3842–3849 Nucleic Acids Research, 1995, Vol. 23, No. 19

Fine-mapping of shotgun template-libraries; an efficient strategy for the systematic sequencing of genomic DNA

Patrik Scholler, Achim E. Karger[1], Sebastian Meier-Ewert[2], Hans Lehrach[2], Hajo Delius[1] and Jörg D. Hoheisel*

Molecular-Genetic Genome Analysis Group, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany; [1]DNA-Sequencing Group, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 560, D-69120 Heidelberg, Germany and [2]Max-Planck-Institut für Molekulare Genetik, Ihnestraße 73, D-14195 Berlin, Germany

Received July 13, 1995; Revised and Accepted August 31, 1995

ABSTRACT

To test the effectiveness of ordering shotgun DNA-templates prior to sequence analysis, the 450 kb left arm of yeast chromosome XII was randomly sub-cloned into a phagemid vector. Clones were ordered by hybridisation to an average map density of one new insert every 125 bp and are currently used for sequencing the chromosomal fragment. An 11.5 kb overlap between the template map and a DNA fragment that had been sequenced earlier allowed an independent evaluation of the strategy's effectiveness. To this end, clones were selected from the map and tag-sequenced from either end, thus comparing the map position with the actual location within the 11.5 kb. Of 65 selected clones, taken mostly at random from a total of 423, 58 mapped on average about a quarter of a clone length around their predicted position, with the other seven being between 0.6 and 1.5 clone length off. 75–86 sequencing reactions on clones selected from the map would have been sufficient for completely sequencing both strands of the 11.5 kb fragment. The results demonstrate the efficacy of such template sorting, considerably assisting sequencing at relatively little cost on the mapping level.

INTRODUCTION

DNA sequencing is progressing at an ever increasing pace, with even large genomes being in the reach of analysis. In most projects, modifications of the enzymatic primer extension method (1) are utilised due to its compatibility to automated systems. For such sequencing, usually cosmid or bacteriophage P1 clones are sub-cloned as small, random fragments for a shotgun strategy or primer walking techniques are employed in an ordered approach. The former method only needs a single primer or two in case of double-strand sequencing. Due to the random selection procedure, however, it produces a relatively large amount of redundant sequence information. In addition, a change of strategy is frequently required to fill remaining gaps. The primer walking technique avoids the redundancy and gap filling problems, yet, a large number of (untested) primers are needed.

Here we describe an approach which combines the advantages of both methods, sequencing in an ordered manner and using a single primer system. The intention is to achieve significant reductions in the amount of sequencing work by putting a comparatively small extra effort into ordering the templates. The technique is applied to most of the 450 kb left arm of chromosome XII of *Saccharomyces cerevisiae* as part of the European Yeast Sequencing Project. From a minimal cosmid set (2) random shotgun sub-libraries were made in a phagemid vector. The clones were ordered by probe hybridisations and subsequently distributed for sequence analysis. The assembled template map should reduce the redundancy of sequencing to a defined minimum, which can be varied locally according to the quality of the results obtained; a higher clone coverage can be used for problematic regions. Moreover, a minimal amount of sequence overlap will be sufficient to allow contigs of sequences to be constructed. Also, a probe-map is produced as a by-product which serves the purposes of a restriction map. Furthermore, the data obtained from the experiments not only provide fingerprint information for the establishment of the template map but, using selected probe sequences, also add structural information on the DNA, such as the position and kind of repeats, for example. Thus, the DNA segment could be assessed before further work-intensive and costly analyses are being carried out. In this manuscript, data are presented that demonstrate the accuracy and effectiveness of the sorting mechanism by comparing part of the assembled template map with sequencing results which were obtained independently.

MATERIALS AND METHODS

Unless stated otherwise, all procedures were performed as described in Sambrook *et al.* (3).

* To whom correspondence should be addressed

Generation of shotgun libraries

DNA of the minimal coverage cosmid set of the left arm of yeast chromosome XII (2) was isolated by a modified alkaline lysis protocol (4). Whenever possible, the insert-DNA was separated from the vector portion by an *Sfi*I digest and a subsequent gel electrophoresis in low-melting-point agarose. Released from the gel by an agarose treatment (Boehringer Mannheim), the insert-DNA was then sonicated for 10 s in a Branson Sonifier 250. Agarose slices containing fragments of 0.8–1.3 kb were isolated from a gel after a second electrophoresis. The agarose blocks were equilibrated in a large volume of water for ~2 h. Placed on glass-wool in self-made spin columns, the DNA was released from the gel by a brief spin in a bench-top centrifuge. Excess water was evaporated before the fragments were blunt-ended with T4 DNA-polymerase following the protocol of Deininger (5).

Plasmid pTZ18R (6) was cut with *Sma*I and dephosphorylated. For each shotgun sublibrary, ~15 ng/µl cosmid DNA were ligated overnight at 15°C to the same mass of vector in 10 µl of 25 mM Tris-HCl, pH 7.5, 10 mM dithiothreitol, 100 mM NaCl, 7 mM $MgCl_2$, 1 mM ATP and 0.25 U T4 DNA-ligase (Amersham). *Escherichia coli* DH5αF'IQ (Gibco-BRL) was transfected with the DNA by electroporation. Bacteria growth was on selection plates containing 100 µg/ml ampicillin and X-Gal. For each cosmid, 1536 individual, white colonies were picked into 384 well microtitre dishes (Genetix), duplicated and stored frozen.

Filter gridding of the libraries

High-density filter arrays of the plasmid libraries were generated as described in detail earlier (7). Clone material was spotted by 384 pin replicators onto Nylon membranes of 22 × 22 cm using a robotic device. Each filter contained 9216 clones, spotted in duplicate for unequivocal detection; with each shotgun library consisting of 1536 clones, a single filter contained six such libraries. Colony growth was on agar plates, after which the DNA was bound *in situ*. For oligomer hybridisations, PCR-products of the individual plasmid inserts were produced in 384 well plates (8; see also next paragraph), spotted onto filters as above and were directly attached to the surface following the manufacturer's instructions.

Clone pool hybridisations

Plasmid inserts were amplified in sealed 384 well plates submerged in the Autogene II waterbath cycler (Grant) using primers which bind to vector sequences flanking the *Sma*I cloning site of pTZ18R: primer I, CGCCAGGGTTTTCCCAGTCACGAC; primer II, CAGGAAACAGCTATGACCATGATTACGAA. Per reaction, 24 µl of 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM of each nucleotide tri-phosphate, 1 µM of each primer and 1 U *Taq* polymerase were inoculated with minute amounts of *Escherichia coli* cells, transferred from the grown cultures by a 384 pin replicator (Genetix). Amplification was done by 32 cycles of 3 min at 95°C and 5 min at 68°C, followed by a final incubation at 72°C for 15 min. For hybridisation, pools made from the PCR-products were used. Since 15 sub-libraries were worked at and each pool contained DNA of each other library for practical reasons, a maximum of eight individual PCR-products were pooled, although higher numbers are possible. The primer portions of the pooled insert-DNAs were cut off with a *Kpn*I + *Xba*I double digest and removed by a very brief gel electrophoresis prior to the labelling reaction. Radioactive labelling by random hexamer priming (9) directly in the gel slice containing the pooled insert-DNAs and the subsequent probe hybridisation were carried out as described in detail earlier (2).

Oligomer hybridisations

Decamer oligonucleotides bearing a specific core sequence of eight nucleotides plus one unspecific base at either end were labelled at the 5′ terminus with [γ-^{32}P]ATP and T4 polynucleotide kinase (7). Hybridisation was carried out in 10 ml 600 mM NaCl, 60 mM Na-citrate, 7.2% (v/v) Na-sarkosyl with 4 nM probe at 4°C overnight using 300 ml bottles in a Biometra OV1 rotator. The filters were washed in boxes using 1 l of the same buffer at 10°C for 15–30 min. Signal detection was by autoradiography. Motif oligomers were usually 16 nucleotides long; they were treated similar to the above protocol, but only 0.6 nM probe was used for hybridisation and washing was carried out at temperatures between 20 and 37°C.

Tag-sequencing of the insert ends

Double-stranded plasmid-DNA was isolated by alkaline lysis (4). Sequencing reactions at the insert ends were carried out on 0.6 µg DNA using the PRISM™ dye terminator cycle sequencing kit of Perkin Elmer according to the manufacturer's instructions and run on an ABI Model 373 automated DNA sequencer.

RESULTS

Hybridisation fingerprinting

For mapping the 450 kb left arm of yeast chromosome XII, a minimal set of 15 cosmids had been selected which represent the entire fragment (2). From each such cosmid a shotgun library of 1536 clones was generated in the phagemid vector pTZ18R, the typical insert size being ~1 kb. Altogether, 23 040 individual bacterial colonies were picked, arrayed in microtitre dishes and grown as high-density filter arrays for hybridisation analyses. Examination of the insert size of the sub-library relevant to this work revealed an average of 915 bp (±205 bp) as determined on a set of 65 clones; the largest clone was 1642 bp, the smallest 508 bp.

The information necessary for ordering the shotgun clones was obtained by probe hybridisations to filters presenting the libraries as ordered arrays of DNA (e.g. Fig. 1). Since a single filter contained 9216 clones (the equivalent of 6 sub-libraries), all shotgun clones fitted on three filters. Identical filter copies were generated for parallel probe hybridisation (7). Typically, a filter could be used for >60 subsequent hybridisation experiments. While not adequate for complex pool or oligomer hybridisations at later stages, hybridisation with unique probes still yielded meaningful data.

Initially, restriction fragments of the original cosmids were hybridised to the filters in order to coarsely pre-sort the shotgun clones into smaller groups. These experiments also identified regions shared between libraries made from overlapping cosmids. Detailed mapping information was then obtained by hybridisations with oligomers and clones picked from the libraries. Oligomers had the advantage that most of them hybridised to more than one of the shotgun libraries thus producing more information per experiment then obtained by a hybridisation of a unique DNA molecule. However, since the position of their binding sites cannot be predicted, redundant information is also generated. Hybridising

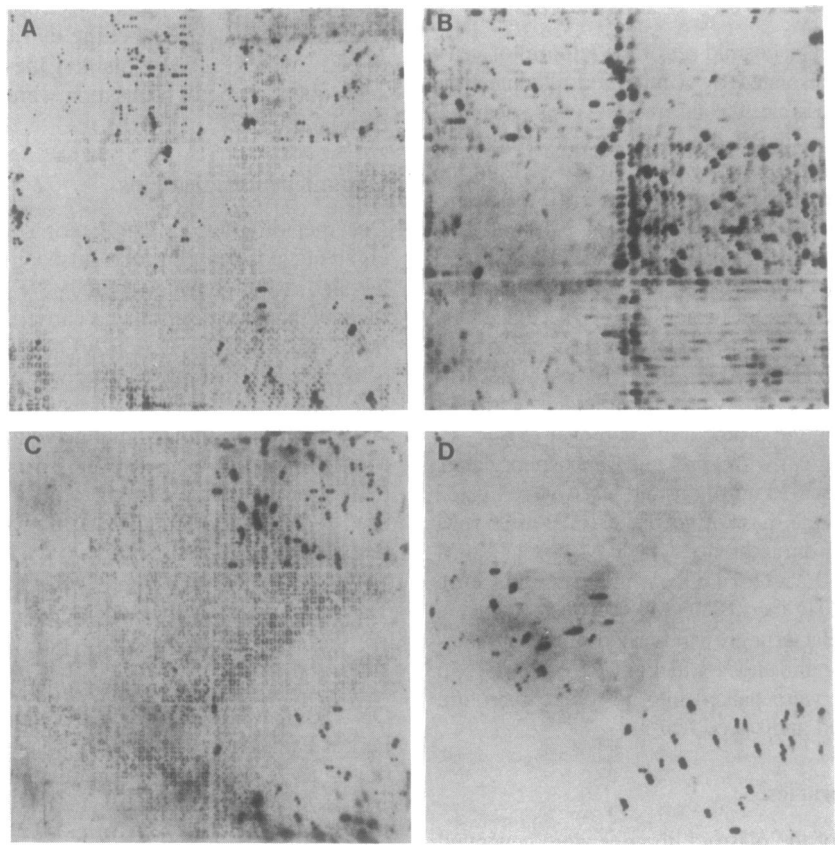

Figure 1. Hybridisation fingerprinting. Each sixth of a filter represents the 1536 shotgun subclones prepared from a cosmid. Positive hybridisation produces two signals because the clones were spotted in duplicate; the angle identifies the microtitre dish in which the relevant clone is stored. (**A**) Hybridisation of an insert-pool. (**B**) Hybridisation of the decamer oligonucleotide NGAAGCCCCN to a filter grid bearing PCR-products of the inserts. (**C**) Hybridisation of a restriction fragment from within an overlap of two neighbouring cosmids; by such means, the respective end of each sub-clone map can be defined. (**D**) Hybridisation of a structure-specific oligomer.

library clones back to the library had the advantage that clones could be selected which were not yet assigned to any contig of the developing clone map, hence falling into gaps. By definition, they produced new, non-redundant information. Rather than hybridising the clones individually, clone pools were actually used. The pools contained one clone from each second sub-library. By this strategy, the higher efficiency of hybridising the clones in pools was combined with the simple data assessment that results from the hybridisations of a unique probe to each sub-library. To avoid cross-hybridisation from the plasmids' vector portion, the pools were made from individually PCR-amplified clone inserts, generated in 384 well microtitre dishes. Figure 2 shows an exemplary gel made from such PCR-products. Some 85% of the inserts could be amplified using the conditions described in the Material and Methods section. Preliminary analyses on some of the remaining clones by a short gel electrophoresis of circular plasmid-DNA indicated that in most cases they contain plasmids lacking sizeable inserts.

Map generation

The hybridisation results were scored manually from autoradiographs (e.g. Fig. 1). For the clone pool hybridisations, a strategy of 'sampling without replacement' was applied which had been used successfully in the cosmid/P1-mapping of the genome of *Schizosaccharomyces pombe* (10). Probes were picked at random from the ever decreasing number of library clones that were not positive in any prior oligomer or pool hybridisation. The process was repeated until all clones with an insert that could be PCR-amplified were hit at least once. By this strategy, the probes, although anonymous, are relatively evenly spaced throughout the genome. The data were analysed using a purpose-written software package (10,11). Briefly, the sorting algorithm calculates a distance between each possible pair of probes from the percentage of clones hybridising to one of the probes but not to both. Unrelated probes, for example, have no clones in common and therefore the percentage equals 1 (the number of clones not shared between the probes is identical to the total number of identified clones), meaning that the probes are at least one clone length apart. For identical probes, as the other extreme, all identified clones hybridise with both probes, so the distance value is 0. Any intermediate case produces a value between 0 and 1. From such distances, the shortest possible linear succession of all probes was calculated by a simulated annealing algorithm and the probes were ordered in contigs. Only subsequently the clones were fitted to the probe map in a procedure similar to STS mapping (10).

In Figure 3 the portion of the resulting plasmid map is shown which overlaps with an 11.5 kb segment of a cosmid that had been

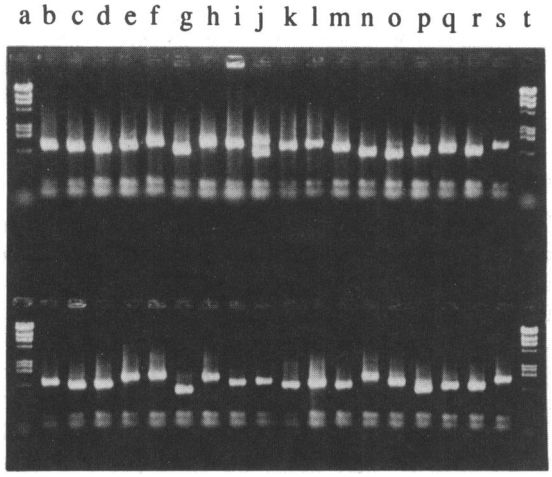

Figure 2. Agarose gel of the PCR-products of shotgun clone inserts. One-fifth of a reaction was loaded to each lane. Lanes (a) and (t) show marker molecules (λ *Hin*dIII plus pUC19 *Fsp*I). In lane j, upper row, two bands are visible indicating the presence of two clones in a single storage well.

sequenced earlier (12). Due to the sub-division of the 450 kb fragment into cosmid-sized and smaller segments, the oligomer signals could be considered locally as being unique and, thus, treated analogous to the pool data. 45 probe hybridisations were sufficient for a contiguous coverage of the 11.5 kb region. However, because of the nature of the probes, being clone pools or oligonucleotides, all but two hybridisation experiments, the two were done for gap closure, did produce mapping data not only for this 11.5 kb region but mainly (7 out of 8 inserts in each pool) in other parts of the 450 kb chromosomal fragment. Therefore, most information obtained from 43 hybridisations contributed to the mapping of other regions and, thus, only a small portion of the actual workload can be assigned to the mapping of this fragment. Based on such considerations, the virtual workload necessary for the ordering of the 11.5 kb adds up to <10 hybridisation experiments, including the two done for gap closure.

In the experiment, the results were scored only once, directly after each individual experiment and were not re-examined after being run through the ordering algorithm; some 180 presumably false positive and negative signals can be seen in the map (Fig. 3). Nevertheless, the sorting algorithm proved to be robust enough to deal with such a background. From similar results from ordering cosmid clones by hybridisation mapping (2), it can be estimated that this error rate could have been reduced by about 80% upon re-examination of the data, thereby improving the map quality. However, a manual review of the autoradiographs was impractical due to the sheer amount of data.

Map evaluation

For an evaluation of the accuracy of the plasmid map, the mapping data of the 11.5 kb portion displayed in Figure 3 were compared to the known sequence of the fragment. This genomic sequence is part of a 36 849 bp sized cosmid-insert which had been sequenced completely by primer walking on *Sau*3A sub-clones in plasmid pBC-KS(+) bearing inserts of 7–15 kb (12). A set of 70 shotgun clones was picked for analysis from the total of 423 covering the area. Apart from a sequencing set selected as described below, clones were picked at random from the entire region (Fig. 3). In some cases direct neighbours with identical hybridisation fingerprints were analysed for comparison.

A few clones were found to be the product of co-ligation events, containing two unrelated DNA-fragments. In all cases, one insert portion mapped inside, the other outside the 11.5 kb overlap (data not presented). The chimeric nature could therefore not be identified from the partial map covering the 11.5 kb fragment only. In a complete map made from an entire cosmid, a chimeric clone would be a unique connection between two distant probes. Such

Table 1.

sequence ranking	map ranking	clone name	sequence start [bp]	sequence end [bp]	predicted map position [bp]
1	2	8 L 8	0	752	585
2	6	7 H12	0	776	637
3	5	8 A13	0	803	566
4	4	5 D 6	30	787	502
5	3	7 N15	0	843	667
6	1	7 I 10	0	857	504
7	9	5 P22	0	1115	1100
8	7	5 N20	458	1279	827
9	8	5 D 8	537	1398	882
10	10	5 M12	596	1396	1290
11	12	8 L14	865	1554	1725
12	11	6 G24	961	1733	1312
13	19	8 K12	722	2364	2840
14	14	6 E15	1141	2100	2060
15	13	8 C23	1158	2517	1833
16	15	6 N 4	1544	2469	2160
17	17	7 P13	1817	2757	2540
18	16	6 I 3	1861	2950	2513
19	18	7 C19	2137	2937	2675
20	20	5 N 6	2411	3526	2866
21	22	5 D10	3026	3836	3410
22	21	5 G10	2864	4071	3083
23	25	5 O13	3199	4072	3763
24	23	8 G19	3270	4183	3573
25	24	8 F 9	3329	4131	3627
26	26	6 O21	3595	4830	3818
27	32	6 P19	3894	4676	4307
28	27	7 D24	3975	4958	4062
29	29	5 H22	4041	5079	4116
30	28	5 F 2	4156	4976	4089
31	31	6 P15	4031	5227	4352
32	30	7 B12	4168	5114	4144
33	33	5 P15	4233	5204	5367
34	34	6 K 4	4418	5191	5449
35	40	8 N20	4680	5755	5602
36	38	6 K11	4983	5593	5911
37	39	5 I23	5140	6036	5938
38	35	7 K 6	5162	6031	5530
39	37	7 O 3	5223	5986	5640
40	41	8 I 11	5229	6230	6163
41	36	7 I 14	5158	6210	5557
42	43	5 P 7	5665	6473	6835
43	42	8 H 3	6138	6987	6781
44	44	8 M24	6188	7048	6862
45	47	6 G10	6301	6935	6559
46	45	5 H 8	6301	7336	7080
47	46	7 P16	6731	7762	7324
48	51	5 M 1	6814	7846	8060
49	49	7 G 3	6772	7935	7651
50	48	5 K17	7134	7874	7570
51	56	5 N18	7304	8237	8494
52	53	5 F 8	7494	8618	8275
53	54	8 P15	7526	8670	8303
54	52	8 I 10	7785	8597	8250
55	55	6 J 21	7839	8810	8436
56	57	8 N 6	8099	8983	8629
57	50	5 G24	8365	9444	8031
58	58	6 I 20	8410	9508	8629
59	59	8 N 1	8852	9868	9555
60	60	7 B 2	9183	10232	8884
61	61	8 O 8	9206	10341	9239
62	62	8 B21	9808	10690	10106
63	63	5 G23	10071	11138	10179
64	64	5 L 9	10544	11567	10234
65	65	5 A14	10443	11451	10641

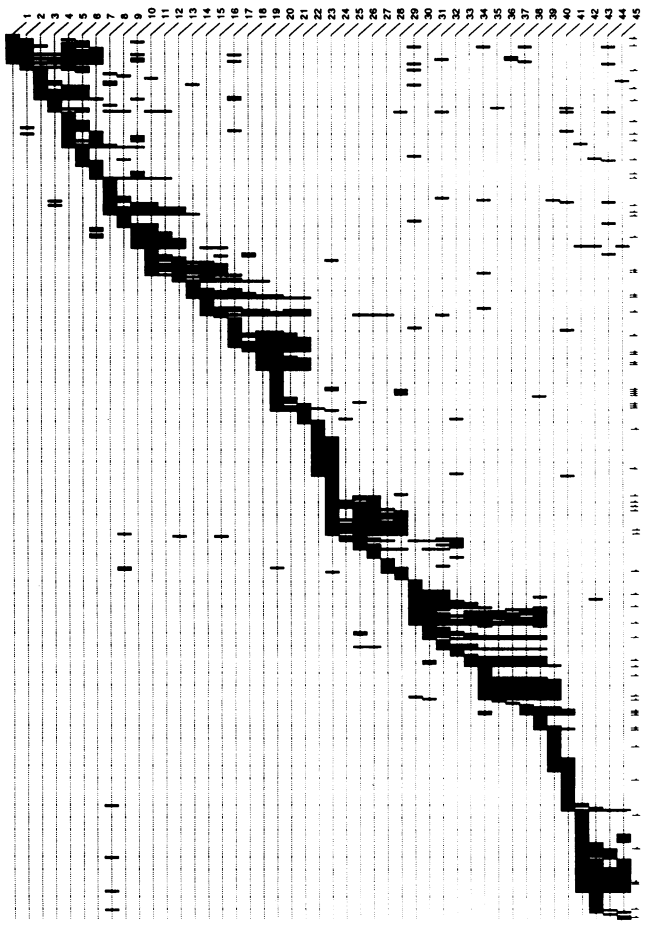

Figure 3. Physical plasmid map covering 11.5 kb. The results of the hybridisation experiments are presented in a two-dimensional matrix: probes correspond to columns (vertical lines 1–45) and clones to rows; for the scale of the figure, the complete list of clone names along the left margin is not printed. Every hybridisation event is recorded by a black bar. The positions of the clones picked for map-evaluation are indicated by marks on the right margin.

data and the respective clone would be ignored in the analysis until confirmed by at least one more clone (10), which is extremely unlikely to happen in case of a random co-ligation event. As a result, chimeric clones are not considered for subsequent analyses such as sequencing.

Each insert end of the selected clones was sequenced by a single reaction, whereby they could be positioned accurately in the known sequence of the 11.5 kb fragment. The congruence between map and actual order is high but varies depending on the algorithm which is used for placing the clones (Fig. 4). The highest similarity was found with the clone order based on the ranking in the list of clones (Fig. 3) that results from the fitting of the clones to the given probe map. Other clone orders, determined by the first, last (not shown) or central probe by which a clone is hit, gave less accurate results (Fig. 4).

Although the distances between the probes are not as constant as the artificially uniform intervals depicted in Figure 3 suggest, still a prediction of the sequence position of the clones' centres was made taking into account the ranking of each clone as well as the insert length as estimated from the number of hybridising probes (Table 1). 58 of the 65 clones mapped on average 230 bp around their actual location. Only a minority of seven clones differed substantially by 645–1297 bp. With regard to the eventual application of the map to an ordered sequencing process, however, the exact position of an individual clone is relatively unimportant as long as the ordering procedure yields a good spatial representation of the DNA-fragment by the shotgun sub-clones.

Selection of a minimal clone set for sequencing

Based on the hybridisation map (Fig. 3), clones for a complete sequence analysis of the 11.5 kb fragment were selected, assuming a reproducible read-length of 500 bp. Since the clone ranking was the best fitting clone order, each 16th clone was chosen from the list starting from the top of the map. On average their spacing is equivalent to ~430 bp, which is nearly 90% of the assumed read-length. Only if a selected clone was followed by a run of more than eight clones with identical hybridisation patterns, these latter clones were ignored. In addition to this, a clone should not show a (false) positive signal outside the main diagonal of positive hybridisation signals in Figure 3, a requirement that needs to be applied the more stringently the fewer probes hit a clone. There are a few clones in the map which showed hybridisation with only two probes, one located within and one placed outside the main diagonal. Since the placing based on two controversial results has a low probability of being correct, the clones should be avoided.

The set of 25 clones selected by the above procedure would have been sufficient for sequencing 98% of the 11.5 kb fragment on at least one strand, leaving three gaps in the single-strand coverage of together only 242 bp (Fig. 5). 56% of the DNA would have been covered on both strands. An increase of the read-length to 750 bp, which is not unreasonable to achieve with today's automated systems, improves the result to continuous single-strand sequence and 90% double-stranded coverage.

The remaining gaps in the double-strand sequence would have been closed by selecting additional clones which map between the initial set, preferentially picking clones that are hit by 3–4 probes and thus selecting for average sized inserts. An additional 25–36 sequencing reactions of 500 bp on 18 extra clones or 8–10 reactions of 750 bp on five clones respectively, would have been sufficient to determine continuously the 11.5 kb sequence on both strands, apart from the short, terminal regions. The imprecision in the given number of reactions is caused by the fact that the orientation of the inserts in relation to the primer binding sites in the vector is unknown, thus one or two sequencing reactions could be necessary to span the gap. The final sequence obtained from 500 bp reads would represent a 1.65- to 1.85-fold sequence redundancy on either strand. The ~1.9-fold redundancy in case of the 750 bp reads is an artificially high value, since the average distance between the selected clones is too small with only 57% of the read length.

DISCUSSION

The mapping and subsequent sequencing of the left arm of yeast chromosome XII is a pilot-project for testing the approach of fine-mapping shotgun template-libraries prior to an ordered DNA-sequencing. Apart from the mere sequence determination, the project's aims are a refinement of the analysis procedures and an assessment of the technique's performance in a practical application to a relatively large segment of chromosomal DNA. In this manuscript, an evaluation of the accuracy of the produced map and its usefulness to sequencing completely a DNA-fragment was made by comparing actual mapping data with existing sequence information.

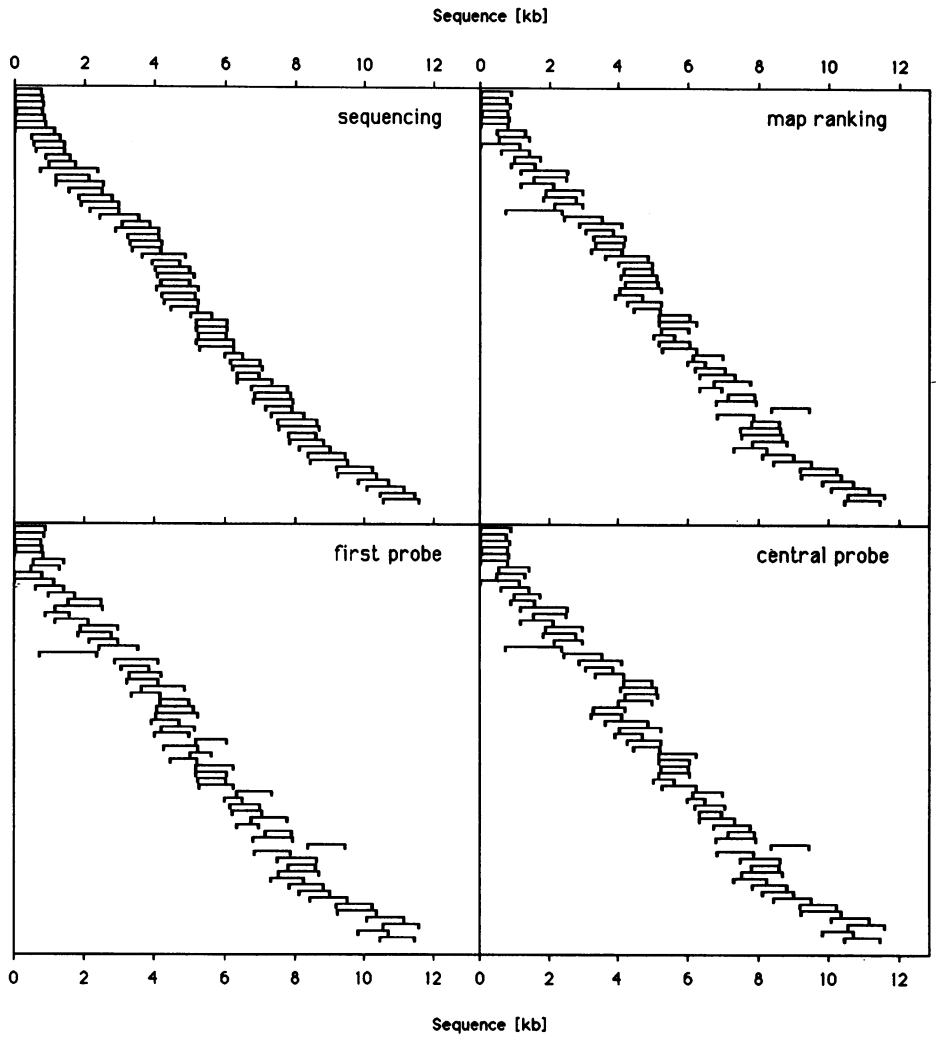

Figure 4. Comparison of map position to actual sequence location of selected clones; the ordering mode is indicated in each panel. The top-left panel shows the exact clone positions within the 11.5 kb fragment as determined by sequencing. Note that no true diagonal placement can be expected, because intentionally many clones had been chosen for analysis that looked similar or even identical from the hybridisation data. The other panels show the clone order as derived from the mapping data. The clone order was determined by the clone ranking in the map or the respective position of the first or the central probe (1–45; see Fig. 3) by which a clone was hit.

The quality of the produced map fulfilled the requirements set by a subsequent sequence determination. Still, improvements can be made on several aspects. A quantitative detection of the hybridisation results adds substantially to the analysis. A normalisation of the signal intensities on the basis of a vector hybridisation, indicating the amount of DNA at each filter position, makes the oligomer hybridisations more reliable, since particularly their signals depend on the quantity of target DNA present at each spot. Thus, even weak signals will be registered as positives when known to occur on spots with little DNA, while false positives caused by a strong unspecific signal can be avoided. Additionally, the frequency of positive clones and thereby the accuracy of the mapping procedure is increased, because a manual analysis is biased towards the strong signals missing less obvious positive clones. For the pool probes, detection of the variations in signal strength produces another form of information on the degree of overlap, thus adding another measure of the distances between probes and clones, on which the mapping algorithm relies upon. In addition, an image analysis system with an electronic record of the signal intensities has the advantage that the raw data of even large projects can be reviewed after an initial round of clone ordering. Due to the assignment of clones to already well established contigs, an analysis of the remaining data allows the removal of incorrect connections at this stage. Also, indifferent signals, such as weak positives, could be identified as being true after re-evaluation. A manual re-examination of mapping data reduced the rate of presumably erroneous scoring events by nearly 80% (2), hence enhancing the map quality notably. For the above reasons, a system for a quantitative analysis of hybridisation results is now in place in the laboratory for future experiments. Using alkaline-phosphatase-labelled DNA, fluorescence signals are produced after hybridisation upon the addition of AttoPhos substrate (13) and recorded by a CCD-camera system.

Another element for improvement is the variation of the insert sizes of the shotgun clones. The relatively wide deviation in the library studied here caused problems in the mapping as well as the

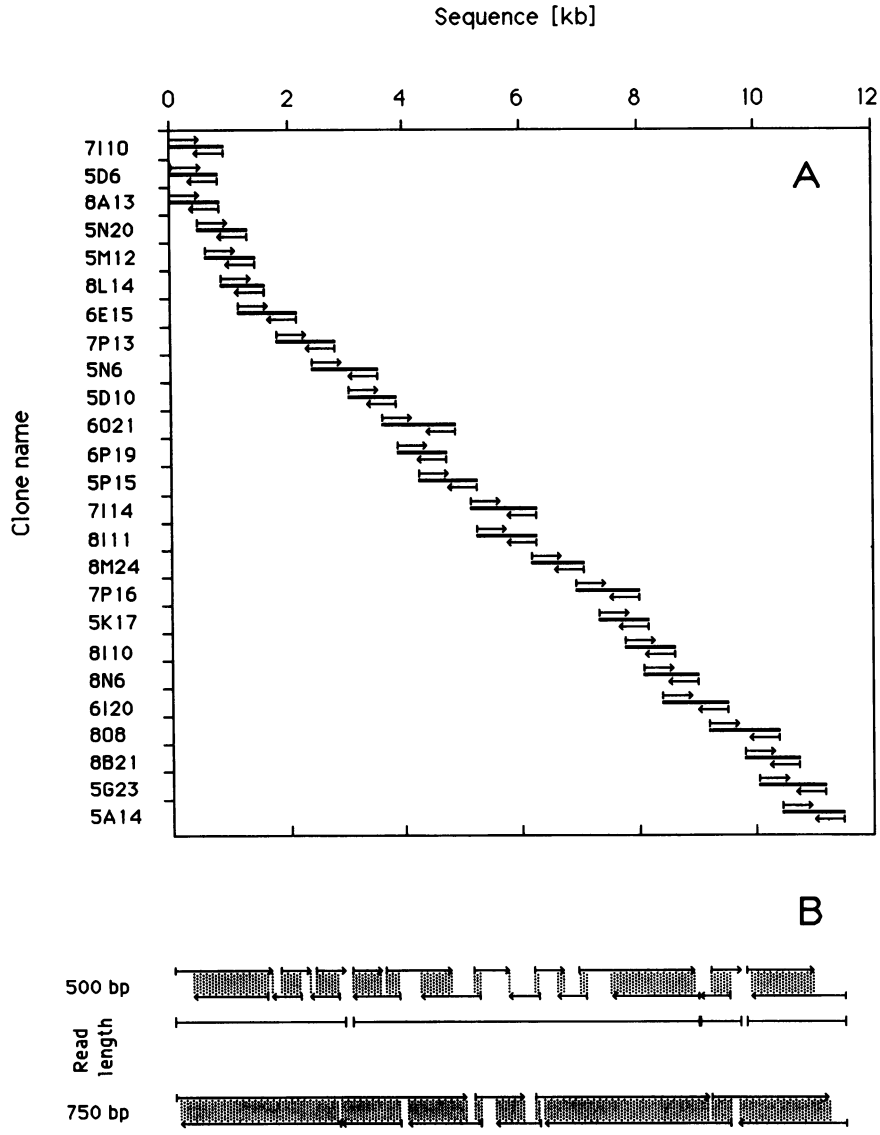

Figure 5. Initial clone set selected for a sequence determination; all distances along the ordinate are to scale. (**A**) Clones were selected from the map as described in the text. Arrows indicate sequencing runs of 500 bp length done on both ends of each clone. The coverage obtained by sequencing either strand and the combined single-strand sequence coverage are shown in (**B**). For a read length of 500 bp, the position and size of the three gaps in the single-strand coverage are shown. Also, the degree of completed double-strand sequence is indicated by hatched areas between the strands.

simulated sequence analysis of the clones; generation of more uniform inserts would raise the efficiency of the method. The distances between the probes would be more homogeneous, in turn allowing a higher accuracy in placing the clones. A higher conformity in length would lead to a reduction in the number of clones that need to be picked for sequencing. To this effect, it also seems to be advantageous to generate a library of two size classes. Using pools of clones of ~1400 bp length as probes, besides the oligomers, would speed up the mapping by reducing the number of probings. Clones of 900 bp, with a variation of no more than 100 bp, would be located in the map, although they might not necessarily span the distance between each pair of probes. Yet, in an initial sequencing round, such as depicted in Figure 5, they could be sequenced completely with reads of 500 bp from either end producing a continuous single-strand sequence coverage of this portion of the map. Only where the short clones would not connect a probe pair, a few additional hybridisations would be necessary to increase the map resolution further. Such a procedure should reduce the overall number of experiments.

In the actual sequencing of the 11.5 kb fragment by the directed method of primer walking on sub-clones (12), 100 reactions of an average length of 380 bp were required to complete a double-strand coverage. This is equivalent to 76 reactions on the basis of a read-length of 500 bp. Starting from the template map, the same fragment could have been sequenced by 75 to 82 reactions, without the need for walking primers which are still a significant cost factor. Shotgun sequencing the fragment, producing an overall ~8-fold redundancy, would have required 184 reactions without the guarantee that all gaps would be closed. The equivalent of 10 hybridisations which were necessary for generating the template

map compares favourably in terms of work and cost with the otherwise necessary sequencing reactions or primer syntheses. It is noteworthy that the value of less than one hybridisation experiment per kilobase was calculated for the analysis of the 450 kb yeast fragment. The efficacy and thus the benefit of template mapping improves considerably with the size of the DNA fragment.

Verification of the sequence on only one strand by using two different chemistries of sequencing reactions, such as T7 DNA polymerase and dye-primers or *Taq* DNA polymerase and dye-terminator conditions (14), rather than by applying a single reaction type to both DNA strands would be advantageous. Such a procedure would reduce the number of reactions further and render the selection and purification of some template-DNA unnecessary. Most efficient would have been a combination of sequence confirmation techniques. Only those parts of the 11.5 kb fragment not being double-strand sequenced after end-sequencing the initial 25 clone selection should have been re-sequenced on one strand only using a second type of chemistry. By such a strategy, a total of 69–72 reactions on 29 clones would be sufficient for a confirmed sequence of the entire fragment.

Preliminary data on shotgun libraries made from two cosmid clones containing human DNA indicate that an analysis of human DNA does not seem much more complex than the experiments with yeast DNA reported here. The oligomer hybridisation is completely unaffected, while the pool probes require a competition reaction prior to hybridisations to the filters in order to suppress hybridisation of repeat sequences. When cosmid maps are generated by means of hybridisation mapping, pre-sorting the relevant shotgun sub-libraries is most likely to be feasible for the proven lack of interfering repeats. Fine-mapping of shotgun libraries made from two additional human cosmid and P1 bacteriophage clones is currently under way.

The mapping of 450 kb in 1 kb fragments can be compared in terms of relative size to the coverage of 225 Mbp in 0.5 Mb YAC clones. Even if this effort seems excessive at first sight, the power of hybridisation mapping is such that it is attractive nevertheless, particularly for large scale projects. The high degree of parallelism in template presentation and probe composition (oligonucleotides and pools) reduces the work necessary to less than one hybridisation experiment per kilobase of sequence. This value is bound to decrease sharply with increasing size of the analysed DNA fragment, because the number of oligomer hybridisations is completely unaffected by size and clone pools larger than those in the present analysis can be used. Although the final number of hybridisations is still substantial for large DNA segments, the effort appears more than compensated by the gains during a subsequent sequence analysis, already with the manual procedures described here. Since there still is a huge potential for an automation of hybridisation techniques, as opposed to the existing gel-based sequencing technology and for a multiplication of the throughput by using different fluorescence labels in parallel, for example, further improvement in efficiency can be expected. It is also noteworthy that for potent, future sequencing methodologies, such as *Sequencing by Hybridisation* (for a short review see ref. 15), a 1 kb template map is not only suitable in size and format but indeed a prerequisite for its practical application.

ACKNOWLEDGEMENTS

The excellent technical support by Sandra Schwarz is gratefully acknowledged. This work was funded by a contract with the European Commission for the European Yeast Genome Sequencing Project.

REFERENCES

1 Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
2 Scholler,P., Schwarz,S. and Hoheisel,J.D. (1995) *Yeast*, **11**, 659–666.
3 Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor University Press, Cold Spring Harbor, NY.
4 Pohl,F.M., Thomae,R. and Karst,A. (1982) *Eur. J. Biochem.*, **123**, 141–152.
5 Deininger,P.L. (1983) *Anal. Biochem.*, **129**, 216–223.
6 Mead,D.A., Szczesna-Skorupa,E. and Kemper,B. (1986) *Prot. Eng.*, **1**, 67–74.
7 Hoheisel,J.D., Lennon,G.G., Zehetner,G. and Lehrach,H. (1991) *J. Mol. Biol.*, **220**, 903–914.
8 Meier-Ewert,S., Maier,E., Ahmadi,A.R., Curtis,J. and Lehrach,H. (1993) *Nature*, **361**, 375–376.
9 Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6–13.
10 Hoheisel,J.D., Maier,E., Mott,R., McCarthy,L., Grigoriev,A.V., Schalkwyk,L.C., Nizetic,D., Francis,F. and Lehrach,H. (1993) *Cell*, **73**, 109–120.
11 Mott,R., Grigoriev,A., Maier,E., Hoheisel,J.D. and Lehrach,H. (1993) *Nucleic Acids Res.*, **21**, 1965–1974.
12 Delius,H., submitted
13 Maier,E., Roest-Crollius,H. and Lehrach,H. (1994) *Nucleic Acids Res.*, **22**, 3423–3424.
14 Koop,B.F., Rowan,L., Chen,W.Q., Deshpande,P., Lee,H. and Hood,L. (1993) *BioTechniques*, **14**, 442–447.
15 Hoheisel,J.D. (1994) *Trends Genet.*, **10**, 79–83.