeast artificial chromosome (YAC) DNA (9,10). About 10% of the 100 Mb *C.elegans* genome map is represented solely as YACs (11,12). The remainder is available as both YAC or cosmid clones (13). The large size of YACs offers additional practical advantages if they can be routinely subcloned for sequencing and other manipulations.

In this article we describe methods which we have used to produce high quality libraries starting with relatively small amounts of YAC DNA. These methods also work well with cosmid DNA. YAC DNA suitable for subcloning is purified by pulsed field gel electrophoresis (PFGE) using a CHEF system (14) and DNA extraction from agarose. Libraries are made by mechanical shearing, end-repair, size fractionation and ligation into M13. These methods have been used to prepare subclone libraries from cosmids and YACs each containing at least 10^5 recombinant clones. Using blunt-end ligation at this stage circumvents the extra step of ligation with linkers. Sequence analysis of the resulting subclones shows a random distribution. The amount of contaminating yeast sequence in the YAC subclones is greatly reduced by secondary PFGE purification of the isolated YAC DNA band using different switching intervals and electrophoresis times.

Furthermore, we show that the generation of single random reads from YAC and cosmid sublibraries at 1-fold average coverage can provide useful insight into the nature of the DNA in that region.

MATERIALS AND METHODS

Preparation of yeast plugs and DNA purification

Agarose plugs of the YAC containing yeast strains were prepared by the method of Huxley (15). A 400 ml culture yielded sufficient plugs to provide purified DNA for construction of two M13 libraries, each containing about 10^5 recombinant clones. Each plug was poured from 0.5 ml agarose mix and measured $2.5 \times 0.8 \times 0.15$ cm^3. The plugs were cut in half longitudinally and the strips ($2.5 \times 0.4 \times 0.15$ cm^3) were loaded end to end across the width of a gel. Each gel measured $21 \times 14 \times 1$ cm^3 and was made from 200 ml of 1% agarose (Low melt, FMC Sea Plaque GTG) in 0.5 × TBE (45 mM Tris–base, 1 mM EDTA, pH 8.0). For each YAC three pulsed field gels were run. DRII and DRIII CHEF gel units (Bio-Rad Laboratories) were used. The exact pulse conditions were determined by the expected size of the YAC DNA and should be chosen to give the optimal separation of bands in the desired region. In general, YACs in the size range 100–500 kb were separated by electrophoresis for 40 h at 6 V/cm with a 30–50 s switch interval. The bands were visualized after ethidium bromide staining under low intensity, long wave UV and the YAC DNA band was carefully excised using a clean razor blade. The excised bands were loaded on a second set of pulsed field gels as described above and further purified using running

* To whom correspondence should be addressed

conditions in which the switching interval and electrophoresis time were altered. The favored conditions for 100–500 kb YACs were 20 h at 6 V/cm with a 50–90 s switch interval or 24 h at 6 V/cm with a 60–120 s switch interval. The purified YAC DNA band was cut from each gel. Strips (0.5 ml) were cut and soaked for a minimum of 2 h with at least one change of 100 mM NaCl, 1 × agarase buffer (1 mM Bis–Tris–HCl, 0.1 mM EDTA, pH 6.5). Each strip was melted in a 1.5 ml eppendorf tube at 68°C for 10 min and then equilibrated at 40°C for 10 min before adding 5 µl β-agarase I (NEB, 1 U/µl) and continuing incubation at 40°C for 2 h. The DNA was extracted twice with an equal volume of saturated, ice-cold phenol and precipitated with ethanol. The DNA was pelleted by centrifugation, washed with 70% ethanol, dried under vacuum and resuspended in 100 µl H_2O.

DNA shearing

The DNA was sheared in a final volume of 1 ml (100 µl DNA, 100 µl 5 M NaCl, 20 µl 0.5 M EDTA, 10 µl 1 M Tris pH 8, 770 µl H_2O) using a French pressure cell at 750 p.s.i. as previously described (16). A further 200 µl H_2O rinse was forced through the cell and added to the sheared DNA which was precipitated with 0.6 vol of propan-2-ol (on ice, 15 min, room temperature 15 min), washed with 70% ethanol, dried and resuspended in 26 µl H_2O.

End repair

End repair was performed by mixing all the sheared DNA with 4 µl of 10 × T4 polymerase buffer (500 mM Tris–HCl, pH 7.5, 100 mM $MgCl_2$, 5 mM DTT), 4 µl 2 mM dNTPs, 4 µl T4 polymerase (USB 5 U/µl), 2 µl Klenow (BCL 5 U/µl). Following incubation at room temperature for 30 min the volume was increased to 500 µl with H_2O, the solution extracted once with phenol and once with chloroform and precipitated with 100% ethanol, washed and dried under vacuum. The pellet was resuspended in 23 µl H_2O and incubated at 37°C, 30 min with 3 µl 10 × kinase buffer (USB, Cleveland, OH), 3 µl 10 mM ATP and 1 µl T4 kinase (USB 30 U/µl).

Size fractionation and ligation

The repair reaction was stopped with 3 µl blue loading dye and immediately electrophoresed on a 0.6% LMP agarose gel in 1 × TAE (40 mM Tris–acetate, 1 mM EDTA, pH 8.0). The 1–2 kb fraction was excised from the gel. The agarose was melted at 68°C, 10 min, the volume increased to 500 µl with H_2O, extracted once with phenol (frozen at –80°C for 15 min prior to centrifugation) and extracted once with chloroform and ethanol precipitated. The DNA pellet was resuspended in 10 µl H_2O.

A ligation reaction with 5 µl DNA, 20 ng *Sma*I cut, phosphatased M13 mp18, 1 µl 10 × ligation buffer (500 mM Tris–HCl pH 7.4, 100 mM $MgCl_2$), 1 µl 10 mM ATP, 1 µl 100 mM DTT (freshly prepared), 1 µl T4 DNA ligase (BCL 5 U/µl) and H_2O to 10 µl was incubated at room temperature for 16 h. The reaction was used to transform electrocompetent *E.coli* JS5 (Bio-Rad Laboratories) using an *E.coli* Pulser (Bio-Rad Laboratories). Transformants were plated on YT media and grown overnight at 37°C.

Template DNA was prepared and sequenced using Applied Biosystems 373A fluorescent DNA sequencers as described previously (8,17,18).

Database comparison

All sequences were compared to the public sequence databases using BLASTX and BLASTN programs for protein and nucleotide similarities respectively (19). In larger contigs likely genes were also identified using GENEFINDER (P. Green and L. Hillier, unpublished).

RESULTS AND DISCUSSION

YAC DNA isolation

It is difficult to obtain sufficient amounts of purified YAC DNA for subsequent manipulations such as subcloning small fragments for sequencing. We therefore determined the feasibility of producing M13 libraries suitable for sequencing from small amounts of YAC DNA.

Following the procedures described we were able to produce M13 libraries from YAC DNA routinely. The libraries contained at least 10^5 recombinants and had a very low background of non-recombinants (blue plaques) when plated. It was important to examine each step of the protocol to minimize any loss of DNA or reduction in efficiency at each of the stages. Even though it is possible to shear the YAC DNA using sonication (we have made sublibraries in this manner), a French pressure cell was preferred since it was more easy to control the size of fraction produced by adjusting the pressure applied on shearing. This was important to help maximize the amount of DNA produced of the required size. Therefore the reaction conditions given here were chosen and adjusted for DNA sheared in this manner.

Pulsed field gel electrophoresis conditions must be determined to give best separation of each YAC DNA from neighboring yeast chromosomes. It is important to minimize yeast contaminants by cutting the YAC band cleanly and conservatively (discussed later in 'yeast contamination'). This results in some loss of DNA. Further loss occurs in the second gel-purification step. Different methods of DNA extraction from the agarose were tried, including freeze/squeeze and melting followed by phenol extraction. However, treatment with β-agarase I as described was clearly the most efficient, yielding consistently higher amounts of DNA. As found previously (15), treatment of all the gel slices together in one tube did not work. Best results were obtained by treating 0.5 ml aliquots in separate 1.5 ml microcentrifuge tubes. Complete agarose digestion is essential for effective phenol extraction and efficient retrieval of the DNA.

Accurate quantitation of small amounts of DNA was difficult. Instead, the least number of gels required to routinely produce libraries in the order of 10^5 recombinants was determined. The DNA from three initial gels was found to be sufficient. No further quantitation was required. We have prepared libraries from YACs and yeast chromosomes ranging in size from 150 to 600 kb with no apparent difference in efficiency.

Distribution

It is important that distribution of the M13 subclones is representative of the original clone. We compared sequence from 458 M13 subclones of the YAC clone Y53B1 against the consensus sequence of the same region that was derived from cosmid subclones. In Figure 1, Y53B1 has been divided into five regions (A–E). Regions A, C and E, which represent 25, 25 and 43% of the YAC length respectively, were sequenced as cosmid clones. Regions B and D which make up the remaining 7% of the

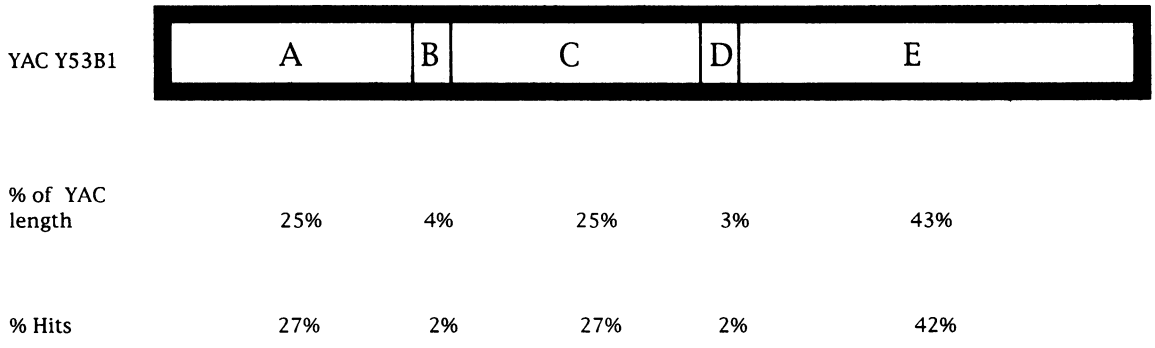

Figure 1. Distribution of 458 subclones of the *C.elegans* YAC Y53B1. The complete sequence of the region covered by the YAC was obtained from M13 libraries of cosmid DNA. The region consisted of three cosmid contigs separated by two YAC spans. The length of each contig is represented as a percentage of the overall length of the region. The percentage of the YAC clones that were homologous to each region by sequence comparison is shown on the lower line.

YAC were sequenced as plasmid clones following rescue from the YAC by homologous recombination in yeast (M. Vaudin, in preparation). Twenty seven percent of the subcones derived from purified Y53B1 DNA hit region A, 27% hit region C, 42% hit region E and a combined 4% hit B and D. Slight under representation in the gap regions may be due to an inordinately high number of repeats in these specific regions (6).

To obtain more information on subclone distribution we prepared an M13 library from a second YAC, Y50B11. This YAC covers 11 overlapping *C.elegans* cosmids, each of which was sequenced as part of the *C.elegans* project. The sequences obtained from 577 YAC subclones (0.6 × coverage) were assembled into a consensus contig of the corresponding overlapping cosmids. The start position of each YAC sequence read on each strand of the contig was derived to determine their distribution. If the start sites were distributed randomly the distances between positions of successive start sites should follow an exponential distribution. In our analysis the two largest gap sizes were 5 and 7 kb. Simulations with 577 random reads and given a YAC size of 350 kb (the size of Y50B11) revealed that a gap size of 5 kb does not significantly deviate from the proposed model.

It appears from this analysis that the method of YAC subcloning is susceptible to similar types of cloning bias that were previously known to occur with cosmids in M13.

Sequence integrity

The final sequence in all studies was determined to be the same as that obtained when M13 subclones were prepared from cosmids and sequenced under the same conditions. The sequence was colinear without deletions, insertions or substitutions. This was determined by visual comparison and assembly of YAC derived sequences into the cosmid projects. Chimeric clones were not found. Therefore the integrity of the sequence from subcloned YACs was not apparently altered, also reflecting the fidelity of the original YAC clones.

Yeast contamination

There is no convenient method for isolation of YAC DNA which is entirely free of yeast chromosomal DNA. We found that a second PFGE step significantly reduces the amount of contaminating yeast DNA. A third step might decrease it further, although an associated loss in yield of YAC DNA would result. Table 1 summarizes the yeast contamination detected in M13 libraries from three different YAC clones. Library A was prepared from DNA purified by a single electrophoresis whereas libraries B and C were prepared using DNA purified by two successive electrophoresis runs. To detect sequences deriving from yeast DNA sequences were compared with the GenBank database. The number of sequences analyzed for each library A, B and C was 686, 952 and 1625 respectively. A major reduction in the total number of yeast sequences identified was achieved when the second preparative gel was employed. Library A contained 28% identifiable yeast sequences while B and C contained 9 and 5.4% respectively. Significant reduction in the amount of yeast chromosome III, 2 μC circle plasmid DNA, ribosomal DNA and 'Y' repetitive element DNA was achieved. Note, however, that the actual amount of yeast sequence present will be higher than this fraction because the entire yeast genome sequence is not yet complete and available in the public databases, though it is difficult to determine the precise level of contamination.

Table 1. Yeast DNA contamination in three M13 sublibraries of YAC DNA detected by sequence comparison to the public databases

YAC	A	B	C
Number sequenced	686	952	1625
% yeast chromosome III	11.5	3.5	0.9
% 2 μ plasmid	4.0	0.4	0.7
% rDNA	1.6	0.3	0.5
% Y repeat	1.5	0.0	0.1
% other yeast	9.4	4.8	3.2
% total yeast	28	9.4	5.4

Library A was made from DNA that had been purified on one gel only. Libraries B and C were made from DNA that had been purified on two gels.

To obtain a more satisfactory estimate of the extent of extraneous yeast sequences in any YAC or yeast chromosomal DNA band excised from a gel we prepared a sub-library from twice purified yeast chromosome III DNA and sequenced 158 subclones at random. These were compared to the yeast sequences in the public databases, which included the entire chromosome III sequence (5), to determine the percentage of sequenced subclones that were specific to chromosome III.

Eighty percent were yeast chromosome III, 2% were other yeast sequences and 18% gave no match. This result indicates that the true level of contamination expected is about 20%.

Since this was a sequencing project, we followed as closely as possible the steps we usually employ by picking and sequencing the M13 clones directly without pre-screening for contaminating sequences. However, it would be perfectly feasible to perform filter hybridization on the M13 clones using labeled yeast to screen out the yeast sequences prior to sequencing, if desired.

Survey sequencing

Single random reads of cDNA clones have been used effectively to detect database similarities, and we have found that random reads of *C.elegans* genomic DNA can be similarly used. Further analysis of the Y53B1 subclones showed that sequencing to 0.5-fold coverage, after the removal of known yeast, identified 43% of the genes in this region (data not shown). 1194 cDNA sequences have previously been placed on the physical map of *C.elegans* (20). With 0.6-fold coverage of Y50B11, nine of the 14 cDNAS which were mapped to this YAC were identified. Thus, by applying this technology, it is possible to survey a region and its genes rapidly, by sequencing a limited number of subclones from random libraries of YACs or cosmids. Such advanced information might not only reveal interesting genes but would provide valuable insight into the nature of such regions, well before they were subjected to complete sequencing, at a fraction of the cost.

Table 2. Survey sequencing of five *C.elegans* cosmids

		No. assembled[a]	No. genes hit/nine total[b]	No. genes hit (cumulative)[c]
C29E4	1	91	9	9
	2	86	8	9
	3	83	9	9
	4	83	9	9
	5	77	8	9
	6	83	9	9
C50C3	1	91	8	8
	2	88	8	8
	3	88	9	9
	4	94	9	9
	5	92	9	9
	6	91	9	9
F54F2	1	98	8	8
	2	98	8	9
	3	100	9	9
	4	98	9	9
	5	99	9	9
	6	98	7	9
R05D3	1	89	7	7
	2	92	7	8
	3	87	8	8
	4	98	9	9
	5	93	8	9
	6	97	8	9
ZC21	1	98	7	7
	2	100	8	8
	3	98	7	9
	4	98	7	9
	5	90	8	9
	6	95	6	9

600 sequences were analyzed for each cosmid. Each sequential batch of 100 sequences, 1–6, equivalent to ~1 × coverage, were assembled into the cosmid database.

[a]The actual number of each batch that assembled (some failed to assemble and were therefore not included in the analysis).

[b]The number of different known or predicted genes in the cosmid that were detected by sequence homology for each assembled batch.

[c]The cumulative number of different genes hit.

The total number of genes present in each cosmid is nine.

The analysis of YAC Y50B11 described is of particular interest when considering the depth of coverage required for surveying a YAC. At 0.6-fold coverage of a 350 kb YAC there can be unsampled regions of up to 7 kb which will therefore not be represented in such a survey approach. Therefore, to demonstrate the potential of surveying a genomic clone at about 1-fold coverage we used the data previously generated in sequencing five *C.elegans* cosmids. Table 2 shows the number of predicted or known genes hit (and the cumulative number hit) with each sequential batch of 100 sequences (with an overlap of at least 50 bases) that were originally entered into the cosmid assembly database. The second column gives the actual number of sequences that assembled into the finished cosmid contig and that were used in the homology search. In general each batch of sequences (equivalent to about 1-fold coverage) sampled all or most of the genes represented. In practice, such an approach would be limited to sampling proteins that are present in public databases and this is likely to be only about half those found by gene analysis of the entire sequence. Potentially, gene finding could be applied to individual reads, but we have as yet not explored this avenue to increase the fraction of genes identified. However sequencing to 1-fold coverage does represent a relatively rapid and simple method to survey the nature of the DNA and the gene content of a cosmid or YAC clone.

Conclusion

We have shown that YAC clones can be routinely subcloned to produce M13 libraries that are complex enough and of sufficient quality for use in shotgun sequence analysis. The sublibraries appear to exhibit similar properties to those prepared from cosmid DNA. Our analysis suggests that, as with the cosmid sublibraries, some regions are better represented in M13 than others and this is generally related to the type of sequence in that region. For instance, inverted repeats do not appear to clone in M13 and this will lead to gaps in otherwise contiguous sequence.

We are currently testing the feasibility of applying these methods to obtain the complete sequence of a YAC clone from M13 subclones.

There is a definite need for methods to create small subclone libraries from YACs. Irrespective of the final application the production of such libraries from relatively small amounts of relatively pure YAC DNA may represent a key starting point to the success of many mapping and sequencing strategies.

ACKNOWLEDGEMENTS

We thank all those involved in the preparing and sequencing of DNA and editing of the sequence data. Support for this work was provided by grant HG00956 from the National Institutes of Health and funds from Washington University.

REFERENCES

1 Gardner, R.C., Howarth, A.J., Hahn, P., Brown-Luedi, M., Shepherd, R.J. and Messing, J. (1981) *Nucleic Acids Res.* **9,** 2871–2888.
2 Anderson, S. (1981) *Nucleic Acids Res.* **9**, 3015.
3 Deninger, P.L. (1983) *Anal Biochem.* **129**, 216.
4 Messing, J. and Bankier, A.T. (1989) *Nucleic Acids Sequencing: A Practical Approach.* IRL Press, Oxford. 1–36.
5 Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P.G., Benit, P. *et al.* (1992) *Nature.* **357**, 38–46.
6 Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J., *et al.* (1994) *Nature.* **368**, 32–38.
7 Bankier, A.T., Weston, K. M. and Barrell, B.G. (1987) *Methods Enzymol.* **155**, 51–93.
8 Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough,R and Waterston, R. (1992) *Nature.* **356**, 37–41.
9 Burke, D.T., Carle, G.F. and Olson, M.V. (1987) *Science.* **236**, 806–812.
10 Burke, D.T. and Olson, M.V. (1991) *Methods Enzymol.* **194**, 251–270.
11 Coulson, A., Waterston, R., Kliff, J., Sulston, J. and Kohara, Y. (1988) *Nature.* **335**, 184–186.
12 Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J and Waterston, R.(1991) *Bioessays* **13**, 413–417.
13 Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7821–7825.
14 Chu, G., Vollrath, D. and Davis, R. (1986) *Science*, **234**, 1582.
15 Huxley, C., Hagino, Y., Schlessinger, D. and Olson, M.V. (1991) *Genomics* **9**, 742–750.
16 Schriefer, L., Gebauer, B.K., Qui, L.Q.Q., Waterston, R.H and Wilson, R.K. (1990) *Nucleic Acids Res.* **18**, 7455–7456.
17 Craxton, M. (1991) *Methods: A Companion to Methods Enzymol.* **3**, 20–26.
18 Halloran,N., Du, Z and Wilson, R.K. (1992) *Methods in Molecular Biology* Vol 10, 297–316
19 Alschul, S.F., Gidh, W., Miller, W., Myers, E.W and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
20 Waterston, R.H., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., Hawkins,T., Wilson, R., Berks, M., Du, Z., Thomas,K., Thierry-Mieg, J and Sulston J. (1992) *Nature Genet.* **1**, 114–123.