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 IIS 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 IIS 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 IIS 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 IIS 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 IIS 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 endoncleases.

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 *Bam*HI and *Bgl*II. 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 *Bam*HI and *Bgl*II suggests that the approach will be applicable to virtually any fragments

*To whom correspondence should be addressed. Tel: +44 151 343 4303; Fax: +44 151 343 1820; Email: rosss@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 $MgCl_2$.

Indexing vectors

The plasmids pINDaa–pINDtt were produced for indexing as described below. Indexing plasmids were cut to completion with *Bam*HI and *Bsr*DI.

Bacterial transformation

Escherichia coli XL-1 Blue (Stratagene) was transformed by the $CaCl_2$ method and recombinants selected using 50 µg/ml ampicillin with 80 µg/ml IPTG and 50 µg/ml X-gal.

Molecular indexing

A complete *Bam*HI and *Bgl*II 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′-gatcgctccagggatcgagcttgagacctgcgactgg in a 120 µl reaction containing 4 U T4 DNA ligase (NEB). Excess adapters were removed by spin chromatography using Sepharose CL-4B (Amersham Pharmacia Biotech). The adapter and end sequences were cleaved from the purified DNA by *Bpm*I digestion and the purification by spin chromatography repeated. One-tenth of the purified material was added to indexing reactions containing equimolar amounts of the indexers and blockers at a 30-fold excess over the fragments to be indexed and ligated. Each second indexer conformed to the general design 5′-ccagtcgcaggtctcaagcaaggatcc**nn**, where **nn** are the indexing bases. The complementary strand for the second indexers was 5′-ggatccttgcttgagacctgcgactgg. Blocking adapters were 5′-tggaaaaagggggag**nn′**, where **nn′** is all two-base combinations except those of the indexers used. The complementary strand of the blockers was 5′-ctcccccttttcc. Dimerised indexing plasmid plus indexed fragments were separated from the remaining material by agarose gel electrophoresis and purified. *Bpm*I digestion was repeated and the enzyme inactivated at 70°C for 20 min. The resultant material was used for a final round of indexing containing equimolar amounts of the third indexer and blocking adapters in a total excess of more than 100-fold over the material used. The third indexers had the design 5′-ccagtcgcaggtctcaagcgacctgcctggag**nn**, where **nn** are the indexing bases. Their complementary strand was 5′-ctccaggcaggtcgcttgagacctgcgactg. Indexed material was serially diluted and dilutions amplified by PCR using 40 pmol/50 µl primer 5′-ccagtcgcaggtctcaagc, 2.5 mM $MgCl_2$, 200 µM dNTPs and 0.2 U *Taq* DNA polymerase (HotStar; Qiagen) according to the manufacturer's instructions, for 32 cycles of 95°C for 30 s, 60°C for 30 s and 72°C for 90 s.

Cloning indexed material

The plasmid vector pUC19 was modified by digestion with *Eco*RI and *Hin*dIII and replacement of the polylinker with the adapter 5′-agcttgaagactgccaggcatgggatcctg and its complementary strand 5′-aattcaggatcccatgcctggcagtcttca, transformed and purified as described for the pINDnn vectors. The modified vector was cut with *Bam*HI and *Bbs*I. Fragments that had been PCR amplified and indexed were purified and incubated at 1 µg per 10 µl with 3 U T4 DNA polymerase 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 *Bam*HI, 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 (17). Remaining sequences were sent to the EBI for BLAST comparisons to known sequences (18). Sites for the restriction endonucleases *Bam*HI, *Bgl*II and *Bpm*I were mapped by the GCG to regions of the genome, plus flanking sequences, that BLAST matched to the sequences of the indexed fragments (19). The mapped regions were examined to determine how indexing had occurred. The GCG was accessed via the HGMP computing services (20).

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 *Bsr*DI and *Bpm*I. 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 *Bsr*DI and one for *Bpm*I 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 *Xmn*I and *Afl*III sites were used to replace the corresponding region in the

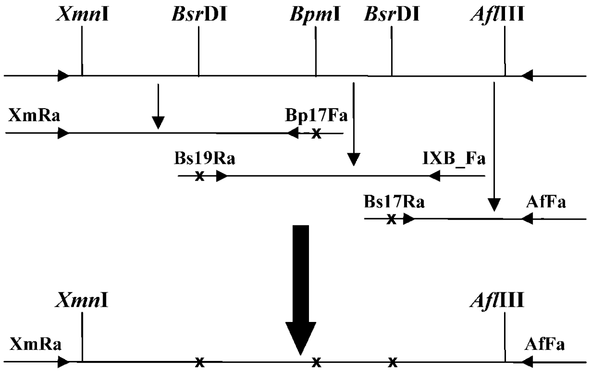

Figure 1. Removal of the *Bpm*I and *Bsr*DI sites of pUC19 by recombination PCR. The indexing plasmids were created by replacing the polylinker region of pUC19 with one containing adjacent sites for *Bpm*I and *Bsr*DI. Sites for *Bpm*I and *Bsr*DI within the ampicillin gene of pUC19 were first removed. Three overlapping regions between the *Xmn*I and *Afl*III sites of pUC19 were amplified using the primer pairs XmRa (5′-TCTCAACAGCGGTAAGATCC) with Bp17Fa (5′-ACGCTCACCGGCACCAGATT), Bs19Ra (5′-CCTGTAGCTATGGCAACAAC) with IXB_Fa (5′-AGTATTTGGTATCTGCGCTC) and Bs17Ra (5′-CTCGCGGTATAATTGCAGCA) with AfFa (5′-GGTAATACGGTTATCCACAG), which altered the sequence within the *Bpm*I and *Bsr*DI recognition sites by introducing neutral substitutions (**x** in figure and underline in sequence). Recombination PCR using primers XmRa and AfFa joined the amplified regions. The new region was then used to replace the original region between *Xmn*I and *Afl*III of pUC19 producing plasmid pIND10. (N.B. not to scale.)

original plasmid. Separate double digestions of candidate modified plasmids with *Bsr*DI or *Bpm*I each with *Xmn*I 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 *Bsr*DI and *Bpm*I 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 *Eco*RI and *Hin*dIII and replaced by an adapter containing the recognition sites for the restriction enzymes *Bsr*DI and *Bpm*I of general design 5′-aattctggag**nn**cattgccgacaaggatcc and the complementary sequence 5′-agctggatccttgtcggcaatg**n′n′**ctccag, where **nn** is one of the dinucleotide pairs aa to tt and **n′n′** is the corresponding, complementary pair. Sixteen plasmid indexers were produced so that digestion of the resultant constructs with *Bsr*DI could give rise to all combinations of two-base overhangs required for indexing. Equimolar amounts of each complementary pair were mixed and heated to 95°C before cooling to ambient temperature. Paired oligonucleotides were ligated at 3 molar excess to 200 ng of the cut vector in a 20 µl reaction. Ligated material was cut with *Xba*I to select against the original plasmid and the reactions used for transformation. Plasmids were produced and cut with *Hin*dIII, *Bsr*DI and *Eco*RI to score for the new polylinker. Candidate plasmids were diluted 1 in 1000 in water and amplified by PCR using the primers 5′-AGGCACCCCAGGCTTTAC and 5′-CCGCACAGATGCGTAAGG. PCR products were purified and their sequences confirmed by using the PCR primers as sequencing primers. The plasmids have been named pINDnn. The vectors in all of our series are conventionally represented with the *Eco*RI site of the original pUC19 plasmid on the left and the *Hin*dIII site on the right. **nn** corresponded to the dinucleotide immediately adjacent to their *Bpm*I site in the direction of the *Bsr*DI site so that the two-base 3′-overhang produced by *Bsr*DI corresponded exactly to the two particular bases found at **nn**. For example, pINDag has the dinucleotide ag in its upper single-stranded 3′-end produced on digestion by *Bsr*DI. In general plasmids that produced pale blue colonies after 24 h at 37°C had the required polylinker sequences.

Molecular indexing of human genomic DNA

Molecular indexing of human genomic DNA that had been digested with the Type II restriction endonucleases *Bam*HI and *Bgl*II was performed as shown in Figure 2. The pINDnn plasmids produced above were used as one of the indexers. Prior cleavage of the plasmids at a unique *Bam*HI site allowed them to dimerise (Fig. 2A), thus facilitating their purification from non-indexed fragments by size separation (Fig. 2B).

Approximately 1 000 000 fragments were produced and sorted in two rounds of indexing. Adapters containing the site for the Type IIS restriction endonuclease *Bpm*I were ligated to the *Bam*HI/*Bgl*II fragments. *Bpm*I was then used to expose sequences that were internal to the 5′-gatc(a/c) ends. The first round of indexing used two indexers, one of which was a pINDnn plasmid (indexer 1, Fig. 2B). Size separation of the dimerised plasmid was achieved by agarose gel electrophoresis. The purified material was re-digested with *Bpm*I to expose further bases for indexing within fragments that had been co-selected by the plasmid indexer. 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 *Bpm*I 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 *Bam*HI 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 *Bam*HI and ligating the adapter 5′-gatcgctccagctgtcgagctt, having a 3′-FAM label, and its complementary strand 5′-aagctcgacagctggagc, 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.ccrt.org.uk). The sequences were compared with the available human sequence to determine whether they had been successfully indexed. The existing human sequence was

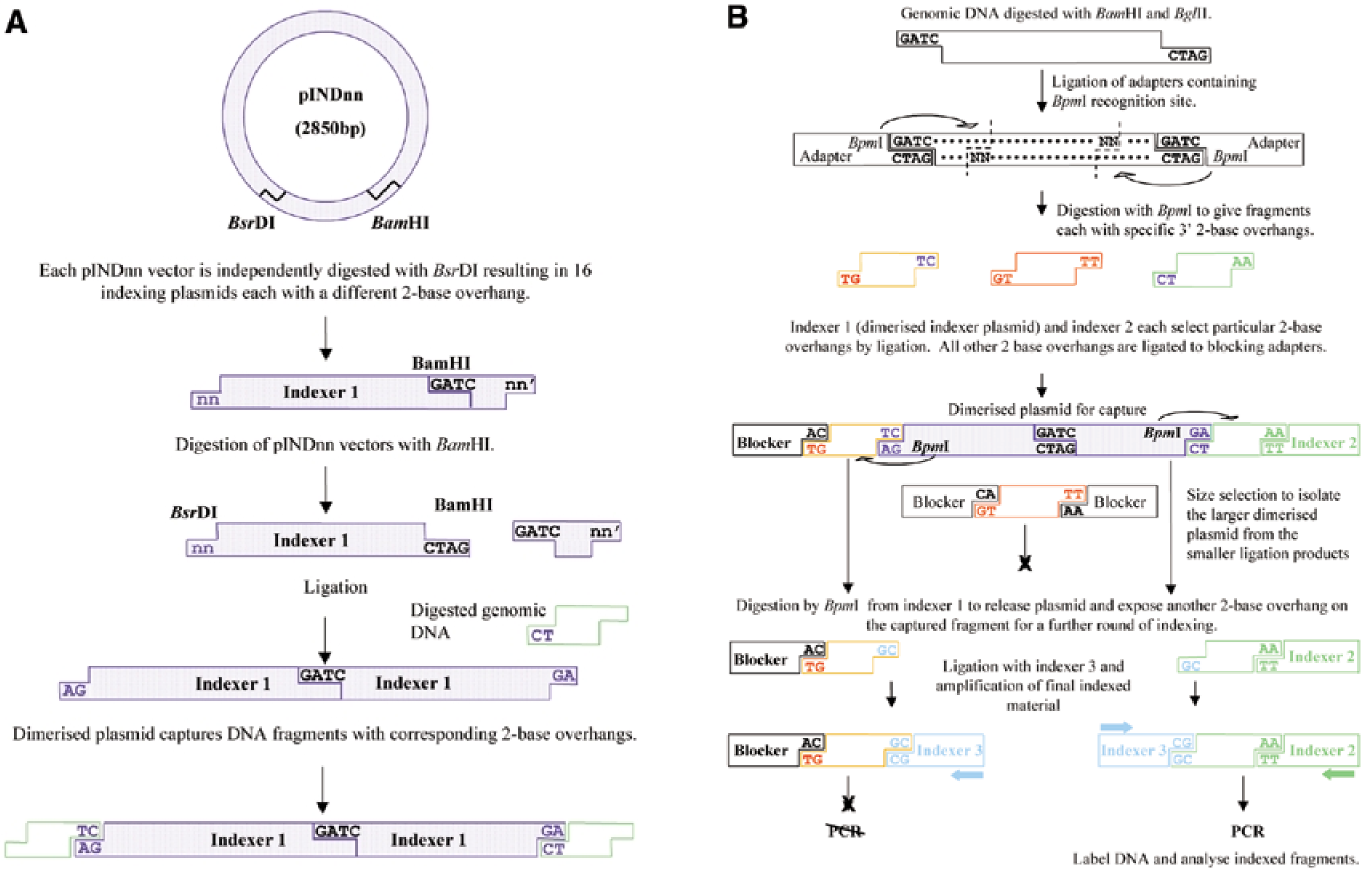
A
pNDnn
(2850bp)
BsrDI
BamHI
Each pINDnn vector is independently digested with BsrDI resulting in 16 indexing plasmids each with a different 2-base overhang.
BamHI
nn
Indexer 1
GATC
nn'
Digestion of pINDnn vectors with BamHI.
BsrDI
BamHI
nn
Indexer 1
CTAG
GATC
nn'
Ligation
Digested genomic DNA
CT
AG
Indexer 1
GATC
Indexer 1
GA
Dimerised plasmid captures DNA fragments with corresponding 2-base overhangs.
TC
AG
Indexer 1
GATC
Indexer 1
GA
CT
B
Genomic DNA digested with BamHI and BglII.
GATC
CTAG
Ligation of adapters containing BpmI recognition site.
Adapter
BpmI
GATC
CTAG
NN
NN
GATC
CTAG
BpmI
Adapter
Digestion with BpmI to give fragments each with specific 3' 2-base overhangs.
TG
TC
GT
TT
CT
AA
Indexer 1 (dimerised indexer plasmid) and indexer 2 each select particular 2-base overhangs by ligation. All other 2 base overhangs are ligated to blocking adapters.
Dimerised plasmid for capture
Blocker
AC
TG
TC
AG
BpmI
GATC
CTAG
BpmI
GA
CT
AA
TT
Indexer 2
Blocker
CA
GT
TT
AA
Blocker
Size selection to isolate the larger dimerised plasmid from the smaller ligation products
Digestion by BpmI from indexer 1 to release plasmid and expose another 2-base overhang on the captured fragment for a further round of indexing.
Blocker
AC
TG
GC
Ligation with indexer 3 and amplification of final indexed material
GC
AA
TT
Indexer 2
Blocker
AC
TG
GC
CG
Indexer 3
Indexer 3
CG
GC
AA
TT
Indexer 2
PCR
PCR
Label DNA and analyse indexed fragments.

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

Table 1. Indexer combinations used to produce the subsets of genomic DNA fragments

Subset	Indexer 1	Indexer 2	Indexer 3
FG1	GG	CA	TC
FG2	CA	GG	TG
FG3	CC	CT	CG
FG4	GA	CT	AA
FG5	CT	GA	AC
FG6	CA	CT	CA

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 *Bam*HI, *Bgl*II or *Bpm*I 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 *Bam*HI or *Bgl*II site or 14 bases from a *Bpm*I 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

```
EM:AP002841

       CTTCCCATGTGTGCTTCTTCATACTATCTCCTCTTTCTGGTTGGCTGACATGGGAGTGAT
149600 +---------+---------+---------+---------+---------+---------
       GAAGGGTACACACGAAGAAGTATGATAGAGGAGAAAGACCAACCGACTGTACCCTCACTA
     BglII
         |    |<   8  >||<    12    >|
       CCAGATCTCTCTATAATTTGCCACTGAAGTGTTTTTGTTTCTCAGTAGCACCTGCCTGTG
149660 +---------+---------+---------+---------+---------+---------
       GGTCTAGAGAGATATTAAACGGTGACTTCACAAAAACAAAGAGTCATCGTGGACGGACAC

       ATGGCTGTGCATTGATAAGTCCTCACGTAGCCTGCATGTGGGAGCTGTCTGTTAAAGATC
149720 +---------+---------+---------+---------+---------+---------
       TACCGACACGTAACTATTCAGGAGTGCATCGGACGTACACCCTCGACAGACAATTTCTAG

       ATCAGTGTGGGACTCTAACTGCGTGGATGAAGCAAAGCCCACTGCTTTCCCTACTACCCT
149780 +---------+---------+---------+---------+---------+---------
       TAGTCACACCCTGAGATTGACGCACCTACTTCGTTTCGGGTGACGAAAGGGATGATGGGA

       AACCTGGCACTTTCCTGTACTGTTATATGAGCAAGAAACAGACTTGTACTGTGTTGAGTC
149840 +---------+---------+---------+---------+---------+---------
       TTGGACCGTGAAAGGACATGACAATATACTCGTTCTTTGTCTGAACATGACACAACTCAG
                                             BpmI
                                                |
       ACTGCATGTTTGGGACTTTTTGTTACTGCAGCCTCTCACACTCTAACTGATACGCTCTCC
149900 +---------+---------+---------+---------+---------+---------
       TGACGTACAAACCCTGAAAAACAATGACGTCGGAGAGTGTGAGATTGACTATGCGAGAGG
                                                |<     14     >|
       AGCTGTGGTCTGTCAGGCTAAAAGCCTGCACTGGGAAAAGGAGGCACAATGAGCTCTCTA
149960 +---------+---------+---------+---------+---------+---------
       TCGACACCAGACAGTCCGATTTTCGGACGTGACCCTTTTCCTCCGTGTTACTCGAGAGAT
```

Figure 3. Representative example of correctly indexed genomic DNA and flanking sequences. The figure shows all possible configurations of indexing. A *Bpm*I indexing adapter ligated to the *Bgl*II site exposes two bases for indexing which are eight bases from the end of the *Bgl*II site. The first indexer ligated at these bases contains a *Bpm*I site which allows two more bases to be exposed for indexing 12 bases along. The *Bpm*I site in the genomic DNA was cleaved at the first *Bpm*I digest and exposes two bases for indexing 14 bases from the site. Restriction sites and indexed nucleotides are shown in bold.

Table 2. Genomic sequences corresponding to the ends of correctly indexed fragments from subset FG1

Accession no.	Start of sequence match	Sequence showing correct indexing at 5′ region of match	Sequence showing correct indexing at 3′ region of match	End of sequence match
AC004882	161126	GTG**CTGGAG**GAAGGGGCCCTGGG**CA**GGG CACGACCTCCTTCCCCGGGACCCGTCCC	GTG**CTGGAG**GATATTGCTGAAGACCTGT CACGACCTC**CT**ATAACGACTTCT**GG**ACA	161377
AC008464	104583	TTT**CA**CTTATCATAATGTCCTCCAGGTT AAAGTGAATAGTATTACAG**GAGGTC**CAA	GAGGCCAGGCTTTCCCAGTTCATTGTTGTTCTCCAG **CT**CCGGTCCGAAAG**GG**TCAAGTAACAACAA**GAGGTC**	104812
AC018371	84110	**CTGGAG**TTTGTGGCTGATGA**GG**GCATCTGCTACC**TC**C GACCTCAAACACCGACTACTCCCGTAGACGATGGAGG	ATGTGACTGCCATCAACTGTAATGAAAAGATCTA TACACTGACGGTAGTTG**AC**ATTACTTT**TCTAGA**T	84351
AC024719	92331	TCT**CTGGAG**AACCCTGACTAATT**CA**GCAC AGAGACCTCAAGGGACTGATTAAGTCGTG	GAGTTTCTTGGTGTCCCCTTCCATTTTTCACTCCAG **CT**CAAAGAACCACA**GG**GGAAGGTAAAAAGT**GAGGTC**	92552
AC067889	38557	AGAA**GG**CGAGGGACCGGC**TC**CTCCAG TCTTCCGCTCCCTGGCCGAG**GAGGTC**	CTTGAGAGGAGAGCTGTTCTCCAGAG GA**AC**TCTCCTCTCGACAA**GAGGTC**TC	38720
AL035690	76650	TCAG**AGATCT**TTCAGGTC**CA**GAT AGTCTCTAGAAAGTCCAGGTCTA	**CTGGAG**GACGTAGCAGTAAACCTCGGCAG GACCTC**CT**GCATCGTCATTT**GG**AGCCGTC	76865
AL157397	90001	**CTGGAG**GTGGTATTTGCTAT**CA**CTTTA GACCTCCACCATAAACGATAGTGAAAT	GAGGAGAGGCCAGGCCACTGAGTGGCCCTACTCCAG **CT**CCTCTCCGGTCC**GG**TGACTCACCGGGAT**GAGGTC**	90524
AP002781	18977	TAAC**CA**GTGGAGTGGATGAACTCCAG ATTGGTCACCTCACCTACTT**GAGGTC**	**CTGGAG**GAAGCCCCAATAGCCCCTCAC GACCTC**CT**TCGGGGTTATCG**GG**GAGTG	19357

Tables 2 and 3 show examples of correctly indexed fragments of human genomic DNA. The original *Bam*HI/B*gl*II or *Bpm*I sites are shown in bold. Correctly seclected 2-base overhangs are also shown in bold.

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 a fragment. A further 20 had no human match or had an ambiguous match to human sequences because of the presence of repetitive sequences. One in 1920 (2/16 × 1/15 × 1/16) fragments should on average be indexed to a particular subset. The observed enrichment was at least 495-fold even if the ambiguous fragments had been indexed incorrectly. Fragments whose identity was unambiguous and that had been indexed correctly suggest a higher success rate. The subset FG2 similarly produced 15 correctly indexed fragments out of 47. Eight 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 *Bpm*I 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.

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 *Bpm*I (column E). The majority of these had been indexed correctly in all other respects. We believe that cleavage by *Bpm*I 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 *Bam*HI 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 *Bpm*I

Table 3. Correctly indexed fragments from subset FG2

Accession no.	Start of sequence match	Sequence showing correct indexing at 5′ region of match	Sequence showing correct indexing at 3′ region of match	End of sequence match
AC004387	9739	TGT**GGATCC**TCTAATAT**GG**GCTT ACACCTAGGAGATTATACCCGAA	TGCAGCTTTGAACTCCTGGGCTCAAGGGATCCTCT AC**GT**CGAAACTTGAGG**AC**CCGAGTTC**CCTAGG**AGA	9991
AC004447	68134	AC**GGATCC**CAGGAGCC**CA**CACATGACATCC**TG**CAG TGCCTAGGGTCCTCGGGTGTGTACTGTAGGACGTC	CCTTCCAATGCAGGGGATCCCAGG GGAA**GG**TTACGTCC**CCTAGG**GTCC	68365
AC00611	1780125	TTG**CTGGAG**GTTAGAGACTAGTT**GG**AGA AACGACCTCCAATCTCTGATCAACCTCT	CACACTCTAATAACTGTACAAAAATAGCGACTCCAGCT **GT**GTGAGATTATTG**AC**ATGTTTTTATCGCT**GAGGTC**GA	78323
AC009112	74773	GAA**GGATCC**AAGAGAGA**GG**GCCA CTTCCTAGGTTCTCTCTCCCGGT	***CTGGAG***CCAGGAGTCAGGCCCTGCATTCAAATCCTGCCTCCG GACCTCG**GT**CCTCAGTCCGGG**AC**GTAAGTTTAGGACG**GAGGC**	74972
AC012590	100723	TA**GGATCC**CATTGATT**CA**TTATTCCAGAGA**TG**TAA ATCCTAGGGTAACTAAGTAATAAGGTCTCTACATT	CCCCCTGACTTTCCATTGCTCCAGCCC GG**GG**GACTGAAAGGTAAC**GAGGTC**GGG	100591
AC016699	137182	CCA**AGATCT**CGCCACTG**CA**CTCCAGCCTGGG**TG**ACA GGTTCTAGAGCGGTGACGTGAGGTCGGACCCACTGT	GGTTCCTGTGCTGGGCGTGCCTCCAGAGT CCAA**GG**ACACGTCCCGCACG**GAGGTC**TCA	137626
AC020705	156518	T**GGATCC**CCTCAGCT**CA**GATGCTTTTTTA**ATG**CCT ACCTAGGGGAGTCGAGTCTACGAAAAAATTACGGA	CTCCTCCAATTTTAAAGTCTCCAGGTT GA**GG**AGGTTAAAATTTCA**GAGGTC**CAA	156518
AC023282	231740	ATT**CTGGAG**CAGAGGTTTTGCTG**AGG**ACAT TAAGACCTCGTCTCCAAAACGACTCCTGTA	TGGG**CTGGAG**CAGAATGTTTTCCATGTCCCT ACCCGACCTC**GT**CTTACAAAAGGT**AC**AGGGA	232000
AC02551	54479	GT**GGATCC**TAATTTCC**CA**GATCCTGGCAGC**TG**CTTG CACCTAGGATTAAAGGGTCTAGGACCGTCGACGAAC	GATTCCCTCTCCTTACCTATCTCCAGGAG CTAA**GG**GAGAGGAATGGATA**GAGGTC**GTC	54771
AC067804	1233008	TTA**CA**CACAGCCACGGC**TG**CTCCAGCGC AATGTGTGTCGGTGCCGAC**GAGGTC**GCG	CAGCAAGCCTGCCCCACCCTAAGGATCCTCA GTCGTTCGGACG**GG**GTGGGATT**CCTAGG**AGT	123218
AC074050	57647	TAC**CTGGAG**GGGCCTGGGCCTGA**GG**AACC ATGGACCTCCCCGGACCCGGACTCCTTGG	GCAACCCCAGAACGGTGACTAAAAGGTGGGACTCCAGGG C**GT**TGGGGTCTTGCC**AC**TGATTTTCCACCCT**GAGGTC**CC	57881
AL138688	47078	CAG**CTGGAG**GGAAGAGACGCGCA**GG**GCA GTCGACCTCCCTTCTCTGCGCGTCCCGT	TCTCACAGGTCTAGGGGTGAGGCTGAAGGATCCAAC AGA**GT**GTCCAGATCCCC**AC**TCCGACTT**CCTAGG**TTC	47214
AL139251	3071	**CTGGAG**ACACCTTCAGAATG**CA**CTGAATTTACCC**TG**TC GACCTCTGTGGAAGTCTTACGTGACTTAAATGGGACAG	GCCCAAGAACTGGCTGGCTCCAGGAT C**GG**GTTCTTGACCGACC**GAGGTC**CTA	3461
AL161660	17826	GGA**AGATCT**GGGGCAGT**GG**GGTG CCTTCTAGACCCCGTCACCCCAC	AACAAAAAACCCAAAATGCCGTGGAACGGACACTCCAGTG TT**GT**TTTTTGGGTTTT**AC**GGCACCTTGCCTGT**GAGGTC**AC	178640
AL359384	103055	CT**GGATCC**CCATTCTG**CA**CCCCCTGAGTGA**TG**GG GACCTAGGGGTAAGACGTGGGGGACTCACTACCC	TATCCATGGTCACAGATCTT ATA**GG**TACCAGTG**TCTAGA**A	103414
AP001357	144168	**CTGGAG**ATGGGCTGGCACCA**CA**ACAAAGTTGCCC**TG**TGC GACCTCTACCCGACCGTGGTGTTGTTTCAACGGGACACG	TATGCCACCCTGGTTCCTACCTCCAGGC ATAC**GG**TGGGACCAAGGATG**GAGGTC**CG	14490
D86998	23584	CT**GGATCC**CCTCCTCC**CA**GGTTTCTTTTCC**TG**CCC GACCTAGGGGAGGAGGGTCCAAAGAAAAGGACGGG	GCTCCAGCCATGCTTTCAGCTCCAGGGC CGA**GG**TCGGTACGAAAGTC**GAGGTC**CCG	23870

Tables 2 and 3 show examples of correctly indexed fragments of human genomic DNA. The original *Bam*HI/*Bgl*II or *Bpm*I sites are shown in bold, as are correctly selected 2-base overhangs.

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). This suggests that such sequences do not markedly affect the process. Of the fragments, 15.1% had no match to human or any other database sequences and were therefore presumed to be unique sequences having originated in parts of the human genome that have yet to be sequenced (Table 5, column F). This figure increased to 26% including fragments that matched known repetitive sequences but which otherwise had no match. We allowed fragments having the original adapter to amplify in the final PCR so that failure of the first *Bam*HI and *Bgl*II cuts could be detected. They were observed only once in 320 fragments. Internal priming and restriction endonuclease failures account for the remaining loss of effi-

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

Indexing history	Frequency of occurrence	Type of fragment indexed from human genomic set M2							
		Example clone ID	Position of repeat	Sequence start	Sequence end	Length/repeat	EMBL start	EMBL end	EMBL ID
100%	10	M2-01-A03	~,~,~	603	679	76	15627	15703	EMU:AC076972
100%	6	M2-01-H09		19644	20266	622	151320	151944	EM:AC005510
			~,M,~	19681	19883	1204			
				20024	20061	Simple_rep			
100%	3	M2-01-G03		16034	16370	336	97122	96787	EM:AC007736
			~,~,E	16258	16379	LINE/CR1			
100%	7	M2-01-E06		10137	10388	251	48579	48329	EMU:AL357555
			~,M,E	10222	10387	SINE/MIR			
				4867	5162	LTR/ERV1			rep
100%	3	M2-01-C03	L,M,R	4872	5161	289	134858	135150	EM:AC021187
				4995	5158	163	43767	43930	
(1s+1),2,3	5	M2-01-E10	~,~,~	10767	10923	156	78623	78778	EMU:HS436C18
1,2,(3s+1)	5	M2-01-B04		2128	2418	290	56637	56348	EM:AC025375
			~,M,~	2294	2335	LINE/L2			
(1ned),2,3	1	M2-01-F03		12862	13012	150	39363	39214	EM:AC016680
			~,M,~	13002	13202	SINE/Alu			
1,2(ned),3	2	M2-01-A02		29	508	479	98237	98713	EMU:AC069303
			E,M,~	29	69	LTR/ERV1			
				70	456	LTR/MaLR			
				174	249	75	20629	20704	EMU:AL354813
				457	508	LTR/ERV1			
100% but	2	M2-01-D09		7827	8419	592	147210	146615	EM:AC020708
2 vector cuts			~,M,~	7836	8250	LTR/ERVL			
(1 rs missing),2,3	2	M2-01-G05		16783	17155	372	127940	127566	EMU:AL157392
			L,~,~	16783	16824	LINE/L1			
(1s+1),2,(3g(c)t)	1	M2-01-E05	~,~,~	9958	10094	136	103050	103185	EM:AC011306
1,(2×),3	1	M2-01-H12		20461	20485	24	122080	122056	EM:AC024490
			~,M,~	20486	20772	SINE/Alu			
				20774	20908	134	194	328	EM:BF103834
1,(2×,3×)	1	M2-01-G06		17199	17277	78	13081	13003	EM:AC009067
			~,M,~	17266	17559	SINE/Alu			
				17570	17618	48	12710	12662	EM:AC009067
NCP	1	M2-01-B03	~,M,E	2014	2095	SINE/Alu			Rep
		M2-01-B08	L,M,R	2981	3051	LTR/MaLR			Rep
NCP	5			3013	3036	23	114762	114739	EM:AL450303
				3049	3078	29	27436	27407	EMU:AC073493
				3052	3129	SINE/Alu			
NCP	11	M2-01-A05		689	1162	473			No match

(N rs 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; *N*ned, position *N* has no corresponding EMBL data; 100%, all restriction sites and indexing as expected; (*N*x), position of incorrect indexer; (*N*s+1), correct index site found 1 base beyond expected; NCP, no conclusions possible; No match, no EMBL match.

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

Source of DNA	Indexer			A	B	C	D	E	F	G	H	I	J	K	L
	1st	2nd	3rd	Enrichment	Indexing efficiency (% 1/1920)	Ligation fidelity (%)	Indexing as expected	17/15 *Bpm*I cut	No EMBL match	Partial EMBL match	Repeat sequence element	Indexing partially determined	One or more ends not indexed	Fragments occuring twice	Total fragments in multiple instances
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 *Bpm*I had cut at onr 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.

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 frag-

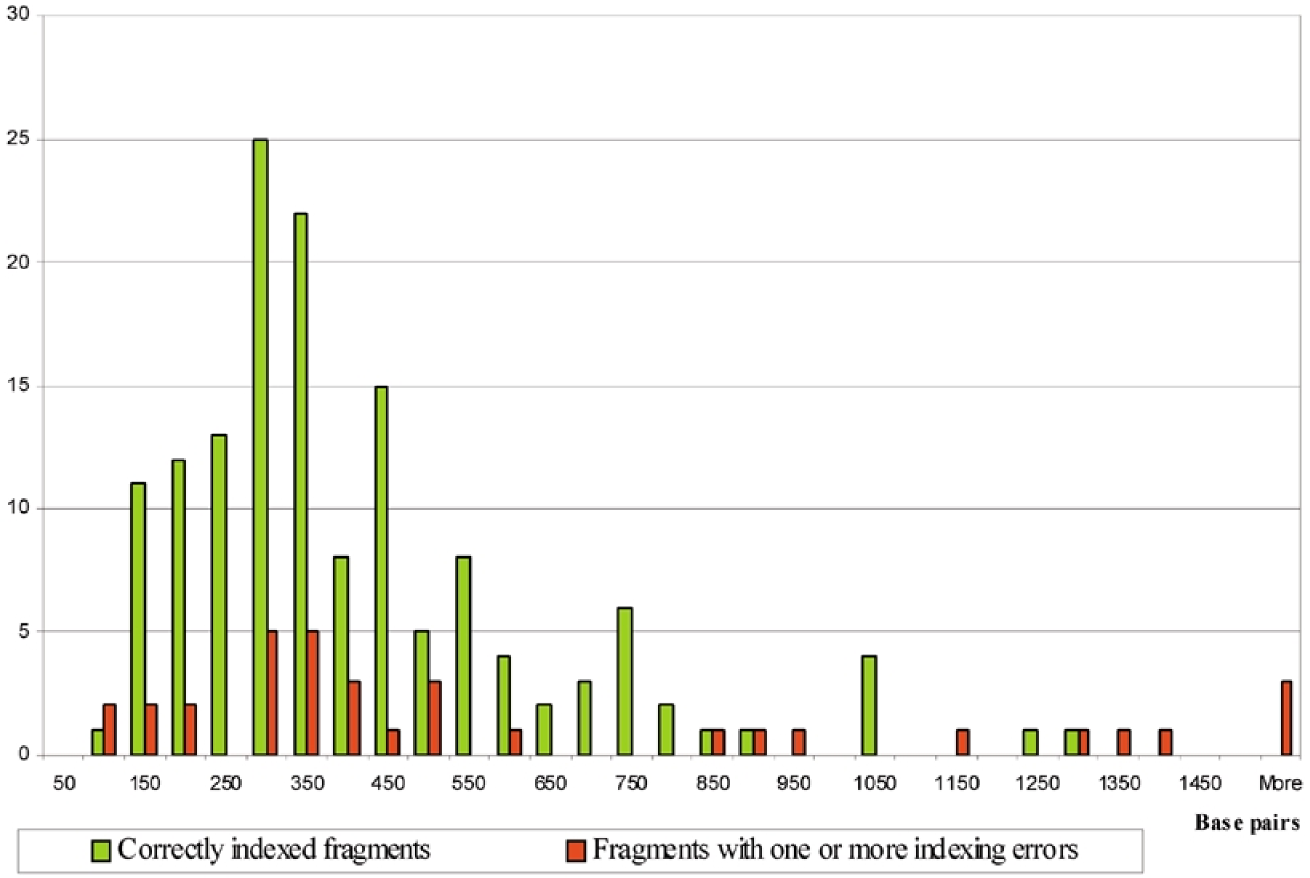

Figure 4. Frequency distribution of sizes of original indexed fragments from the subsets F1, F2, M1 and M2. Independent clones of fragments indexed originally from human genomic DNA were sequenced and the sequences compared to the EMBL database to find the corresponding genomic human sequences. The sizes of the originally indexed fragments were determined from the positions of presumed indexed sites found in the EMBL sequences corresponding to the clones

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 *Bpm*I 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 *Bpm*I 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 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 (3.2×10^9) (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

1. Liang,P. and Pardee,A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.
2. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
3. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M., Roth,R., George,D., Eletr,S., Albrecht,G., Vermaas,E., Williams,S.R., Moon,K., Burcham,T., Pallas,M., DuBridge,R.B., Kirchner,J., Fearon,K., Mao,J. and Corcoran,K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
4. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
5. Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y., Dairkee,S.H., Ljung,B.M., Gray,J.W. and Albertson,D.G. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.*, **2**, 207–211.
6. Lindblad-Toh,K., Winchester,E., Daly,M.J., Wang,D.G., Hirschhorn,J.N., Laviolette,J.P., Ardlie,K., Reich,D.E., Robinson,E., Sklar,P., Shah,N., Thomas,D., Fan,J.B., Gingeras,T., Warrington,J., Patil,N., Hudson,T.J. and Lander,E.S. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **4**, 381–386.
7. Unrau,P. and Deugau,K.V. (1994) Non-cloning amplification of specific DNA fragments from whole genomic DNA digests using DNA 'indexers'. *Gene*, **145**, 163–169.
8. Kato,K. (1995) Description of the entire mRNA population by a 3′ end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Res.*, **23**, 3685–3690.
9. Ivanova,N.B. and Belyavsky,A.V. (1995) Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Res.*, **23**, 2954–2958.
10. Kato,K. (1996) RNA fingerprinting by molecular indexing. *Nucleic Acids Res.*, **24**, 394–395.
11. Sibson,D.R. and Starkey,M.P. (1997) Increasing the average abundance of low-abundance cDNAs by ordered subdivision of cDNA populations. *Methods Mol. Biol.*, **69**, 13–32.
12. Mahadeva,H., Starkey,M.P., Sheikh,F.N., Mundy,C.R. and Samani,N.J. (1998) A simple and efficient method for the isolation of differentially expressed genes. *J. Mol. Biol.*, **284**, 1391–1398.
13. Wu,D.Y. and Wallace,R.B. (1989) Specificity of the nick-closing activity of bacteriophage T4 DNA ligase. *Gene*, **76**, 245–254.
14. Pritchard,C.E. and Southern,E.M. (1997) Effects of base mismatches on joining of short oligodeoxynucleotides by DNA ligases. *Nucleic Acids Res.*, **25**, 3403–3407.
15. Kim,S.C., Podhajska,A.J. and Szybalski,W. (1988) Cleaving DNA at any predetermined site with adapter-primers and class-IIS restriction enzymes. *Science*, **240**, 504–506.
16. Sibson,D.R. (1995) Adaptored sequencing. European patent application no. 95 906394.2.
17. Staden,R., Beal,K.F. and Bonfield,J.K. (1998) The Staden package. In Misener,S. and Krawetz,S.A. (eds), *Computer Methods in Molecular Biology*, Vol. 132, *Bioinformatics Methods and Protocols*. Humana Press, Totowa, NJ, pp. 115–130.
18. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. Genetics Computer Group (1991) *Program Manual for the GCG Package, Version 7*. GCG, Madison, WI.
20. Rysavy,F.R., Bishop,M.J., Gibbs,G.P. and Williams G.W. (1992) The UK Human Genome Mapping Project online computing service. *Comp. Appl. Biosci.*, **8**, 149–154.
21. Shaw-Smith,C.J., Coffey,A.J., Huckle,E., Durham,J., Campbell,E.A., Freeman,T.C., Walters,J.R. and Bentley,D.R. (2000) Improved method for detecting differentially expressed genes using cDNA indexing. *Biotechniques*, **28**, 958–964.
22. Cho,S.-H. and Kang,C. (1990) DNA sequence-dependent cleavage sites of restriction endonuclease *Hph*I. *Mol. Cell*, **1**, 81–86.
23. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.