Cell, Vol. 73, 109–120, April 9, 1993, Copyright © 1993 by Cell Press

High Resolution Cosmid and P1 Maps Spanning the 14 Mb Genome of the Fission Yeast S. pombe

Jörg D. Hoheisel, Elmar Maier, Richard Mott, Linda McCarthy, Andrei V. Grigoriev, Leonard C. Schalkwyk, Dean Nizetic, Fiona Francis, and Hans Lehrach
Imperial Cancer Research Fund
Lincoln's Inn Fields
London WC2A 3PX
England

Summary

Gridded on high density filters, a P1 genomic library of 17-fold coverage and a cosmid library of 8 genome equivalents, both made from S. pombe strain 972h⁻, were ordered by hybridizing genetic markers and individual clones from the two libraries. Yeast artificial chromosome (YAC) clones covering the entire genome were used to subdivide the libraries, and hybridization of short oligonucleotides and DNA pools made from randomly selected cosmids provided further mapping information. Restriction digests were generated as an independent confirmation of the clone order. The high resolution clone map was aligned to the genetic map and the physical NotI and YAC maps. The usefulness of the various mapping techniques and cloning procedures could be assessed upon the different data sets.

Introduction

The fission yeast Schizosaccharomyces pombe is a unicellular fungus with a genome of about 14 Mb, which is organized into three chromosomes of 3.5, 4.6, and 5.7 Mb. It has all the basic characteristics of a eukaryotic genome, yet it contains relatively few repetitive sequences. Over 500 nonallelic genetic markers have been identified, with over 250 markers mapped to the chromosomes (Lennon and Lehrach, 1992). Biologically, S. pombe is of interest because it is similar to humans in some respects and is likely to have many functions that are conserved in mammalian cells. It therefore serves as an important biological model system, as in the intense studies carried out on cell division cycle control (e.g., Booher and Beach, 1989; Kinoshita et al., 1990; Enoch and Nurse, 1991; Forsburg and Nurse, 1991). Other biological problems in S. pombe have also received much attention, including signal transduction, recombination, mechanism of mitosis, and control of meiosis and cell differentiation (reviewed by Hayles and Nurse, 1992). Because of this inherent biological interest in S. pombe, the moderate size of its genome, the relative paucity of repeat sequences, and the high density of genetic markers, the analysis of the S. pombe genome was chosen as a test case for the application of various hybridization protocols developed for large-scale high resolution genome mapping.

Hybridization permits the simple and fast identification of homologous DNA and RNA stretches, and large numbers of clones can be analyzed in parallel. One of its main features is the fact that any piece of DNA can be either the probe or the probant, so that each experiment can actually be carried out either way, whichever is more convenient to analyze or produces more information. On this basis, a system has been developed that provides a tool to generate a physical map based on clone libraries made from large genome portions or an entire genome and to assign other mapping data toward these ordered libraries. As stages in this process, a large number of different reference libraries for various organisms, including humans, has been created, and a range of hybridization procedures has been established (Craig et al., 1990; Lehrach et al., 1990; Nizetic et al., 1991; Hoheisel et al., 1991; Ross et al., 1992).

Most large-scale mapping efforts on the cosmid level reported earlier, in particular, the extensive analysis of the Caenorhabditis elegans genome, relied on data from gel fingerprinting techniques (Coulson et al., 1986). Although a strong analogy exists between the types of information generated by gel and hybridization fingerprinting methods, there are large differences with respect to the experimental manipulations involved (Lehrach et al., 1990). In general, hybridization techniques are more work effective, owing to much simpler handling requirements and the analysis of large clone numbers in parallel. Gel fingerprinting, on the other hand, is less sensitive to the occurrence of repeat sequences than clone hybridization. The hybridization of short oligomers combines these features in a single technique.

For the experiments reported here, genomic libraries from the S. pombe strain 972h⁻ were established in yeast artificial chromosome (YAC), bacteriophage P1, and cosmid vectors. Owing to the relatively small size of the S. pombe genome, the mapping analysis was mainly based on single clone hybridizations. Hybridizations with clone pools and oligonucleotides were performed as an additional source of data, assisting in the determination of the clone order as much as the comparison of the results obtained with the different probe types. The parallel use of the three libraries not only contributed substantially toward the resolution of the clone order on all levels but also allows an evaluation of the effectiveness of the three cloning procedures.

Results

Mapping Strategy and Procedure

Following the approach of integral genome mapping (Lehrach et al., 1990), utilizing different libraries in parallel so that information produced on one mapping level can instantly be used to assist the analysis on another level, work was simultaneously started with the YAC and cosmid libraries, the complete map of the former having been reported earlier (Maier et al., 1992). Owing to technical problems, the genomic P1 library only became available at a later stage. DNAs of known genetic markers and randomly

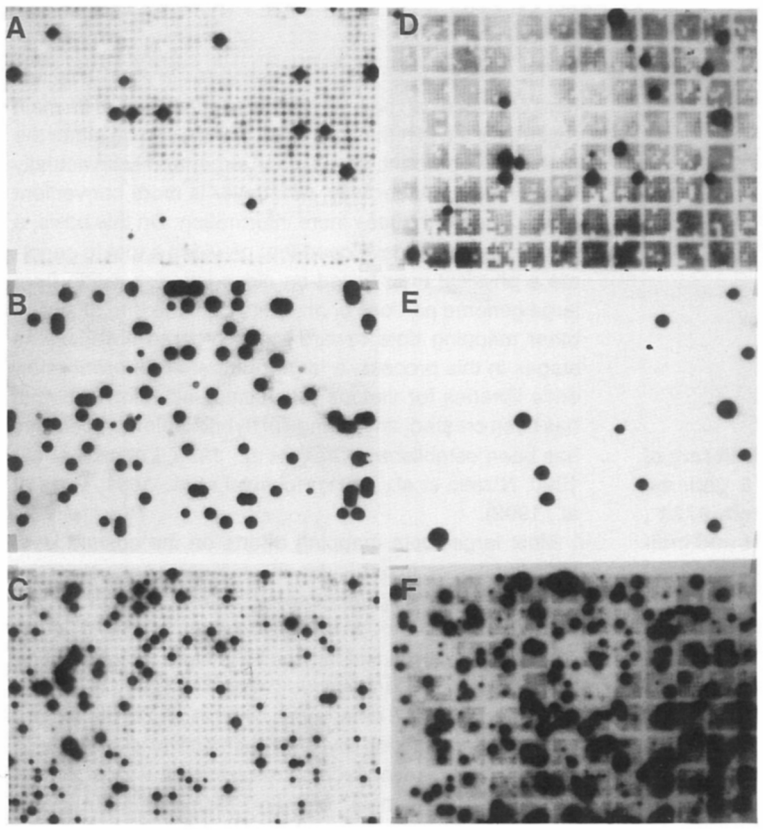


Figure 1. Autoradiographs of Probe Hybridizations onto Cosmid and P1 Filters

Probes made from single cosmid clones (A and D), a pool of 12 cosmids (B), and a genetic marker DNA (E) were hybridized. (C) and (F) show the results obtained with a single YAC clone and a cosmid containing several copies of the rDNA repeat, respectively.

selected cosmid clones were isolated, labeled by random hexamer priming (Feinberg and Vogelstein, 1983), and hybridized to the library filters. Rather than use a short probe made from both insert termini of the cosmids (Hoheisel et al., 1991), thus avoiding any potential cross-hybridization caused by most of the genomic insert DNAs, we chose the technique of labeling the whole cosmid, and thus detecting repeat sequences in the S. pombe genome, for two reasons. First, since the physical map is a step toward the eventual sequence analysis, it should indicate the locations of repeats, so that procedures can be designed for either sequencing through such DNA stretches or avoiding them. Second, the work on S. pombe was carried out as a test case for the mapping of larger genomes, in particular, those of mammals. Therefore, the analysis tools developed should be able to deal with some level of repeated sequences, which are very abundant in larger genomes.

On the YAC level (Maier et al., 1992), cosmid probes were picked at random until the majority of the genome was covered by contigs. Then, YACs representing the already mapped part of the genome were hybridized to the cosmid library to identify the cosmids located in the gaps, which in turn could then be used as probes to close the map. This proved to be an important prerequisite for an efficient ordering process. At the same time, the entire cosmid library of about 23-fold genome coverage was thus subdivided into small portions covered by the respective YAC clones. Five remaining undetected overlaps were bridged by hybridizing rescued YAC ends to the YAC and cosmid filters. The identified cosmids were again used to confirm the YAC–YAC overlaps defined by the rescued probes.

For the cosmid ordering, various hybridization procedures were applied. A strategy of "sampling without replacement" was followed with the single clone hybridizations (Figure 1A). Only clones not hit in any of the earlier marker and cosmid hybridizations were used in the next round of experiments, until 99.5% of a subset of 3072 cosmids, representing about 8 genome equivalents, had been hit at least once. During this phase, the efficiency of the probe selection was significantly increased as compared with a totally random selection by coarsely ordering the as yet unhit cosmids on the basis of the data obtained from the YAC–cosmid hybridizations (Figure 1C). Since the probings were done in batches of 30 to 60 parallel hybridizations, evenly spaced clones could be picked as probes, thus avoiding unnecessary redundancy.

In addition to the single clone hybridizations, cosmids were individually grown and prepared in pools of 12 clones each. Clones were arranged in a two-dimensional pooling scheme, so that each clone was present in two different pools, which allows each hybridization signal to be related to a particular clone (Evans and Lewis, 1989). To reduce background signal, end-specific probe was created by a polymerase extension of two radioactively labeled oligonu-

cleotide primers annealing to the vector directly adjacent to either end of the insert (Figure 1B; Hoheisel et al., 1991). The pool hybridizations were found to be as effective as single clone hybridizations, but hybridization results of a much higher quality were required, considerably increasing the experimental effort to be put into each experiment. Repeat sequences in the probe pools made discerning the relation between the observed signals and the individual clones responsible for them more complicated or even impossible, thus transforming these and all intersecting pools into multilocus probes, rather than a group of unique probes.

Thirty-two oligonucleotides 11–12 nt long were also hybridized. Oligonucleotide hybridization has been proposed as an efficient means of ordering clones by hybridization fingerprinting (Lehrach et al., 1990) and has been demonstrated to work effectively on a small virus genome (Craig et al., 1990). Its major advantages over other hybridization procedures are that the number of probes necessary is independent of the genome size and that the hybridizations are far less affected by repeat sequences in the DNA. Because oligomers assay for the presence of a short sequence, they offer more detailed clone ordering information than hybridizations of cosmids to cosmids can. The sequences of the oligomers used on S. pombe were based on simple tandem di-, tri-, and tetranucleotide repeats or were random with 30%–50% G or C content, avoiding certain features known to be highly frequent in the Escherichia coli DNA, which is always present in the colony spots of the in situ filters. Annealing was performed at 8°C and at high probe concentration, whereby a duplex was formed indiscriminately on all clones. Signal discrimination was achieved subsequently during the filter washing, the conditions having been determined empirically for each oligonucleotide. Owing to their different nature as multilocus probes, the oligonucleotide results were initially not directly included in the analysis but were used to check the probe order obtained by other means.

Contig Assembly

In the analysis, more accurate and longer clone contigs were obtained by ordering the probes first and then fitting the clones to the probe order. A probe–probe comparison based on the hybridization frequencies of a relatively large number of clones has a higher information content than a clone–clone comparison based on the hybridization data of the much smaller number of probes. Using newly designed algorithms and software tools described in detail elsewhere (Mott et al., submitted), the distance between each possible pair of probes was calculated from the percentage of clones hybridizing only to one probe but not to both, formally the same distance used by Cox et al. (1990) for positioning markers on the basis of DNA fragmentation in radiation hybrid cells. The probe order was derived by algorithms using simulated annealing to produce a set of probe tagged sites, named in analogy to the concept of a sequence tagged site (Olson et al., 1989). Contiguous arrays of probes were formed, which were kept separate whenever the distance between two neighboring probes exceeded a given value. Usually a threshold was used that was roughly equivalent to a 2 clone link between a pair of probes. In the final map, however, such a low value was only accepted if supported by additional mapping information, such as the parallel existence of the cosmid and P1 maps. Another algorithm exploited a set of heuristic rules, producing an almost identical order of probes. Since these ordering processes were based on the assumption that each probe is unique, an optional filtering process was implemented so that clones with probable well-to-well contaminations or containing known repeat sequences (e.g., ribosomal DNA [rDNA]), both of which often cause false connections, were detected and excluded from further calculations.

Gap Closure and P1 Mapping

After the sampling without replacement strategy was finished, the cosmids were assembled in some 140 contigs, which is in good agreement with computer simulations (data not shown). These contigs were ordered and oriented, where possible, by taking account of the genetic markers, the cosmid probe tagged site order on the YAC map (about 1 probe per 100 kb; Maier et al., 1992), and the YAC–cosmid hybridizations. Potential contig links, indicated, for instance, by single clone overlaps between neighboring contigs or by the pool or oligomer data, and suspected undetected overlaps due to the lack of a probe on the clone overlaps, were checked by hybridizing either clones thought to bridge the relevant gap or terminal clones of the 2 contigs, quickly reducing the number of contigs to 75 and eventually to 72.

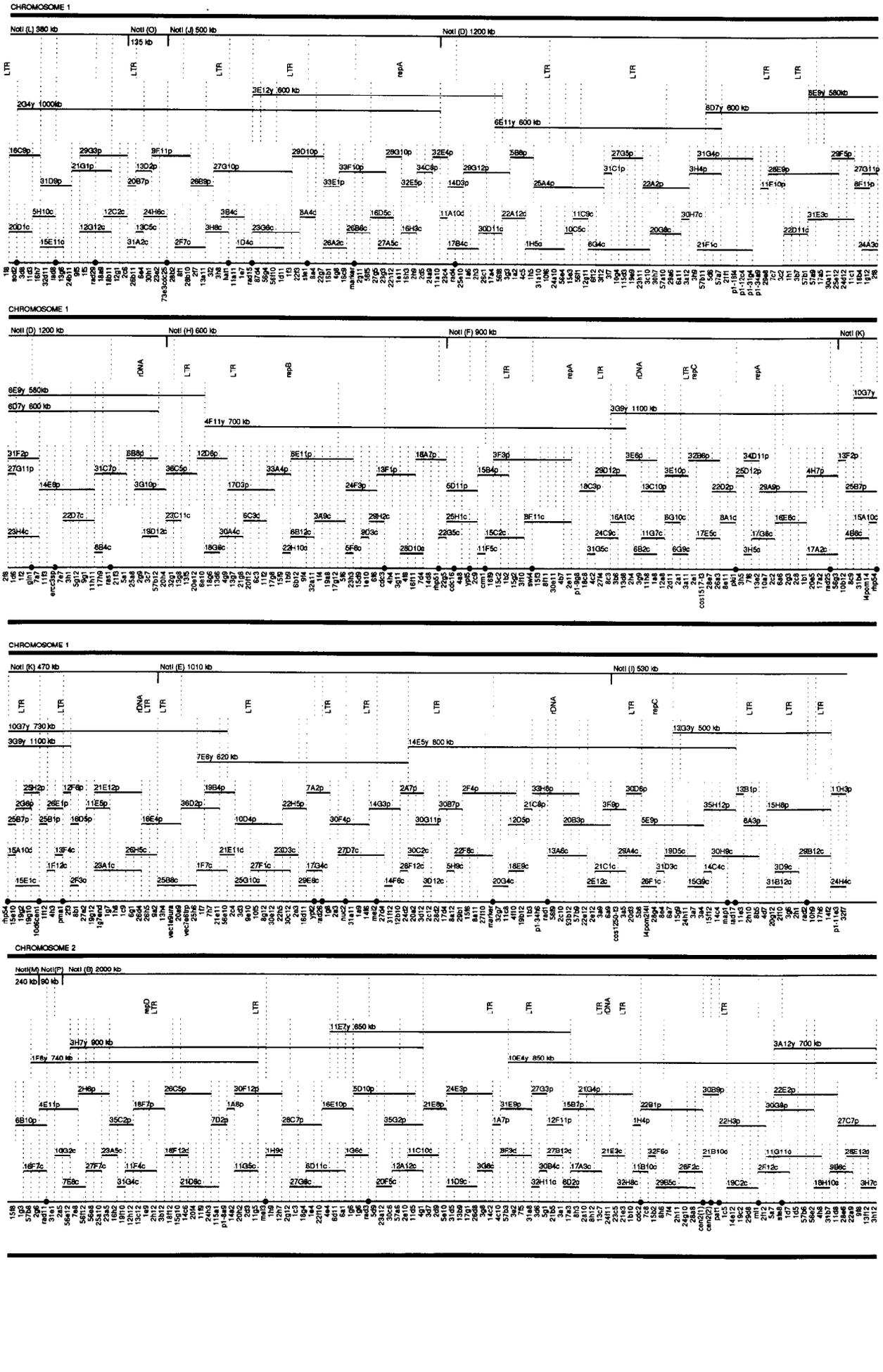
As a quality control and confirmation of the cosmid map, a parallel map was built with the P1 library, ordering the clones by hybridizing a subset of the probes already used on the cosmids. Wherever necessary, additional probes were hybridized to make the probe order as determined with the P1 library valid on its own account. Any discrepancy between the two maps was solved by comparisons with other available mapping data, in particular, the YAC map, followed by further probe hybridizations as verification.

The remaining gaps in the cosmid map were known from the YAC data to be no longer than an average P1 clone. To speed up the gap closure process, the bridging of the remaining breaks was carried out with the P1 library clones, owing to their greater length (about twice that of a cosmid), instead of extending work to the remaining 15 genome equivalents available in cosmids.

Physical Clone Map of S. pombe

Figure 2 shows the physical map of the entire S. pombe genome. The established probe order is based on the combined results from the YAC, P1, and cosmid libraries. A set of clones from each of these libraries is shown (25 YACs, 209 P1 clones, and 241 cosmids). Clone and probe names are given, so that any region can be described immediately by a list of names. Note that clones that have only one probe in common do not necessarily overlap. Since the probes are clones themselves, in some cases the listed clones only overlap with either end of the relevant probe, not with each other. Therefore, the probes must be

Cell
112


CHROMOSOME 1
NotI (L) 380 kb
NotI (O)
135 kb
NotI (J) 500 kb
NotI (D) 1200 kb
LTR
repA
YAC
2G4y 1000kb
3E12y 600 kb
8E9y 580kb
8D7y 600 kb
6E11y 600 kb
P1
Cosmid
CHROMOSOME 1
NotI (D) 1200 kb
NotI (H) 600 kb
NotI (F) 900 kb
NotI (K)
rDNA
repB
repC
6E9y 580kb
6D7y 800 kb
4F11y 700 kb
3G9y 1100 kb
10G7y
CHROMOSOME 1
NotI (K) 470 kb
NotI (E) 1010 kb
NotI (I) 530 kb
10G7y 730 kb
3G9y 1100 kb
7E6y 620 kb
14E5y 800 kb
13G3y 500 kb
CHROMOSOME 2
NotI(M) NotI(P)
NotI (B) 2000 kb
240 kb
90 kb
repD
11E7y 850 kb
3H7y 900 kb
1F8y 740 kb
10E4y 850 kb
3A12y 700 kb

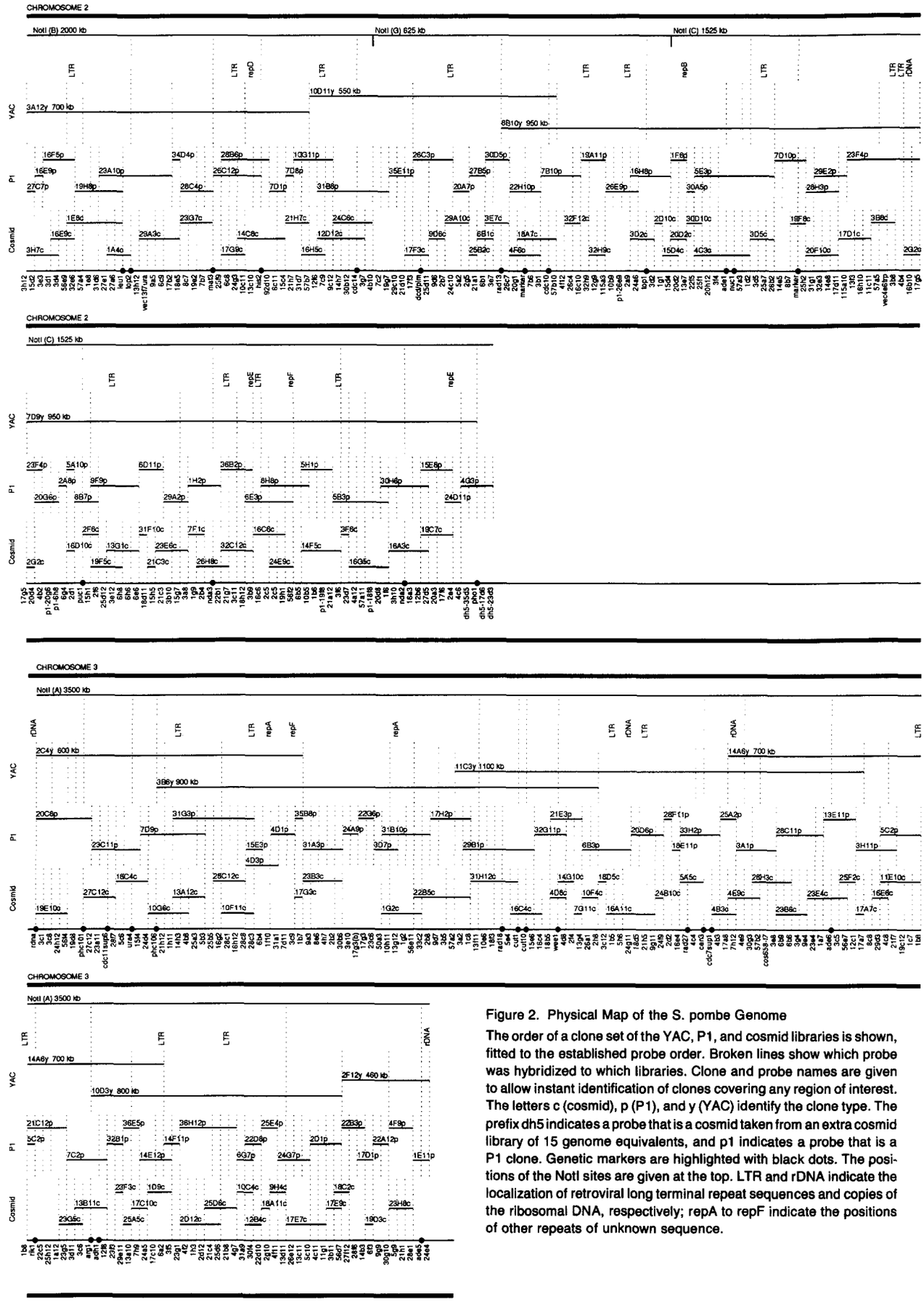


Figure 2. Physical Map of the S. pombe Genome

The order of a clone set of the YAC, P1, and cosmid libraries is shown, fitted to the established probe order. Broken lines show which probe was hybridized to which libraries. Clone and probe names are given to allow instant identification of clones covering any region of interest. The letters c (cosmid), p (P1), and y (YAC) identify the clone type. The prefix dh5 indicates a probe that is a cosmid taken from an extra cosmid library of 15 genome equivalents, and p1 indicates a probe that is a P1 clone. Genetic markers are highlighted with black dots. The positions of the NotI sites are given at the top. LTR and rDNA indicate the localization of retroviral long terminal repeat sequences and copies of the ribosomal DNA, respectively; repA to repF indicate the positions of other repeats of unknown sequence.

Table 1. Results of the Contig Assembly Process

Chromosome	Contig Gaps in			Number of Probes to Contigs		
	Cosmids	P1 Clones	Combined	Cosmids	P1 Clones	Combined
I	31	3	3	84[a]	51	95
5700 kb				67	76	87
				42	40	48
				79	58	97
				(272)[b]	(225)	(327)
				[20.9 kb][c]	[25.3 kb]	[17.4 kb]
II	24	0	0	(231)	(171)	(279)
4600 kb				[19.9 kb]	[26.9 kb]	[16.5 kb]
III	14	0	0	(120)	(109)	(158)
2500 kb + 1000 kb rDNA				[20.8 kb]	[22.9 kb]	[15.1 kb]
Total	69	3	3	623 probes	505 probes	764 probes

[a] Each row is the number of probes to 1 of the 4 contigs of chromosome I.
[b] Numbers in parentheses, subtotals of probes to the entire chromosome.
[c] Numbers in brackets, average probe distances on the chromosome.

included in any list of clones spanning a certain region.

For most of the genome, the probes are relatively evenly spaced along the genome. Only in regions of uncertainty, i.e., mainly around repetitive sequences, was the probe number significantly increased above average (e.g., the area between probes 28c3 and 13g12 on chromosome III). Altogether, 764 probes were hybridized to the P1 and cosmid libraries, with all the probes used for the creation of the YAC map included in this number. There were no major differences in the average distance between probes among the three chromosomes (Table 1). The main cause for the relatively high number of breaks in the cosmid map is the insufficient coverage by cosmids of about 8 genome equivalents and, resulting from this fact, the strategy of closing gaps by hybridizations to the P1 library rather than by work-intensively exhausting the 8-fold cosmid coverage to its limits or by extending work to the remaining 15 genome equivalents. However, interruptions of the cosmid map are also due to the presence of rDNA sequences throughout the genome (see below). The greater insert length of the P1 clones gave them a better chance to span the 10.4 kb rDNA sequence.

After filtering out the clones that are positive for the rDNA and centromeric clones and eliminating apparent well-to-well clone contaminations, the remaining clones are found on average by three probes each (P1, 2.93; cosmid, 3.03). Twenty-seven percent of the cosmids and 29.5% of the P1 clones show connections to other regions of the genome by an average of 1.4 probes. This rate of cross-hybridization of a rather distinct portion of the clones is probably caused to a large extent by repeated misscoring of clones with a high copy number. This effect is most clearly visible with the ~2.7% of the cosmid library that consists of vector concatemers with no genomic insert, which showed up strongly in most hybridizations. Another factor, of course, could be the presence of relatively short repeated sequences that cross-hybridize. Sequences as short as 25 bp have been found to be responsible for such a signal (J. D. H., unpublished data).

Relation to Physical and Genetic Mapping Data

The map obtained with the cosmids and P1 clones is in agreement with the YAC clone order (Figure 2; Maier et al., 1992) and aligns to the NotI restriction map (Fan et al., 1989). The locations of most of the 75 genetic markers on the map are consistent with the genetic map. The few inconsistencies previously detected by the YAC mapping (Maier et al., 1992) were confirmed: *rad11* was found to be on the short arm of chromosome II, and *wee1* is located next to *cut1* and *cut10* on the short arm of chromosome III. The order of *nda2* and *nda3* on the long arm of chromosome II is reversed, so that *nda3* is closer to the centromere.

In addition to these findings, the marker *pho1* (Elliott et al., 1986), thought to map in between *nda3* and *nda2*, was mapped distal to *nda2*. Based purely on the P1 hybridization data, the ordering algorithm actually produced a loop at the end of the long arm of chromosome II. Its stem was created by a strong cross-hybridization (repE; Figure 2) between probe 3b9 (distal to *nda3*) and probes 17f6 and 2e4 (proximal to *pho1*). The region around *pho1* was thereby actually left unconnected. Taking into account the probe order of the YAC map and to a lesser extent that of the cosmid map, however, the branch containing probes 17f6 and 2e4 was cut off, thereby creating the map shown in Figure 2. Additional circumstantial evidence for the very terminal position of *pho1* was provided by the poor and variable cloning efficiency of the genomic DNA surrounding the marker. No cosmid containing *pho1* was identified in the total cosmid library of 23-fold genome coverage. Screening another cosmid library of an additional 15 genome equivalents (unpublished data) did detect three clones, however. The genomic insert DNA used for this second library had been partially cut with about twice the frequency used the first time. Nearly identical cleavage conditions had been used for the P1 library, in which three clones were found, too. This phenomenon can be explained by the fact that there are few MboI sites situated more terminal than *pho1*, causing the strong underrepre-

sentation of the region and an increasing probability of cloning with a higher frequency of cleavage.

The marker *adh1*, for which there was no mapping information available (Lennon and Lehrach, 1992), is located next to *arg1* and *rik1* on the long arm of chromosome III.

Repeat Sequences and Ambiguities

The 10.4 kb rDNA repeat (Schaak et al., 1982), retrotransposable regions (Levin et al., 1990), and centromeric sequences (Chikashige et al., 1989) are the three most frequently occurring repeat sequences in S. pombe. Apart from the 70 to 100 copies of rDNA clustered at both ends of chromosome III, there was cross-hybridization to rDNA by cosmid and P1 clones across the entire genome. Based on the hybridization results, however, the locations of only 8 copies of the repeat are unambiguous (Figure 2). There are 4 copies on chromosome I, 2 on chromosome II, and 2 in the middle of chromosome III. Using a probe for the long terminal repeat (LTR) sequence of retroviral elements (Levin et al., 1990), 52 locations were identified (Figure 2): 25 on chromosome I, 20 on chromosome II, and 7 on chromosome III. Twenty-three cosmids (0.8% of the clones) and 37 P1 clones (1.1%) from the centromeric regions showed cross-hybridization. Owing to the long insert sizes of the P1 clones and with the help of cosmids identified by chromosome-specific centromere probes and confirmed locations on the YAC map, all three centromeres could be spanned with the P1 library.

Apart from these major repeats, six other repeated elements were identified. One is located at four positions in the genome (repA; Figure 2). There was very strong cross-hybridization between the clones detected by probes 1e11 (distal to *rad4*, chromosome I), 2e11 (proximal to *swi4*, chromosome I), 13e2 (proximal to *pki1*, chromosome I), and 1f10 and 13g12 (distal to *rad18*, chromosome III). At the latter position, this repeat was responsible for the best probe order, as calculated by the ordering algorithm, consistently creating a branching point in the map, with one branch leading to a dead end. The probe order in this region was manually adapted, taking into account the known probe order and distances of the YAC map and the genetic distances between genetic markers surrounding this area, so that it fit these criteria but was still in agreement with the hybridization data obtained from the cosmid and P1 clones, although it does not represent the best fit purely based on this data set. The analysis was made even more difficult by the presence of another repeat sequence (repF) linking 56f2 (distal to *nda3*, chromosome II) to 3c9 (proximal to 1f10, chromosome III). Further connections were found between (repB) 15f9, 1b9 (distal to *cdc5*, chromosome I), and 13e7 (distal to *top1*, chromosome II); (repC) 2e1 (distal to *pki1*, chromosome I) and 28g4 (proximal to *rad17*, chromosome I); and (repD) 1e9 (distal to *mei3*, chromosome II) and 13c10 (*his2*, chromosome II).

While the maps of chromosomes II and III are continuous, there are three contig breaks in chromosome I. The gap between probes p1-31g4 and p1-34a9 (distal to *gln1*) might actually be bridged by a single clone. These data, however, cannot be considered valid by themselves and could not be confirmed by other means. The gap between probes 2e11 and p1-9g8 (proximal to *swi4*) could be caused by the repeat sequence repA. The intensity of cross-hybridization is strong, suggesting several copies of the repeat in this position, which might prevent the connection between the probes. For the third break, at probes 8b1 and 27e2 near the centromere, there is no apparent explanation. The lack of connecting clones could be due to unclonability in E. coli, the absence of MboI sites in this region, or sheer coincidence, although the latter is unlikely, considering the coverage of the libraries used.

Reference System

Filters containing all the library clones have been distributed freely during the period of mapping, and clones identified by the users have been provided. Apart from linking the probes to the developing physical map and vice versa, other laboratories could already work on a more detailed (e.g., sequence) analysis of the genomic region of their interest even before the physical map was completed. Their markers were thus automatically positioned in the S. pombe genome without any additional effort by the user. Also, since one particular genome is being used as a reference by a large number of groups, the data generated are more easily comparable.

Reference filters of all three libraries, as well as identified clones, representing either the entire library or the minimal set of clones shown in Figure 2, are being distributed without restriction. The minimal set of clones spanning the genome is also available under the condition that the data generated will eventually be reported to the reference data base set up in our laboratory. The information stored in this data base will be freely accessible through the reference library system (Lehrach et al., 1990). Restriction patterns generated with the enzymes BamHI, PstI,

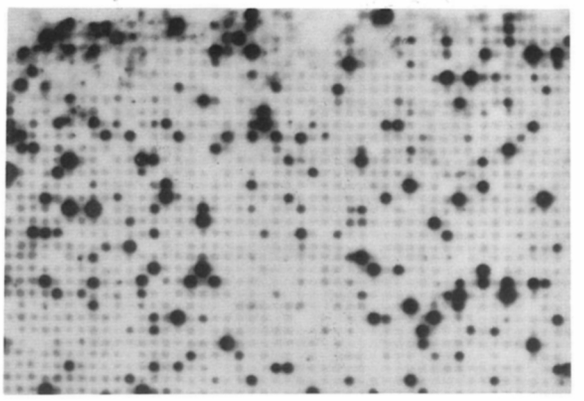


Figure 3. Hybridization of the Oligonucleotide GCTTCTCAAAG to a High Density Filter

The experiment was actually carried out on a filter containing all 8640 cosmids. Only one-sixth of the filter is shown. Probe annealing was carried out as described in Experimental Procedures. The filter was washed at 25°C for 3 hr. Film was exposed for 6 hr.

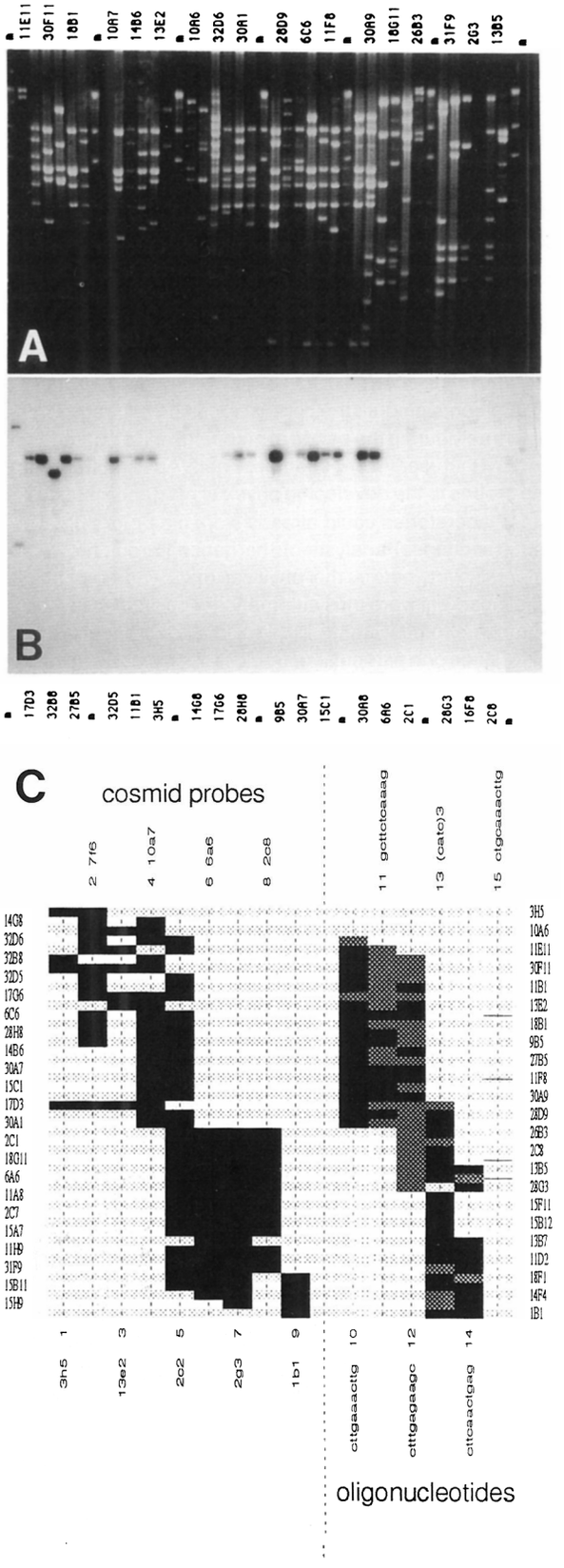


Figure 4. Assessment of the Quality of Oligomer Hybridization

(A) BamHI–PstI digests of clones from part of the contig represented in (C) (part of the region between *pki1* and *rad25* on chromosome I). Preparations are presented from all clones assigned by an earlier version of the map to this section by more than one positive cosmid probing.

(B) Southern blot of (A), probed with the undecamer GCTTCTCAAAG (lane 11 in [C]). The two smaller fragments at the left of the gel are due to the fact that the relevant genomic fragment is linked to the vector arm in these clones. The appearance of a second band in the leftmost lane is due to incomplete digestion of the DNA. The gap in oligomer signal between clones 10A6 and 32D6 is due to mispositioning of these clones in the early version of the map on which the clone order in the gel was based. m, DNA marker.

(C) Representation of the manually scored oligonucleotide probing data from high density filters (closed), additional cosmids containing the same fragments detected on Southern blots (stippled), and false positives (horizontal lines in lane 15). Probe names are given at the top and bottom, clone names at the right and left.

and HindIII on a set of clones covering the genome are stored in this data base. Owing to a comparatively high cleavage frequency, the patterns produced by the latter enzyme, in particular, might allow the positioning of clones, for example cDNAs, without even hybridizing them to the reference filters.

Assessment of the Oligomer Hybridizations

To test the effectiveness of the oligomer fingerprinting approach, a set of oligonucleotides was hybridized to high density filters of the S. pombe cosmids (e.g., Figure 3). Two types of oligonucleotides were used for hybridization. The first type was dodecamers based on simple tandem di-, tri-, and tetranucleotide repeats. The second type consists of random sequence undecamer oligonucleotides with 30%–50% G or C content, avoiding the nearest-neighbor sequence compositions known to be highly frequent in the E. coli genome. The latter set had been selected using a higher order Markov chain selection algorithm. Hybridization data were scored by hand; hence, normalization for the amount of DNA in each filter spot could not be performed. Out of a larger group of oligonucleotides, 32 oligomers produced discriminable signal, with an average signal frequency of 3.8% of the clones. To assess the quality of these data, we prepared DNAs from the complete set of 100 cosmids in 2 contigs, digested them with BamHI and PstI, and hybridized Southern blots of them to the oligomer probes that scored positive in these regions and, as a control, a few of the negative ones (e.g., Figures 4A and 4B). Figure 4C presents 1 of the 2 contigs actually ordered according to the oligomer data. This caused a slight change in the clone order as determined from the single cosmid hybridizations; the probe order was unaffected, however. The apparently missing values in the cosmid–cosmid hybridization data are due to nongrowth of the clones of the filters, variable clone length, or unscored weak signals. Most of the latter show up in the hybridization with 13e2, which is a repetitive probe (repA), so only the few unequivocally unique signals are shown.

In the majority of cases, the oligomer signals from the high density filter clearly correspond to hybridization with a particular restriction fragment. In many cases, additional cosmids with the same fragment were found by Southern

hybridization. These missed positives were the most common discrepancy, with high density filter probings missing from none to more than half of their targets and an overall false negative frequency of 26% for this data set. In several cases, more than one oligomer detects a particular BamHI–PstI fragment, for example, CTTGAAACTTG (Figure 4, lane 10), GCTTCTCAAAG (lane 11), and CTTTGAGAAGC (lane 12). Between them, the pattern of undetected positive clones is not the same. The potentially more troublesome false positives are much less frequent, with the tested set of oligomers giving rise to 19 scattered false positive signals out of a total of 234 scored in the two areas. The observed pattern of discrepancies probably reflects a bias inherent in manual scoring, in that only the more obviously positive colonies are likely to be noticed.

Discussion

The feasibility of large-scale genome mapping by hybridization fingerprinting clone libraries has recently been demonstrated on the YAC level with the coverage of the entire S. pombe genome (Maier et al., 1992) and the euchromatin of the human Y chromosome (Foote et al., 1992). While an ordered YAC library bridges the gap between the levels of linkage and cytogenetic mapping on one hand and of recombinant DNA technology on the other, a finer analysis is necessary for more detailed studies of the genomic DNA and ultimately for its sequencing. The work presented here shows that hybridization techniques are a powerful means to that end.

Most basic placement information was obtained from hybridizations of entirely anonymous clones. The frequency of hybridization allows the calculation of the distance between probes and hence the determination of their order as a probe tagged site map. The advantage of this approach is the fact that no information is needed about a clone used as a probe before its application as an ordering tool. The very same raw data that determined the probe order are then used again to fit a clone coverage to the generated probe tagged site map. This employment of one data set for two purposes is one of the reasons for the efficiency of the methodology; two additional reasons are the ability to relate results from any level of analysis and the capacity to examine large clone numbers in parallel. In comparison with the gel fingerprinting of C. elegans, about one-quarter as many experiments was sufficient for the initial characterization of a region in cosmids (620 for 14 Mbp:17,000 for 100 Mbp; Coulson et al., 1991), with far less experimental effort involved.

The results obtained with the clone pools were effectively comparable with the single clone hybridizations. However, a higher quality of the hybridization results was required. For analysis on the cosmid and/or P1 level of a genome the size of the unicellular fungus or smaller, the advantage of a reduced number of experiments is balanced or outweighed by the considerable increase in experimental effort. For larger genomes, however, pooling offers a way of ordering at least those regions that do not contain too many repeat sequences. Particularly the use as probes of transcribed sequences such as cDNAs and exon-trap clones will allow the mapping of the genomic DNA in parallel with the simultaneous positioning of the transcribed sequences on this map. Work on this basis is under way in our laboratory with, for instance, two Drosophila cDNA reference libraries of good representation whose clones have been checked for repeat sequences and various genomic libraries of Drosophila melanogaster (Hoheisel et al., 1991); work has also been started with a wide variety of human libraries.

As illustrated by the experiments on S. pombe, the main obstacle for a practical application of oligonucleotide fingerprinting for high resolution mapping in cosmids and P1 clones on a large scale is the speed and accuracy with which these data can be harvested. To utilize fully the oligonucleotide fingerprinting technique, the rate of false negatives has to be reduced. The signal strength is firmly dependent on the quantity of target DNA present at each spot. The DNA amounts, however, vary strongly on account of the cosmid copy number in the cells as well as the cell number itself. Signal intensities have hence to be normalized on the basis of a vector hybridization, indicating the amount of cosmid DNA at each filter position, to obtain reliable data. In addition, the frequency of positive clones and thereby the efficiency of the procedure would be increased, because a manual analysis is biased toward the strong signals, missing the large number of less obvious positive clones. Thus, improvement in automated image processing is absolutely necessary for the analysis of oligomer data and is the focus of our current efforts on the process of data gathering.

For the establishment of the clone order with oligonucleotides, expansion of the analysis software is still necessary. On the experimental side, for instance, the YAC-based subdivision of cosmid and P1 libraries will simplify the reconstruction of an oligonucleotide map, because multilocus oligonucleotide probes can then be locally treated as single-copy probes in regions defined by hybridizations with YACs. The battery of 32 oligomers used on S. pombe is not large or informative enough to give the cosmid order by itself. However, it demonstrates the informative value of this work-saving approach, especially when projected for use on larger genomes. Although the search for suitable oligonucleotides proved to be more cumbersome than expected, a set of about 150 oligonucleotides would allow fingerprinting of cosmid and P1 libraries made from very large genomes with less experimental effort than has been put into the mapping of S. pombe.

The parallel use of three clone libraries and the instant relation of the different analysis levels greatly helped the progress of the mapping. The YAC map provided a framework that supported the more detailed analysis in P1 and cosmid clones and made the resolution of difficult regions easier or possible at all, a fact also observed during the C. elegans analysis (Coulson et al., 1991). Cosmid cloning, on the other hand, provided an extremely robust system, combining a reasonable insert size with very simple handling. The P1 clones proved to be a valuable alternative or complement to cosmid cloning. Although the library construction at the time was not as straightforward as cosmid

cloning, these problems have been overcome with later modifications. Owing to their average insert length of about double that of a cosmid, P1 clones were more successful in spanning repeat sequences, particularly the 10.4 kb rDNA. However, as far as experimental work for (further) analysis that requires DNA preparation is concerned, cosmids are advantageous for their higher yield and purer DNA.

For a large-scale mapping project aiming at complete coverage, a minimum of about 10 genome equivalents seems to be essential, a number suggested by preliminary calculations on the basis of the S. pombe results. Although such good coverage was not necessary to obtain data on all regions of the genome with high probability, any degree of redundancy proved to be extremely helpful and was certainly essential for resolving problematic areas. Since library construction and data acquisition by hybridization from larger libraries are less work intensive and time consuming than a subsequent investigation of unresolved regions, higher coverage should be sought.

The high resolution physical map of S. pombe provides instant access to any segment of the genome. The probe order for all three libraries—YAC, P1, and cosmid—was determined independently, based on the relevant set of raw data. The congruity between these data sets, as well as their coherence to the genetic and the NotI information, illustrates the accuracy of the clone coverage obtained. Apart from the known repeat sequences, for which there were specific probes available (long terminal repeat and rDNA), the locations of unknown repeat sequences in the genome were identified merely on the basis of the hybridization results. Because of the lack of an automated image analysis that could take into account different signal intensities rather than rely on yes–no digitizing information, and because of the lack of problem-oriented software, only strong cross-hybridization events were identified. Further developments in this direction might lead to a more detailed analysis of the genome structure from the basic hybridization data.

The physical clone map will readily assist in the sequence determination of the S. pombe genome by directly serving as a substrate for standard sequencing strategies. Alternatively, a fragment library of 0.5–1.0 kb DNA fragments could be created and ordered by hybridizations on the basis of the cosmid and P1 maps and additional internal order information from oligomer hybridizations. The necessary experimental tools for polymerase chain reaction amplifying tens of thousands of such fragments within a week (Meier-Ewert et al., 1993) and spotting them on a support medium for hybridization are in place. This set of fragments, because they are ordered, would allow an efficient approach toward the sequencing of the genome. The sequence analysis of S. pombe and its comparison with Saccharomyces cerevisiae (Oliver et al., 1992), C. elegans (Sulston et al., 1992), and humans should provide extremely valuable information for the understanding of their biologies and define the gene set that underlies the basic eukaryotic cell and is likely to be common to all eukaryotic organisms.

Experimental Procedures

Preparation of the Clone Libraries

Cells of the S. pombe strain 972h$^-$ were grown to saturation in M9 minimal medium (Sambrook et al., 1989). For cosmid cloning, the DNA was prepared according to Herrmann and Frischauf (1987) and cloned into Lawrist4 vector following a protocol detailed earlier (Hoheisel et al., 1991), yielding clones with an average insert size of 37.5 kb. The 8640 individual clones picked and stored in microtiter dishes represent the genome about 23-fold.

For cloning into bacteriophage P1, the S. pombe cells were embedded in agarose, and the chromosomal DNA was prepared as described by Larin and Lehrach (1990), the final DNA concentration being about 2 μg per block. The DNA was partially digested with MboI to produce fragments of ~80 kb on average. After dephosphorylation with calf intestinal alkaline phosphatase, blocks were treated with proteinase K and phenylmethylsulfonyl fluoride (Larin et al., 1991). Vector arms were generated from pAd10sacB11 (Pierce et al., 1992) by linearization with ScaI, dephosphorylation, and BamHI cleavage at the cloning site. The ligation reaction was performed in agarose with an 8 M excess of vector arms and 2.5 U/μl T4 DNA ligase (New England Biolabs) at 16°C overnight. The reaction was terminated by addition of EDTA to 20 mM. The ligation mixture was melted at 68°C for 15 min and loaded evenly into a trough of a 1% low melting point gel in 0.5× TBE buffer (Sambrook et al., 1989). The gel was electrophoresed in a contour-clamped homogeneous electric field apparatus for 16 hr, using a pulse time of 4 s at 180 V. DNA of limiting mobility was isolated by agarose treatment and packaged in vitro as described (Sternberg et al., 1990). Clones were recovered in the E. coli strain NS3145 and selected on 2YT agar containing 30 μg/ml kanamycin and 5% sucrose. A total of 3456 clones with an average insert size of 70 kb (17 genome equivalents) were picked individually into microtiter dishes.

Clones were robotically arrayed in high density grids of 9,216 (i.e., 3 × 3,072) or 20,736 (6 × 3,456) clones onto 22 × 22 cm Hybond N$^+$ membranes and grown overnight on agar plates containing 30 μg/ml kanamycin. Cosmid incubation was performed at 37°C, while P1 clones were grown at room temperature, followed by an induction on fresh plates containing 1 mM isopropyl-β-D-thiogalactopyranoside for 6 hr at 37°C. Filter processing was carried out as described (Hoheisel et al., 1991).

Hybridizations

Cosmid, P1, and marker DNAs were prepared by an alkaline lysis procedure (Pohl et al., 1982). Probe DNA was radioactively labeled by random hexamer priming (Feinberg and Vogelstein, 1983) or, in the case of cosmid pools, by a primer extension reaction, as described in detail elsewhere (Hoheisel et al., 1991). No competition reaction was carried out. The library filters were prehybridized in 0.5 M sodium phosphate (pH 7.2), 7% SDS, 1 mM EDTA, and 0.1 mg/ml yeast tRNA at 65°C for 2 hr to overnight. Hybridization was performed in the same buffer at 65°C overnight, and the probe concentration was 2 × 10^5 to 5 × 10^5 cpm/ml. The filters were briefly rinsed in 40 mM sodium phosphate (pH 7.2), 0.1% SDS at room temperature. The same buffer at room temperature was then added, and the filters were washed by rocking slowly in a water bath at 65°C for about 20 min. Only filters hybridized with identical probe were washed together. The filters were briefly blotted dry, and film was exposed for 2 hr to overnight at −70°C, using an intensifying screen. To strip the probe off the filters, 1–2 liters of 5 mM sodium phosphate (pH 7.2), 0.1% SDS at room temperature was poured into a box containing up to 60 filters and kept in a water bath at about 90°C for 30 min, after which the procedure was repeated. Following this procedure, filters continued to produce reasonable results after ~30 cycles.

DNA oligomers were labeled at the 5′ end with [γ-^{32}P]ATP and T4 polynucleotide kinase. Hybridization was carried out in 600 mM NaCl, 60 mM sodium citrate, 7.2% (v/v) sodium Sarkosyl, and about 20 pM probe at 8°C overnight. The filters were washed in the same buffer at temperatures between 8°C and 28°C over various time periods.

Data Analysis

Hybridization results on autoradiographs were either digitized using specially adapted software (A. R. Ahmadi, unpublished data) or read

in manually. Data analysis was carried out on a SUN SPARC-II workstation, using software written in C and Prolog. The data base handles hybridization data from a variety of sources, e.g., optical densities from computer-based image analysis, x and y coordinates from the digitizer, and text files of clone names. These input formats are translated into a canonical format, listing the positive clones for each experiment merged for each round of analysis. The software tools developed allow a quality control of the data, from a simple graphical listing of the hybridization results according to entry date, to checking for experimental artifacts (e.g., high probe concentration produced an increased background signal on some clones), to the identification of probable well-to-well contaminations in the library microtiter dishes, to rather complex comparisons between the various clone maps (Mott et al., submitted).

Alignment to the NotI map

S. pombe genomic DNA was digested to completion with NotI. The fragments were separated by contour-clamped homogeneous electric field gel electrophoresis and were transferred onto filters. Clones from the developing physical maps were hybridized to the digests (Figure 2). From these results and from hybridizations with the minimal set of YAC clones, the positions of the NotI sites could be narrowed down to relatively small areas. Clones covering those regions were isolated, hybridized discriminately with a NotI octamer (Hoheisel et al., 1990), and/or digested with NotI to determine the locations of the sites accurately.

Acknowledgments

Janet Roden made herself irreplaceable in organizing the laboratory. We thank Marion Denter and Tobias Hatter for their technical help during some periods of this work. Huw Griffith and Ali Reza Ahmadi were helpful with computation. We also thank John Armstrong, Antony Carr, Richard Egel, Susan Forsburg, Kathy Gould, Daan Muris, Olaf Nielsen, Tom Patterson, Viesturs Simanis, and Mitsuhiro Yanagida for providing genetic probes of S. pombe and Paul Nurse for the S. pombe strain 972h$^-$ and comments on the manuscript. E. M. was funded by a European Community Research Fellowship.

Received November 18, 1992; revised January 15, 1993.

References

Booher, B., and Beach, D. (1989). Involvement of a type 1 protein phosphatase encoded by *bws1*$^+$ in fission yeast mitotic control. Cell *57*, 1009–1016.

Chikashige, Y., Kinoshita, N., Nakaseko, Y., Matsumoto, T., Murakami, S., Niwa, O., and Yanagida, M. (1989). Composite motifs and repeat symmetry in S. pombe centromeres: direct analysis by integration of NotI restriction sites. Cell *57*, 739–751.

Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., and Waterston, R. (1991). YACs and the *C. elegans* genome. Bioessays *13*, 413–417.

Coulson, A. R., Sulston, J., Brenner, S., and Karn, J. (1986). Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA *83*, 7821–7825.

Cox, D. R., Burmeister, E., Price, R., Kim, S., and Myers, R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. Science *250*, 245–250.

Craig, A., Nizetic, D., Hoheisel, J. D., Zehetner, G., and Lehrach, H. (1990). Ordering of cosmid clones covering the herpes simplex virus type I (HSV-I) genome: a test case for fingerprinting by hybridization. Nucl. Acids Res. *18*, 2653–2660.

Elliott, S., Chang, C., Schweingruber, M. E., Schaller, J., Rickli, E. E., and Carbon, J. (1986). Isolation and characterisation of the structural gene for secreted acid phosphatase from *Schizosaccharomyces pombe*. J. Biol. Chem. *261*, 2936–2941.

Enoch, T., and Nurse, P. (1991). Coupling M phase and S phase: controls maintaining the dependence of mitosis on chromosome replication. Cell *65*, 921–923.

Evans, G. A., and Lewis, K. A. (1989). Physical mapping of complex genomes by cosmid multiplex analysis. Proc. Natl. Acad. Sci. USA *86*, 5030–5034.

Fan, J.-B., Chikashige, Y., Smith, C. L., Niwa, O., Yanagida, M., and Cantor, C. R. (1989). Construction of a NotI restriction map of the fission yeast *Schizosaccharomyces pombe* genome. Nucl. Acids Res. *17*, 2801–2818.

Feinberg, A. P., and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. *132*, 6–13.

Foote, S., Vollrath, D., Hilton, A., and Page, D. C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. Science *258*, 60–66.

Forsburg, S. L., and Nurse, P. (1991). Cell cycle regulation in the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Annu. Rev. Cell Biol. *7*, 227–256.

Hayles, J., and Nurse, P. (1992). Genetics of the fission yeast *Schizosaccharomyces pombe*. Annu. Rev. Genet. *26*, 373–402.

Herrmann, B. G., and Frischauf, A.-M. (1987). Isolation of genomic DNA. Meth. Enzymol. *152*, 180–183.

Hoheisel, J. D., Craig, A. G., and Lehrach, H. (1990). Effect of 5-bromo and 5-methyl deoxycytosine on duplex stability and discrimination of the NotI octadeoxynucleotide: quantitative measurements using thin-layer chromatography. J. Biol. Chem. *265*, 16656–16660.

Hoheisel, J. D., Lennon, G. G., Zehetner, G., and Lehrach, H. (1991). Use of high coverage reference libraries of *Drosophila melanogaster* for relational data analysis: a step towards mapping and sequencing of the genome. J. Mol. Biol. *220*, 903–914.

Kinoshita, N., Ohkura, H., and Yanagida, M. (1990). Distinct, essential roles of type 1 and 2A protein phosphatases in the control of the fission yeast cell division cycle. Cell *63*, 405–415.

Larin, Z., and Lehrach, H. (1990). Yeast artificial chromosomes: an alternative approach to the molecular analysis of mouse developmental mutations. Genet. Res. *56*, 203–208.

Larin, Z., Monaco, A. P., and Lehrach, H. (1991). Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proc. Natl. Acad. Sci. USA *88*, 4123–4127.

Lehrach, H., Drmanac, R., Hoheisel, J., Larin, Z., Lennon, G., Monaco, A. P., Nizetic, D., Zehetner, G., and Poustka, A. (1990). Hybridization fingerprinting in genome mapping and sequencing. In Genome Analysis, Volume 1: Genetic and Physical Mapping, K. E. Davies and S. M. Tilghman, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 39–81.

Lennon, G. G., and Lehrach, H. (1992). Gene database for the fission yeast *Schizosaccharomyces pombe*. Curr. Genet. *21*, 1–11.

Levin, H. L., Weaver, D. C., and Boeke, J. D. (1990). Two related families of retrotransposons from *Schizosaccharomyces pombe*. Mol. Cell. Biol. *10*, 6791–6798.

Maier, E., Hoheisel, J. D., McCarthy, L., Mott, R., Grigoriev, A. V., Monaco, A. P., Larin, Z., and Lehrach, H. (1992). Yeast artificial chromosome clones completely spanning the genome of *Schizosaccharomyces pombe*. Nature Genet. *1*, 273–277.

Meier-Ewert, S., Maier, E., Ahmadi, A., and Lehrach, H. (1993). An automated approach to generating expressed sequence catalogues. Nature *361*, 375–376.

Nizetic, D., Zehetner, G., Monaco, A. P., Gellen, L., Young, B. D., and Lehrach, H. (1991). Construction, arraying and high density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. Proc. Natl. Acad. Sci. USA *88*, 3233–3237.

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). The complete DNA sequence of yeast chromosome III. Nature *357*, 38–46.

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common

language for physical mapping of the human genome. Science *245*, 1434–1435.

Pierce, J. C., Sauer, B., and Sternberg, N. (1992). A positive selection vector for cloning high molecular weight DNA by the bacteriophage P1 system: improved cloning efficiency. Proc. Natl. Acad. Sci. USA *89*, 2056–2060.

Pohl, F. M., Thomae, R., and Karst, A. (1982). Temperature dependence of the activity of DNA-modifying enzymes: endonucleases and DNA ligase. Eur. J. Biochem. *123*, 141–152.

Ross, M. T., Hoheisel, J. D., Monaco, A. P., Larin, Z., Zehetner, G., and Lehrach, H. (1992). High density gridded YAC filters: their potential as genome mapping tools. In Techniques for the Analysis of Complex Genomes, R. Anand, ed. (New York: Academic Press), pp. 137–154.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular Cloning, A Laboratory Manual: Second Edition (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Schaak, J., Mao, J., and Söll, D. (1982). The 5.8S RNA gene sequence and the ribosomal repeat of *Schizosaccharomyces pombe*. Nucl. Acids Res. *10*, 2851–2864.

Sternberg, N., Ruether, J., and deRiel, K. (1990). Generation of a 50,000-member human DNA library with an average DNA insert size of 75–100 kb in a bacteriophage P1 cloning vector. New Biol. *2*, 151–162.

Sulston, J., Du, Z., Thomas, K., Wison, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, R., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., and Waterston, R. (1992). The *C. elegans* genome sequencing project: a beginning. Nature *356*, 37–41.