Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects

U. Radelof*, S. Hennig, P. Seranski1, M. Steinfath1, J. Ramser, R. Reinhardt, A. Poustka1, F. Francis2 and H. Lehrach

Max-Planck-Institut für Molekulare Genetik, Ihnestraße 73, 14195 Berlin, Germany, 1Abteilung Molekulare Genomanalyse, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany and 2Institut Cochin de Génétique Moléculaire, INSERM U 129, CHU Cochin Port-Royal, 24 rue du Faubourg Saint-Jacques, 75014 Paris, France

Received August 14, 1998; Revised and Accepted October 13, 1998

ABSTRACT

Large-scale genomic sequencing projects generally rely on random sequencing of shotgun clones, followed by different gap closing strategies. To reduce the overall effort and cost of those projects and to accelerate the sequencing throughput, we have developed an efficient, high throughput oligonucleotide fingerprinting protocol to select optimal shotgun clone sets prior to sequencing. Both computer simulations and experimental results, obtained from five PAC-derived shotgun libraries spanning 535 kb of the 17p11.2 region of the human genome, demonstrate that at least a 2-fold reduction in the number of sequence reads required to sequence an individual genomic clone (cosmid, PAC, etc.) can be achieved. Treatment of clone contigs with significant clone overlaps will allow an even greater reduction.

INTRODUCTION

Since the foundation of the Human Genome Organisation (HUGO) in 1989 (1) only a few per cent of the human genome have been sequenced (S.Beck, http://www.ebi.ac.uk/~sterk/genome-MOT/). Although the sequencing rate has increased dramatically during that period, completion of the project by 2005 will require either appropriate increases in funding or the use of new methods (2,3).

In spite of a number of alternative proposals for directed sequencing strategies, like deterministic sequencing (4), transposon-facilitated sequencing (5–8), primer walking and primer ligation (9), most sequence information has been generated by traditional shotgun sequencing. As an inherent part of this method longer sequences have to be subdivided into shorter, overlapping sequence stretches. If that subdivision is random, as in the case of traditional shotgun sequencing, an unequal representation of different parts of the sequence will be expected due to sampling effects, requiring oversampling to ensure a minimal coverage of under-represented regions. This situation can be made considerably worse by biological effects, e.g. different cloning efficiencies of different sequence stretches. Typically >2000 sequence reads/100 kb are generated from randomly chosen shotgun clones and assembled. In most cases a number of gaps remain in the consensus sequence after that pure shotgun phase, as well as regions of weak data quality, and subsequently directed approaches, e.g. primer walking, are used as the finishing step. Completed shotgun projects show an 8–12-fold average coverage per base final sequence, which is significantly more redundant than necessary to achieve consensus sequence data of sufficient quality. In addition, it is a common situation in large-scale sequencing projects that the target region is spanned by overlapping genomic clones (cosmids, PACs, etc.) and it is often difficult to find a set of those clones which cover long sequence stretches with a minimal amount of overlap. The resulting redundancy in the overlapping regions is twice as high as in the non-overlapping regions.

As a very useful advance, a subset of shotgun clones with no or little overlap can be selected from shotgun libraries, using automated facilities (10) to generate and analyse high density filter arrays. A sampling without replacement method was introduced by Hoheisel et al. (11) and applied to shotgun clone selection by Scholler et al. (12). In this strategy individual clones or pools of clones of fixed length are used as hybridization probes. The number of experiments (clone–probe tests) is therefore proportional to \(N^2\), the square of the number of clones analysed in each individual shotgun library. If clone pools are used as hybridization probes, the effort is reduced by a constant factor. The approach requires the generation of new probes for each new library and therefore requires a quite significant upstream effort. Moreover, it will often have difficulties with repeat sequences in the probes and the procedure works sequentially. The result of one hybridization experiment has to be analysed before the next one can be carried out.

*To whom correspondence should be addressed. Tel: +49 30 84131203; Fax: +49 30 84131380; Email: radelof@mpimg-berlin-dahlem.mpg.de
As an alternative approach we describe in this paper a method called preselection by oligonucleotide fingerprinting (PfOF). This method is based on the use of short oligonucleotides as hybridization probes on high density clone filter grids, a strategy to characterize clones and clone overlaps (13). It can be used both in the establishment of clone maps of large insert clones (14) and in the characterization of much shorter clones, e.g. cDNA clones (15–19).

As has been pointed out before, oligofingerprinting in itself can be considered as a kind of sequencing technique. There have even been proposals to use it as an approach to determine sequences completely (20–29). In fact, oligofingerprinting is particularly well suited to be used in combination with gel-based sequencing techniques. There exists a continuum from gel-based sequencing of randomly chosen shotgun clones without any preselection to sequencing by hybridization. Weaknesses in both approaches can be overcome by a powerful combination of the two.

MATERIALS AND METHODS

Generation of shotgun libraries

PAC DNA is prepared as described (30), purified by alkaline lysis and caesium chloride banding and then sheared by sonication. The resulting DNA fragments are end repaired, size selected, ligated into Smal-digested and dephosphorylated pUC18 vector and transferred by electroporation into Escherichia coli (strain KK2186). The bacterial suspension is plated out on 22 cm × 22 cm LB–agar plates containing ampicillin, X-gal and IPTG. Plates are afterwards incubated for 12 h at 37°C and stored for better development of the blue colour for 24 h at 4°C. Well-separated, white colonies are picked by a robotic picking system (Genetix or Linear Drives) originally developed by us (31,32). For each 100 kb to be sequenced, ~2600 colonies are picked. About 3000 colonies/h are transferred into 384-well plates containing 2YT medium, ampicillin and HMFM freezing solution. After incubation at 37°C overnight, plates are replicated and stored at −80°C.

Generation of PCR products

The hybridization of short oligonucleotides requires highly purified target DNA. This is generated by an automated PCR approach on several shotgun libraries in parallel. PCR amplifications are carried out in 384-well microtitre plates (Genetix), in an automated waterbath system developed in-house, allowing up to 51 840 PCR amplifications/run. Using disposable plastic 384-pin inoculation devices (Genetix), a small amount of the bacterial suspension is added to a 40 µl reaction volume containing 50 mM KCl, 10 mM Tris–HCl, pH 8.5, 1.5 mM MgCl2, 200 µM dNTPs, 100 ng of each PCR primer (M13 forward, 32mer, gettacgcagcttgcaaggggaggtg; M13 reverse, 32mer, cccagccgctagccctgtcgcctg) and 0.5 U Thermus aquaticus (Tag) DNA polymerase. After inoculation, the microtitre plates are sealed with a 0.45 µm thick plastic foil with a heat sealer designed for this purpose (Genetix). PCR is performed for 30 cycles consisting of 10 s at 94°C, 10 s at 65°C and 4 min at 72°C.

Spotting of PCR products

High density filter arrays of PCR products from shotgun clones are generated robotically as described previously (19). Each 22 cm × 22 cm nylon membrane carries 27 648 different clone spots as duplicates and in addition 2304 spots of genomic salmon sperm DNA. These spots yield signals in every oligohybridization experiment and are necessary as guide spots for the automated image analysis. To obtain a quality assessment of the hybridization data, PCR products from previously sequenced shotgun clones are spotted on each filter. The hybridization signals of these clones can thus be directly compared with those predicted from the DNA sequences. Twenty filter copies are prepared for parallel hybridization experiments.

Oligonucleotide hybridization

Using a computer program developed in-house, a set of 100 8mer oligonucleotides, best suited for characterization of genomic DNA, were selected out of a set of >250 oligonucleotides used in our laboratory for characterization of cDNA libraries. The selection is based on the following idea. The highest information value of a single hybridization experiment could be achieved using an oligonucleotide (or even a pool of different oligonucleotides) that has a hybridization probability of 50% to all clones in the shotgun libraries in question. Therefore this probe divides all clones into two partitions of the same size (clones with/without a hybridization signal). The ideal set would consist of probes each having that hybridization probability. In addition, every single probe would, together with a second one, divide all clones into four partitions of the same size and together with a third one into eight partitions of the same size, etc. In practice we use several megabases of genomic sequences from commonly available databases, cut them into pieces of typically sized shotgun clones, e.g. 1–2 kb, and search for the 8mers with maximal hybridization probabilities. Following the ideas mentioned above, the computer program selects oligonucleotides, from scratch or from an already existing oligonucleotide collection, which partition the whole simulated clone set as equally as possible.

Since 10mers hybridize more reliably than 8mers, each probe in reality comprises a pool of all 16 10mers sharing the same 8mer core sequence with N residues at the 3′- and 5′-ends (NXXXXXXXXX). The oligonucleotides are labelled at the 5′-end by a kinase reaction using [γ-33P]ATP (Amersham International) and T4 polynucleotide kinase (New England Biolabs). Each probe is used in a separate hybridization experiment. Using 20 filter copies 20 hybridizations are carried out in parallel. The hybridizations are performed overnight at 4°C in hybridization bottles (built in-house) containing 12 ml 600 mM NaCl, 60 mM sodium citrate, 7.2% Na Sarkosyl with a probe concentration of 2.5 nM. Afterwards, 10 filters are washed at a time in 1 l of the same buffer for 20 min at 4°C. To evaluate the total amount of DNA which has been spotted for each clone on the filter, one additional hybridization is carried out with an 11mer oligonucleotide matching the plasmid vector sequence common to all PCR products.

The intensities of the hybridization signals are measured by phosphor storage autoradiography (Molecular Dynamics, Sunnyvale, CA). The system is at least 10 times more sensitive and faster than conventional film-based autoradiography and allows linear measurement of the hybridization signal over a larger range (33). The PhosphorImager scans have a 16 bit grey scale resolution and a resolution of 88 or 176 µm/pixel. The result is subsampled to an 8 bit 1024 × 1024 image. It requires ~5 min to scan a 22 cm ×
22 cm hybridization image, allowing the subsequent scanning of many filter images a day.

Re-arraying

Clones selected for sequencing are collected with a re-arraying robot and forwarded to our in-house sequencing unit. The robot, developed in our department, routinely re-arrays >600 clones/h without cross-contamination and with a yield of >97%, i.e. <3% of the bacterial clones fail to grow in the daughter plates.

Sequencing

The sequencing reactions are carried out using the dye primer technique on an ABI catalyst robot using 1 µl of the PCR product and 3 µl of the ThermoSequenase mix (Perkin Elmer) for each of the four A, C, G and T reactions. Energy transfer primer (0.1 pmol for A and C and 0.2 pmol for G and T reactions, respectively) M13(–40) or M13(–28) is added to the ThermoSequenase mix before starting the sequencing run. Samples are pooled and precipitated according to ABI’s instructions and analysed on ABI 377XL DNA sequencers. Data are processed using ABI's sequence analysis software v.3.0 and v.3.1, but with the Perkin Elmer manual lane tracking kit according to the manufacturer’s instructions.

Image analysis

Hybridization images obtained from the PhosphorImager are transferred to a DEC alpha UNIX workstation. An image analysis program determines raw hybridization intensities for each clone and probe and subtracts the average background from the signals. A normalization routine compensates for (i) different overall hybridization intensities (maxima and minima) from different clones and (ii) different masses of different clones. The final output is a hybridization matrix containing normalized intensities for all clones and probes. An example is given in Table 1. Each row of this matrix represents the oligofingerprint of one clone. Programs for hybridization data analysis on high density matrices were written in our laboratory and are still under development.

Preselection

The aim of the preselection is to avoid unnecessarily high sequencing redundancy. Therefore we search for shotgun clones representing a minimum tiling path along the pool of more or less randomly distributed shotgun clones representing the entire sequence of the original genomic clone. The clones required have minimal sequence overlaps, indicated by maximally dissimilar hybridization patterns.

Single clones can be identified by their fingerprint vector \( \mathbf{F}_N \), which contains the hybridization intensity for oligos \( j = 1, \ldots, K \) on clone \( N \). A simple measure for the similarity of two vectors is their scalar product

\[
S_{NM} = \mathbf{F}_N \cdot \mathbf{F}_M = \sum_{j=1}^{K} F_{Nj} \cdot F_{Mj}
\]

Two vectors (clones) can be regarded as maximally dissimilar if \( S_{NM} = 0 \), i.e. they have no oligonucleotide match in common, and as maximally similar if \( S_{NM} = 1 \) (for normalized fingerprint vectors).

Once the scalar product for each clone pair is calculated the construction of a low redundancy set can be done using the following series of steps:

(i) start with an arbitrary clone
(ii) add to selected clone set
(iii) find the clone with minimal scalar product to all clones in selected clone set
(iv) go to (ii)

The selection of a typically sized set from a shotgun library containing 2600 clones for a 100 kb PAC is completed in a few minutes on a standard UNIX workstation.

Table 1. Excerpt of a typical fingerprint matrix containing the hybridization intensities of each clone and probe (oligonucleotide)

<table>
<thead>
<tr>
<th>Oligo 1</th>
<th>Oligo 2</th>
<th>Oligo 3</th>
<th>Oligo 4</th>
<th>Oligo 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00000</td>
<td>2.87352</td>
<td>0.00000</td>
<td>3.21158</td>
<td>0.00000</td>
</tr>
<tr>
<td>Clone 2</td>
<td>Clone 3</td>
<td>Clone 4</td>
<td>Clone 5</td>
<td>Clone 6</td>
</tr>
<tr>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>2.53546</td>
<td>0.00000</td>
</tr>
<tr>
<td>Clone 7</td>
<td>Clone 8</td>
<td>Clone 9</td>
<td>Clone 10</td>
<td>Clone 11</td>
</tr>
<tr>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>3.38061</td>
</tr>
<tr>
<td>Clone 12</td>
<td>Clone 13</td>
<td>Clone 14</td>
<td>Clone 15</td>
<td>Clone 16</td>
</tr>
<tr>
<td>1.18321</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Clone 17</td>
<td>Clone 18</td>
<td>Clone 19</td>
<td>Clone 20</td>
<td></td>
</tr>
<tr>
<td>0.00000</td>
<td>3.04255</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data are filtered with respect to background noise and are normalized.

RESULTS

Simulation experiments

Different computer simulations were carried out in order to compare the efficiency of the preselection under various conditions with the standard shotgun approach. The influence of the shotgun clone insert size, the insert size distribution and the repeat content of the genomic region in question have been investigated. For this purpose, arbitrarily chosen human genomic sequences of 100 kb length were extracted from a publicly available database (http://www-eri.uchsc.edu/chr21) and randomly cut into pieces of typical shotgun clone sizes. Some arbitrarily chosen areas were set to over- or under-represented regions based on typical assemblies of sequenced shotgun libraries. Each virtual shotgun library consisted of 2000 clones. Theoretical oligofingerprints
were generated using the same set of 8mer oligonucleotides applied in the real experiments. Hybridization ‘intensities’ were set to 1 in cases where the oligonucleotide sequence matched the clone sequence and to 0 otherwise. The real situation is more complicated since 7 (one mismatch) and even multiple 6 (two mismatches) matches yield strong signals (M.Wiles, personal communication) and fractional numbers of signal intensities are used.

In all simulations, shotgun clones were selected using the selection algorithm given above. The same numbers of clones were taken by a random process simulating shotgun sequencing. All clones selected were ‘virtually’ sequenced from both sides with a read length of 600 bases. After assembly, the consensus sequence was measured and compared (Figs 1–3). Each point in the curves represents an average value of 50 statistically independent selected clone sets.

In the first simulation experiment (Fig. 1) we examined the influence of the repeat richness of the genomic region (cosmid, PAC, etc.) to be sequenced. For this we used a 100 kb database sequence with a repeat content of 52% (ALU, LINE, MER, etc.) in comparison with an artificially repeat-free sequence of the same length. This sequence was constructed by combining several repeat-masked database sequences. In both cases, shotgun clones of fixed size (1.5 kb) were used.

In the second experiment (Fig. 2), the same 52% repeat sequence as above was ‘shotgunned’ into clones of either fixed or Gaussian distributed insert length.

In the third experiment (Fig. 3), again the 52% repeat sequence was used to consider the impact of the shotgun clone insert size using shotgun clones of different but fixed sizes.

The differences in efficiency of the PrOF method in all test cases are very small, indicating that the influence of these parameters is weak, and demonstrating the robustness of the fingerprinting approach. In the region of 97% coverage of the entire genomic sequence, where ‘gap closure’ usually starts, the PrOF approach required much less than half the number of sequence reads compared with random selection in all cases considered.

A few remarks should be made concerning the influence of repeat regions on PrOF efficiency. ALU repeats, which are the most frequent repeat family in human genomic DNA, have a length of 300–400 bp. Since typical shotgun clone sizes are in the range 1–2 kb, there is enough sequence information left to distinguish clones from different ALU-containing regions by their fingerprints, as long as sufficient oligos are used. LINE elements, which form another frequent human repeat family, can be found in various lengths of up to 7 kb. However, the large diversity of LINEs along the human genome makes it possible again to distinguish the respective clones of different LINE regions by their fingerprint, especially when clones only partly cover a LINE region. On average, a high repeat content of a genomic region will slightly reduce the effectiveness of the PrOF method, as shown in Figure 1. A severe problem, of course, would arise for genomic sequencing projects covering duplicated regions of many kilobase lengths. In principle, there is no easy way to map clones back to the correct region solely by their fingerprint in those cases. However, when working on single cosmid or PAC/BAC libraries, it is unlikely to run into those problems and it should be emphasized that the simple shotgun approach would face the same problem in the sequence assembly step.
Pilot experiment

In order to test the efficiency of the PrOF strategy for handling experimental data, we used an already sequenced cosmid shotgun library containing ~40% repetitive sequences (ALU, MER, etc.). Figure 4 shows the assembly of 426 clones covering a consensus sequence of ~45 kb. The assembly does not contain the finishing data produced by primer walking. Large fluctuations in coverage clearly reflect a situation typical in shotgun projects, with regions both heavily over- and under-represented and even gaps in the consensus sequence due to statistical and biological effects.

In the conventional shotgun approach a large number of randomly chosen clones are sequenced in order to increase the probability of obtaining sequences in under-represented regions. However, this strategy also increases the mean coverage to unnecessarily high values. In our example the average coverage is 11-fold, with maximal local coverage around 30-fold. The generation of so many sequence reads and the additional gap closure makes the process much more expensive than it need be, blocks sequencing capacity and wastes time.

All shotgun clones of this library were PCR amplified, spotted on filters and oligofingerprints were created as described above. As a quality check of the experimental fingerprint data, we compared the calculated similarity of the clones using hybridization data with the real clone overlap detected by sequencing. The observed relationship is nearly linear, as shown in Figure 5.

For a direct comparison of the PrOF approach with the random approach used in the standard shotgun procedure, we selected certain numbers of clones out of the same clone pool either based on oligofingerprints or randomly (Fig. 6). Again, as in the simulations, in the region of ~97% coverage, the PrOF method is ~2-fold more effective than random selection (Table 2).

Each point of the curves in Figure 6 represents an average of 50 statistically independently selected clone sets. In each single experiment, a different result is achieved. In one experiment possibly 300 reads are needed to achieve 97% coverage, while in another 270 or 330 could be necessary to cover the same consensus sequence. The range of variation at a fixed set size is given in Figure 7 for both methods. The PrOF method clearly shows a much narrower variation. The certainty of getting a specific coverage in a single experiment is much greater in comparison with the random approach.
Figure 8. Graphical representation of the number of reads (x-axis) in the same order as they were actually selected and sequenced. The percentage of the genomic region covered by the respective number of reads is given on the y-axis.

Table 2. Number of reads required to gain a certain percentage of the genomic sequence covered are given for the PrOF approach and random selection

<table>
<thead>
<tr>
<th>Coverage (%)</th>
<th>Random (reads)</th>
<th>PrOF (reads)</th>
<th>Random/PrOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>286</td>
<td>164</td>
<td>1.74</td>
</tr>
<tr>
<td>96</td>
<td>542</td>
<td>248</td>
<td>2.18</td>
</tr>
<tr>
<td>97</td>
<td>588</td>
<td>276</td>
<td>2.13</td>
</tr>
<tr>
<td>98</td>
<td>685</td>
<td>364</td>
<td>1.88</td>
</tr>
</tbody>
</table>

Ratios of reads required are also shown.

**Application in large-scale sequencing**

Currently we are applying the preselection strategy to a large-scale sequencing project spanning a 1.5–2 Mb region of the 17p11.2 region of the human genome. In the first experiment we are using five shotgun libraries derived from PACs between 70 and 130 kb in size, 535 kb in total. All amplified clones are spotted on one filter (20 filter copies). In addition clones from four already sequenced cosmid-derived libraries are spotted on the same filter as controls. After the hybridization of 100 oligonucleotides (20 in each step in parallel, using 20 filter copies) and the computational analysis of 82 hybridization images (18 low quality images rejected), the selected clones were robotically re-arrayed and sequenced from both sides.

In four out of five preselection projects, we obtained almost the same results as in the simulations and the pilot experiment. Figure 8 depicts the results from three of these projects in direct comparison with three typical shotgun projects (also PAC-derived) carried out in this laboratory. In order to normalize the results to a common scale, the number of all sequence reads is divided by the respective PAC size and multiplied by 100 kb. Again, as is shown in Table 3, in the projects where the PrOF strategy was used, only half the number of sequence reads are necessary, compared with the standard shotgun projects, to get the same consensus sequence length.

<table>
<thead>
<tr>
<th>Coverage (%)</th>
<th>Shotgun (reads)</th>
<th>PrOF (reads)</th>
<th>Shotgun/PrOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>771</td>
<td>416</td>
<td>1.85</td>
</tr>
<tr>
<td>96</td>
<td>1132</td>
<td>581</td>
<td>1.95</td>
</tr>
<tr>
<td>97</td>
<td>1263</td>
<td>614</td>
<td>2.05</td>
</tr>
<tr>
<td>98</td>
<td>1523</td>
<td>677</td>
<td>2.25</td>
</tr>
<tr>
<td>99.5</td>
<td>2003</td>
<td>851</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Ratios of reads required to cover the same consensus sequence length are also shown.

**DISCUSSION**

We describe here the first results of a powerful combination of oligonucleotide fingerprinting and shotgun sequencing. To select optimal sets of shotgun clones prior to sequencing, clones from shotgun libraries could be ordered into contigs, based on the results of an oligofingerprinting experiment (13). This however requires an unacceptably large number of hybridization experiments and would partly generate information on exact overlaps between clones, which is then independently generated again in the sequencing procedure. Thus we use fingerprinting data to achieve a considerably more modest goal, the definition of a minimally overlapping set of clones based on their hybridization patterns (oligonucleotide fingerprints). Sequences are then assembled by the standard analysis tools for shotgun data.

Sequence information generated and oligofingerprinting results can be combined to select clones in regions of weak quality sequence data, for bridging or extending into gap regions, and can therefore aid in gap closure. In this step, we can take advantage of already sequenced clones and the control clones with known sequence to determine effective recognition patterns of the oligonucleotides, which in practice can include specific types of mismatches, especially end mismatches, though with reduced signals (data not shown). This application is at the moment mostly limited by the software, which is being continuously improved. However, we do not expect to do the finishing without any primer walking. Gap regions often contain ‘unusual’ sequences, such as sequences of low complexity, e.g. simple repeats. Those sequences may be not only reduced in the shotgun libraries, but completely absent.

Even with the simple analysis software used in the experiments described, the PrOF approach has resulted in significant cost reductions and throughput improvements in large-scale sequencing. We have been able to demonstrate both in simulations and large-scale experiments that the number of clones to be sequenced in shotgun projects can be significantly reduced. The reduction can be increased further if genomic regions spanned by overlapping genomic clones are being sequenced, because shotgun clones are distinguished solely by their oligofingerprint and selected with the same average redundancy in the overlap region of two libraries as for the non-overlapping regions.

The method has proved to be more efficient than a sampling without replacement strategy due to a more favourable scaling behaviour (NlogN instead of N²), the use of a standard set of probes for all experiments and, as shown in this paper, a reduced sensitivity to the effect of repeat-rich genomic regions, shotgun clone insert sizes and insert size distributions.
A main advantage of our protocol is the rapid handling of many shotgun libraries in massively parallel experiments. Moreover, once the technical facilities required are available in a sequencing laboratory, the preselection costs, including all materials (as described above) and salaries, are ∼5% of the cost of traditional shotgun sequencing if one filter (capacity ∼900 kb) is handled as in the experiment described here. However, the cost per filter is greatly reduced if multiple filters are handled in parallel. In our cDNA projects, four different filters are routinely hybridized in one hybridization bottle, using the same amount of chemicals as used here for one filter. It is feasible for one skilled person to perform the oligofingerprinting of batches of shotgun libraries representing a total sequence length of >3.5 Mb in parallel within 2 months, including all working steps from the initial PCR to the re-arraying of the selected clones. This additional effort and cost at least doubles the sequencing throughput independent of the sequencing technology used, because less than half the number of clones now have to be sequenced.

We expect the technique to also be useful in very large-scale sequencing projects, e.g. in whole genome shotgun sequencing projects proposed for the human genome by Weber et al. (3) and planned now by Venter et al. (3) after criticism by Green (35). To be able to approach such large projects, further improvements in the software, but also in the throughput of the oligofingerprinting pre-screening (clone picking, PCR, spotting and hybridization, e.g. use of fluorescently labelled oligonucleotides and fully automated hybridization) will still be helpful. A number of automated improvements in this direction are currently under development.

ACKNOWLEDGEMENTS

The excellent technical support by Christina Steffens and Susanne Jung is gratefully acknowledged. Discussions with Sebastian Meier-Ewert always revealed deep insights into the problems of oligofingerprinting technology. Without the strong support of the automation and informatics group of our department this work would not have been possible. We also thank Michael Wiles, John O’Brien, Georgia Panopoulou, Matthew Clark, Leo Shalkwyk and Holger Eickhoff for critical reading of the manuscript and valuable comments. The project was funded by a contract with Dr Karl Thomae GmbH.

REFERENCES