Molecular indexing of human genomic DNA

D. Ross Sibson* and Fiona E. M. Gibbs

J. K. Douglas Laboratories, Clatterbridge Cancer Research Trust, Clatterbridge Hospital, Bebington, Wirral CH63 4JY, UK

Received May 9, 2001; Revised and Accepted August 8, 2001

ABSTRACT

Molecular indexing sorts DNA fragments into subsets for inter-sample comparisons. Type IIIS or interrupted palindrome restriction endonucleases, which result in single-stranded ends not including the original recognition sequence of the enzyme, are used to produce the fragments. The ends can then be any sequence but will always be specific for a given fragment. Fragments with particular ends are selected by ligation to a corresponding indexing adapter. We describe iterative indexing, a new process that after an initial round of indexing uses a Type IIIS restriction endonuclease to expose additional sequence for further indexing. New plasmids, pINDnn, were produced for novel use as indexing adapters. Together, the plasmids index all 16 possible dinucleotides. Their large size can be increased by dimerisation in vitro and allows the isolation of indexed material by size separation. Fragments produced from human genomic DNA by Type II restriction endonucleases were sorted using six bases in total to a possible enrichment of 1920-fold. By comparison with the public human sequence databases, fidelity of indexing was shown to be high and was tolerant of repetitive sequences. Genome-wide comparisons on a candidate or non-candidate basis are made possible by this approach.

INTRODUCTION

There is a great deal of interest in comparing sequences found in different situations for variations in copy number, internal variation or modification to gain insights into phenotypic differences. Large-scale comparisons have the greatest power. Global patterns of gene expression have been compared by a variety of approaches, including differential display (1), microarrays (2), massively parallel sequence signatures (3) and serial analysis of gene expression (4). Microarrays have also been used for measuring variations in genomic copy number (5) and scanning for sequence differences (6). Genomic DNA has great complexity. It is therefore difficult to scan for and isolate variation between different sources. Molecular indexing of DNA was first developed for this purpose (7), but its usefulness has been restricted to comparing global patterns of gene expression (8–12).

Fragments for indexing are most conveniently produced by Type IIIS or interrupted palindrome restriction endonucleases. The activity of these enzymes leaves single-stranded overhangs which do not overlap the original recognition site. Therefore, the overhangs can be any combination of bases but will always be the same on a given fragment. Adapters with complementary ends can be designed to ligate to any sequence of interest so that it can be isolated by solid phase capture and/or PCR.

A limitation of indexing has been its fidelity. For example, there are 4^6 possible sequences of 6 bases in length which could be used to sort fragments into a corresponding number of subsets with concomitant enrichment of the fragments in the subset. Any loss of fidelity, for example on ligation of the indexing adapters or during purification of the indexed fragments, will reduce the enrichment. The fidelity of ligation decreases as the length of the overlapping regions to be joined increases (13,14). However, the information in longer regions must be accessed to achieve high levels of indexing. Type IIIS restriction endonucleases will cut from an adapter into adjacent sequence (15). This has been exploited for serial analysis of gene expression (SAGE) and for sequencing (3,4,16). We therefore used indexing adapters that contain the site for a Type IIIS restriction endonuclease so that on ligation to fragments of interest, further sequences of the fragments could be exposed by cleavage with the enzyme. Multiple rounds of indexing short overhangs could thus be performed to the desired extent. There was also an added benefit that the process could be started at the overhangs produced by Type II restriction endonucleases.

Indexing adapters are usually biotinylated to allow their capture by streptavidin-coated beads. However, restriction endonucleases cannot easily remove indexed fragments from the beads. Novel capture was therefore achieved by constructing a plasmid for use as the first indexer. The plasmid with any fragments that had been indexed could then easily be purified from non-indexed fragments by size separation. We report here the results of indexing human genomic DNA that had originally been digested by the Type II restriction endonucleases BamHI and BgIII. Three rounds of indexing, in which each indexer selected for a two-base sequence, were used to access the information in six possible bases. Combining the information obtained during each round of indexing determines the sequence of the indexed part of the selected fragments. The use of fragments produced by the Type II restriction endonucleases BamHI and BgIII suggests that the approach will be applicable to virtually any fragments.

© 2001 Oxford University Press

*To whom correspondence should be addressed. Tel: +44 151 343 4303; Fax: +44 151 343 1820; Email: ross@ccrt.co.uk
of interest from a complex nucleic acid or nucleic acid population.

MATERIALS AND METHODS

Restriction digests
Restriction endonucleases were purchased from New England Biolabs (NEB) and used in accordance with the recommendations of the manufacturer.

Ligation
Ligations were performed with 0.1 U ligase (NEB) per 10 µl for 16 h at 16°C unless stated otherwise.

DNA purification
Except where indicated, fragments of DNA and PCR products were purified by ion exchange chromatography (Qiagen) according to the manufacturer’s recommendations.

PCR
PCR for plasmid modification and plasmid screening was performed in all cases with 0.1–1 ng pUC19 at 94.5°C for 5 min, followed by 32 cycles of 94.5 and 65°C for 30 s each and 72°C for 1 min. A final incubation of 72°C for 10 min was performed. Reactions (50 µl) contained 0.2 mM dNTPs, 25 pmol each primer, 2.5 U AmpliTaq Gold (Perkin Elmer) and 2.5 mM MgCl2.

Indexing vectors
The plasmids pINDaa–pINDtt were produced for indexing as described below. Indexing plasmids were cut to completion with BamHI and BsrDI.

Escherichia coli
Escherichia coli XL-1 Blue (Stratagene) was transformed by the CaCl2 method and recombinants selected using 50 µg/ml ampicillin with 80 µg/ml IPTG and 50 µg/ml X-gal.

Molecular indexing
A complete BamHI and BglII double digest of 10 µg human genomic DNA (Sigma) was purified and ligated to a 30-fold excess of the adapter 5′-ccagtcgcaggtctcaagctcgatccctggagc and its complementary strand 5′-gatcgctccagggatcgagcttgagacctgagctgcagctcgacctggcagtcttca, transformed and purified as described for the pINDnn vectors. The modified vector was cut with BamHI and BbsI. Fragments that had been PCR amplified and indexed were purified and incubated at 1 µg per 10 µl with 3 U T4 DNA ligase and 0.2 mM dATP to resect back the four bases at their 3′-termini. Reactions were at 37°C for 30 min. Fragments were re-purified, cut with BamHI, purified and ligated to 1 µg prepared vector. Ligations were used for transformation.

Sequence analysis
Transformants were picked directly into PCR reactions which were performed as described except that they used the M13 forward and reverse primers at 20 pmol per 50 µl. PCR products were purified and sequenced using dye terminator chemistry with the megaBACE sequencer (Amersham Pharmacia Biotech). The Staden programs were used to remove the vector and cloning adapters and non-recombinant sequences.

RESULTS AND DISCUSSION

Construction of plasmids for use as indexing adapters
The parent plasmid for all of the vectors was pUC19. We wished to replace its polylinker with one containing unique, adjacent sites for the Type IIS restriction endonucleases BsrDI and BpmI. The former was used to produce a two-base overhang corresponding to the indexing end and the latter for cutting into the indexed fragments following capture. First, the two sites for BsrDI and one for BpmI in the ampicillin gene were removed by using PCR to introduce neutral substitutions. Separate PCR products were produced for each change and then joined by recombination PCR (Fig. 1). Flanking XmnI and AflIII sites were used to replace the corresponding region in the
original plasmid. Separate double digestions of candidate modified plasmids with BsrDI or BpmI each with XmnI were performed and the products analysed by agarose gel electrophoresis to determine the positions of any differences compared to pUC19. Plasmids lacking some or all of the original BsrDI and BpmI sites were observed. One of these plasmids, pIND10, having lost all three sites, was used for further work.

The polylinker of the pIND10 vector was removed by digestion with EcoRI and HindIII and replaced by an adapter containing the recognition sites for the restriction enzymes BsrDI and BpmI of general design 5′-aattctggagmcattgcgcacaagtcc and the complementary sequence 5′-gcggtagctcgttcagctggcg, having a 5′-overhang produced by HindIII, and ligating the adapter. Each indexer was specific for a two-base sequence to achieve selection of six bases in total. We used T4 DNA ligase, having already indexed the BpmI fragments of bacteriophage λ with high fidelity (data not shown). Taq DNA ligase may have greater fidelity but is less efficient when the overlapping regions to be joined are short (21). Blocking adapters having in combination all ends except those of the indexers were used to prevent non-indexed ends from joining so that chimeric fragments which may participate in indexing did not result. The second and third indexers but not the plasmid were used for final selection by PCR. In this way only fragments having been indexed by all three indexers should have been selected. Subsets were produced using the indexers listed in Table 1. A BamHI site in the second indexer was used for cloning or labelling of the indexed material. Indexed material was fluorescently labelled by digesting the DNA with BamHI and ligating the adapter 5′-gatcgtcgtcgtcgtcg, having a 5′-FAM label, and its complementary strand 5′-aagtcgtcgtcgtcgtcg, leaving a 5′-gatc overhang. Discrete peaks characteristic of the subsets were reproducibly observed when the material was analysed on a megABACE capillary electrophoresis system (data not shown).

**Validation of indexing by sequence comparison**

Male and female human genomic DNAs were indexed independently. We wished to compare fragments in the subsets with the available human sequence to confirm that the fragments had been indexed correctly. Fragments from each subset were therefore independently cloned and clones picked at random for sequencing. Sequences are available via our Web site (www.ccrf.org.uk). The sequences were compared with the available human sequence to determine whether they had been successfully indexed. The existing human sequence was
assumed to be correct even when known to be unfinished. The organisation of the human sequence from which a correctly indexed fragment had been obtained could be predicted (Fig. 3). We expected sites for one of the restriction endonucleases BamHI, BglII or BpmI to be found at each end of the sequence corresponding to the correctly indexed fragment in the genomic sequence to which it matched. The expected first and second pairs of indexed bases should be eight bases from a BamHI or BglII site or 14 bases from a BpmI site. Twelve bases should separate the sites for the first and third pairs of indexed bases. Failure to index correctly is indicated by any combination of: (i) unexpected bases at the site of those indexed; (ii) incorrect location of the indexed bases; (iii) the presence of any other bases at the site of those indexed.

![Figure 2. (A) Dimerisation of plasmid pINDnn for use in indexing. (B) Plasmid pINDnn-mediated indexing of genomic DNA.](image)

![Figure 3. Representative example of correctly indexed genomic DNA and flanking sequences. The figure shows all possible configurations of indexing. A BpmI indexing adapter ligated to the BglII site exposes two bases for indexing which are eight bases from the end of the BglII site. The first indexer ligated at these bases contains a BpmI site which allows two more bases to be exposed for indexing 12 bases along. The BpmI site in the genomic DNA was cleaved at the first BpmI digest and exposes two bases for indexing 14 bases from the site. Restriction sites and indexed nucleotides are shown in bold.](image)
Tables 2 and 3 show examples of correctly indexed fragments of human genomic DNA. The original BamHI/BglII or BpmI sites are shown in bold. Correctly selected 2-base overhangs are also shown in bold.

<table>
<thead>
<tr>
<th>Accession no.</th>
<th>Start of sequence match</th>
<th>Sequence showing correct indexing at 5' region of match</th>
<th>Sequence showing correct indexing at 3' region of match</th>
<th>End of sequence match</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC004882</td>
<td>161126</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>161377</td>
</tr>
<tr>
<td>AC008464</td>
<td>104583</td>
<td>TGTTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>104812</td>
</tr>
<tr>
<td>AC018371</td>
<td>84110</td>
<td>TGGTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>84351</td>
</tr>
<tr>
<td>AC024719</td>
<td>92331</td>
<td>TGGTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>92552</td>
</tr>
<tr>
<td>AC067889</td>
<td>38557</td>
<td>TGGTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>38720</td>
</tr>
<tr>
<td>AL035690</td>
<td>76650</td>
<td>TGGTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>76865</td>
</tr>
<tr>
<td>AL157397</td>
<td>90001</td>
<td>TGGTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>90524</td>
</tr>
<tr>
<td>AP002781</td>
<td>18977</td>
<td>TGGTCTCTTTAACTGTAATTTCTCTTACTCCTCTCTCT</td>
<td>CTGGTAGGAATATTTCTGGAAGACCTGGTAGTTCTCT</td>
<td>19357</td>
</tr>
</tbody>
</table>

The fragments in Table 4 were classified where possible according to the relative success of their indexing and also according to the arrangement of known repetitive sequences in the fragments. The details for a clone corresponding to each general classification are listed in each case.

Indices were observed. Those with the same general classification are listed in each case.

Observed enrichment comparing likely successes and failures based on the database matches was calculated as 686- to 1411-fold (Table 5, column A) and the overall efficiency ranged between 35.7 and 76.3% (column B). Accurate ligation of indexers was observed for between 68.4 and 99% of the events depending on the indexed population (column C). A population of up to 21% of the fragments, dependent on the subset, had one of their correctly indexed sites at one base distant from that predicted for BpmI (column E). The majority of these had been indexed correctly in all other respects. We believe that cleavage by BpmI had occurred at 17/15 because when the site of the first indexer was affected the position of the third indexer was always found at the new expected site. Displacement of the expected cleavage site has been reported for certain Type IIS restriction endonucleases which are thought to measure distance along the DNA rather than the actual number of bases (22). We are examining subsets in which the observed fragments would be expected. It can then be determined whether such events are rare, selected by the process, or whether they are a common phenomenon. Cleaving at 17/15 was not observed when the residual BamHI site intervened, suggesting a sequence-specific phenomenon, but no pattern was obvious. We would have reported higher efficiencies of indexing if we had included in the total of correctly indexed fragments the fragments for which a one base slip had occurred in the location of the BpmI DNA. The fragments in Table 4 were classified where possible according to the relative success of their indexing and also according to the arrangement of known repetitive sequences in the fragments. The details for a clone corresponding to each general classification are listed in each case.

Uncut restriction sites; (iv) the absence of expected restriction sites in the human sequence but presence of the indexing adapter or the PCR primer alone in the clone. PCR priming on an internal part of a fragment is inferred by (iv). Results for the indexer combinations FG1 and FG2 are shown in Tables 2 and 3, respectively. The first subset produced eight fragments that had been indexed entirely correctly. Three fragments had an indexer missing at one of their ends and presumably arose by PCR priming not from an indexer but within the internal region of match. A further 20 had no human match or had an ambiguous origin. This corresponds to an enrichment of 612-fold. Interestingly, two of the fragments lacked an indexer, which suggested internal priming, and the remainder had an ambiguous origin. This corresponds to an enrichment of 612-fold. Interestingly, two of the fragments had been indexed correctly if it was assumed that BpmI had cut one base further on than expected (Table 3, AC020705 and AC023282). Similar results were obtained for the remaining FG subsets.

Having found correctly indexed fragments in the FG subsets, further subsets were analysed to determine their indexing efficiencies. The results for the indexer combination first TT, second AG, third GT with male genomic DNA are shown in Table 4 and summarised in Table 5 for this and other combinations of indexers with independent samples of male and female DNA. The fragments in Table 4 were classified where possible according to the relative success of their indexing and also according to the arrangement of known repetitive sequences in the fragments. The details for a clone corresponding to each general classification are listed in each case.
cleavage site. Repeat sequence DNA can occur at any combination of either end or the middle of an indexed fragment. When fragments consisted entirely of repetitive sequences, it was not always possible to determine whether indexing had been successful because there were multiple possible origins for the fragment (column H), for example clone M2-01-B08 (Table 4). It was rare for the repeat sequence to have been indexed incorrectly when it was possible to identify the actual location of an indexed fragment containing repeat sequences in the genome data, for example clone M2-01-C03 (Table 4).

Tables 2 and 3 show examples of correctly indexed fragments of human genomic DNA. The original BamHI/BglII or BpmII sites are shown in bold, as are correctly selected 2-base overhangs.
Integrity after the ligation errors. Complex patterns of cleavage were discernible, for example M2-01-D09 (Table 4), where the corresponding genomic sequence suggested that cleavage from the indexer had occurred twice. Apart from the entirely repetitive sequences, five fragments occurred twice and one fragment three times in the indexed subsets (Table 5, columns K and L).

### Table 4. Frequency of types of fragments with respect to indexing success and presence of repeat sequences as isolated from male genomic DNA by the indexer combination first TT, second AG, third GT

<table>
<thead>
<tr>
<th>Indexing history</th>
<th>Frequency of occurrence</th>
<th>Type of fragment indexed from human genomic set M2</th>
<th>Example clone ID</th>
<th>Position of repeat</th>
<th>Sequence start</th>
<th>Sequence end</th>
<th>Length/repeat</th>
<th>EMBL start</th>
<th>EMBL end</th>
<th>EMBL ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% 10</td>
<td>M2-01-A03</td>
<td>603</td>
<td>679</td>
<td>76</td>
<td>15627</td>
<td>15703</td>
<td>EMU:AC076972</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% 6</td>
<td>M2-01-H09</td>
<td>19644</td>
<td>20266</td>
<td>622</td>
<td>151320</td>
<td>151944</td>
<td>EM:AC005510</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% 3</td>
<td>M2-01-G03</td>
<td>16034</td>
<td>16370</td>
<td>336</td>
<td>97122</td>
<td>96787</td>
<td>EM:AC007736</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% 7</td>
<td>M2-01-E06</td>
<td>10137</td>
<td>10388</td>
<td>251</td>
<td>48579</td>
<td>48329</td>
<td>AL357555</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% 3</td>
<td>M2-01-C03</td>
<td>4872</td>
<td>5161</td>
<td>289</td>
<td>134858</td>
<td>135150</td>
<td>EM:AC021187</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% 5</td>
<td>M2-01-E10</td>
<td>10767</td>
<td>10923</td>
<td>156</td>
<td>78623</td>
<td>78778</td>
<td>HS436C18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% 3</td>
<td>M2-01-B04</td>
<td>2128</td>
<td>2418</td>
<td>290</td>
<td>56637</td>
<td>56348</td>
<td>EM:AC025375</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1s+1),2,3</td>
<td>M2-01-A03</td>
<td>603</td>
<td>679</td>
<td>76</td>
<td>15627</td>
<td>15703</td>
<td>EMU:AC076972</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2,(3s+1)</td>
<td>M2-01-B08</td>
<td>2981</td>
<td>3051</td>
<td>LTR/MaLR</td>
<td>Rep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(N is missing), restriction site not found; (n), intervening base at indexed site; L, M, R, repeat sequence at left end, middle and right end of sequence; E, M or –, repeat sequence at one end of sequence; rep, repeat sequence with unambiguous EMBL match; Rep, repeat sequence with no unambiguous EMBL match; 1,2,3, 1st, 2nd and 3rd indexed sites, respectively; NCP, no conclusions possible; No match, no EMBL match.
Different individuals contributed the multiple instances in half of the cases. The fragments occurring multiple times had always been indexed correctly, were all found to have an unambiguous location in the genome and did not contain any known repetitive sequence. This suggests an indexed population comprising discrete, correctly indexed fragments and a complex background of incorrectly indexed fragments, the latter contributed by the different types of errors reported above.

Standard methods of detecting indexed fragments for comparative purposes include gel electrophoresis and hybridisation. The least efficient indexing that we observed was 35.7% (Table 5, column B). This corresponds to an enrichment of at least 686-fold. Provided that incorrectly indexed fragments contribute a background with a uniform distribution of sizes and no obvious biases for selected parts of the genome, we would not expect detection of the correctly indexed fragments by either approach above to be compromised. The cloned fragments whose indexing fate was known were used to determine the sizes of the genomic fragments that had most likely been indexed. Fragments from the first four subsets of Table 5 were used. Their size distribution together with that of fragments considered to have been indexed incorrectly is shown (Fig. 4). Incorrectly indexed fragments are clearly in the minority and have a size distribution similar to that of the correctly indexed fragments. They would therefore not be expected to affect comparative studies. Fragments of below 100–150 bp were rare, consistent with the size selection used. The mean size was 435 bp. This presumably represents a bias against the larger fragments during the PCR and cloning. The overall size distribution is ideal for comparative studies using conventional or automated fluorescent gel electrophoresis.

Other reports of molecular indexing cannot be directly compared to our approach because they have targeted cDNA (8–12). The range of abundances at which different cDNAs occur makes it impractical to calculate the fidelity of their indexing. cDNA is also a less stringent test of indexing than genomic DNA because the lower complexity of the former reduces the opportunity for PCR and ligation errors. Reports for which it is possible to estimate the fidelity of indexing suggest lower levels than we obtained. The same fragment of cDNA for skeletal muscle was found in two independent indexed subsets, suggesting indexing failure in at least one case (12). Ligation fidelity of 100% has been observed for cDNA indexing, but this was using a low ratio of indexer to cDNA and there was an associated increase in internal priming during PCR (21). The overall fidelity of indexing in this case was 21.3%. The highest fidelity of indexing in the same report was 32.5%. Four bases had been selected in both cases, so a lower overall fidelity would have been expected if, as in our report, six bases had been targeted. A further complication in interpreting other reports is that their fidelities concern fragments in multiple instances.

Table 5. Summary of indexing results with respect to fidelity of the indexing steps and unambiguous fragments indexed more than once

| Source of DNA | 1st | 2nd | 3rd | Enrichment | Indexing | Ligation | 17/15 | No | Partial | Repeat | Indexing | One or more | Fragments |
|--------------|-----|-----|-----|------------|----------|---------|------|---|--------|--------|----------|not indexed| in total|
| Female ga ct aa | 1371 | 71.4 | 99.0 | 25 | 3 | 5 | 5 | 16 | 2 | 10 | 0 | 2 | |
| Male | 1411 | 73.5 | 97.3 | 36 | 8 | 2 | 3 | 13 | 0 | 13 | 1 | |
| Female tt ag gt | 1280 | 66.7 | 91.4 | 20 | 6 | 7 | 9 | 0 | 2 | 10 | 0 | 4 | |
| Male | 1465 | 76.3 | 95.9 | 29 | 10 | 11 | 2 | 6 | 3 | 8 | 1 | |
| Female ac cc tc | 686 | 35.7 | 89.5 | 5 | 2 | 2 | 8 | 0 | 1 | 9 | 1 | 4 | |
| Male | 754 | 39.3 | 81.3 | 11 | 2 | 12 | 6 | 5 | 2 | 17 | 0 | |
| Female ca cg tg | 960 | 50.0 | 68.4 | 8 | 0 | 2 | 3 | 0 | 1 | 8 | 1 | 3 | |
| Male | 1280 | 66.7 | 75.0 | 8 | 1 | 2 | 7 | 4 | 1 | 4 | 0 | |
| Female gg at aa | 1200 | 62.5 | 70.8 | 15 | 0 | 0 | 11 | 8 | 0 | 9 | 0 | 0 | |
| Male | 891 | 46.4 | 77.8 | 13 | 2 | 18 | 6 | 11 | 0 | 15 | 0 | |

(A) Enrichment assuming 1/1920 fragments selected.
(B) Indexing efficiency assuming 1/1920 fragments selected.
(C) Percentage of indexed ends found to have the expected indexed bases at the indexed point in the genomic sequence.
(D) Number of fragments found to have been indexed correctly.
(E) Number of fragments found to have been indexed correctly and BpmI had cut at one base further than expected.
(F) Number of fragments with no significant match to sequence databases.
(G) Number of fragments found to only partially match sequence databases.
(H) Number of fragments found to consist entirely of repeat sequence DNA.
(I) Number of fragments for which the sequence database was not complete.
(J) Number of fragments with one or more ends not indexed.
(K) Number of correctly indexed fragments with unambiguous genomic location that were isolated twice.
(L) As (K) except number of fragments occurring more than once in combined male and female indexed subsets of the same type.

NB. Fragments can count towards the totals in more than one row of a column.
ments that have been gel purified following indexing. This ignores the contribution of the background which we have found to contain a significant proportion of incorrectly indexed fragments. The overall indexing fidelities of between 35.7 and 76.3% that we report for selection of six bases from genomic fragments sets a standard for other approaches. Now that we have identified the factors affecting overall fidelity, we expect further improvements to be made, for example by changes to the PCR.

There are a minimum of 60 and 47 unique, correctly indexed fragments for the combined male and female data for the first two subsets of Table 5, respectively. These were the subsets for which the most fragments had been isolated. The small numbers of fragments occurring more than once (Table 5, columns K and L) suggests a significantly greater total population. Performing additional rounds of indexing to decrease the number of fragments obtained can facilitate comparison by, for example, electrophoresis. This can be simply achieved by including a \( BpmI \) site in one of the final indexers to expose more internal sequence. The exposed sequence can be labelled using an adapter that is specific for one of the 16 possible two-base overhangs, with a consequent increase in resolution. Alternatively, PCR can be postponed until all rounds of indexing are complete. A maximum of five independent indexing steps, each selecting a two-base sequence, would be required to yield approximately a single indexed fragment per final indexed subset (10^6/16^5). It is expected that increased indexing would be unnecessary for detection of indexed fragments by hybridisation to corresponding arrays of indexed fragments since the overall complexity of the population obtained after three independent rounds of indexing (~100 000 bp) is within the detection limits of such systems.

Our approach yielded human sequences that previously had not been reported and we have also found that \( BpmI \) may cleave up to a few bases beyond its expected site. The approach is now being adopted for screening breast cancer samples for allelic imbalance with regard to loci of tumour suppressors and oncogenes. Use of Type II restriction enzymes with sensitivity to methylation will also make it possible for us to screen for changes in methylation status.

**CONCLUSION**

We have produced a series of plasmids pINDnn and shown that they can be used to index fragments of human genomic DNA and purify the indexed fragments. Iterative indexing from the

![Figure 4. Frequency distribution of sizes of original indexed fragments from the subsets F1, F2, M1 and M2. Independent clones of fragments indexed originally from human genomic DNA were sequenced and the sequences compared to the EMBL database to find the corresponding genomic human sequences. The sizes of the originally indexed fragments were determined from the positions of presumed indexed sites found in the EMBL sequences corresponding to the clones.](image-url)
sites for Type II restriction endonucleases has been demonstrated. Enrichment $>495$-fold and up to 1411-fold was observed. Increased rounds of indexing would be expected to achieve greater resolution. It is likely to be possible to design adapters containing the site for a Type IIS restriction endonuclease for use with the ends produced by most restriction endonucleases to access particular regions of defined fragments or particular features such as methylation. In order to study most regions of the genome it will no longer be necessary to clone or isolate with PCR primers corresponding to the region of interest. The regions of interest may be accessible in principle by a suitable combination of restriction enzymes and indexing adapters.

The approach is already suitable for screening human genomic DNA (3.2 $\times$ 10$^9$) (23) and could be used for global comparisons by differential display-type screens or coupled with microarrays. The detection sensitivity of the latter would be enhanced (5). In contrast to some other types of screening, indexing retains the selected targets as amplicons for further investigation.

ACKNOWLEDGEMENT

We are especially grateful for the support and facilities provided by the Clatterbridge Cancer Research Trust.

REFERENCES