2960–2966 *Nucleic Acids Research, 1997, Vol. 25, No. 15*

Representation of cloned genomic sequences in two sequencing vectors: correlation of DNA sequence and subclone distribution

Stephanie L. Chissoe*, Marco A. Marra, LaDeana Hillier, Ryan Brinkman, Richard K. Wilson and Robert H. Waterston

Department of Genetics and Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA

Received June 6, 1997; Accepted June 19, 1997

DDBJ/EMBL/GenBank accession nos U41279, U51244

ABSTRACT

Representation of subcloned *Caenorhabditis elegans* and human DNA sequences in both M13 and pUC sequencing vectors was determined in the context of large scale genomic sequencing. In many cases, regions of subclone under-representation correlated with the occurrence of repeat sequences, and in some cases the under-representation was orientation specific. Factors which affected subclone representation included the nature and complexity of the repeat sequence, as well as the length of the repeat region. In some but not all cases, notable differences between the M13 and pUC subclone distributions existed. However, in all regions lacking one type of subclone (either M13 or pUC), an alternate subclone was identified in at least one orientation. This suggests that complementary use of M13 and pUC subclones would provide the most comprehensive subclone coverage of a given genomic sequence.

INTRODUCTION

A collaborative effort to sequence the genome of the nematode *Caenorhabditis elegans* began in 1990. Since then, over 62 Mb of finished genomic sequence data has been completed by the Genome Sequencing Center and the Sanger Centre (*C.elegans* Mapping and Sequencing Consortium, unpublished). Various strategies were investigated early in the project (1), although the majority of the data has been produced using a strategy comprised of initial random subclone sequencing followed by directed sequencing of specific regions ('subclone-directed sequencing') (2,3).

In this strategy, random subclones are generated by ligation of sonicated genomic clone DNA fragments into sequencing vector, and single sequence reads are obtained from one end of the random clone. M13 has been the predominant subcloning vector, primarily because it allows the use of a robust, inexpensive template preparation method which yields high quality sequence data (4). The sequence reads corresponding to each genomic clone are then assembled via computer to generate sequence alignments, and consensus sequences are determined for each alignment. For most cosmid projects, the average redundancy of the consensus sequence (the number of times the consensus sequence is represented by sequence reads) following the random sequencing phase is ~6-fold. Assuming a truly random sampling, this should result in 99.8% representation of the clone sequence in sequenced contigs (5,6). Furthermore, assuming true randomness at this redundancy, >99.99% of the original clone sequence would be contained within random subclones of average size 1500 bp (as distinct from the average of 400 bp of the subclone represented in the sequence contigs). Here, any regions not adequately represented in the initial random read could easily be recovered by directed approaches.

Despite these predictions, after sequencing random M13 subclones to a redundancy of six, we frequently notice gaps in the sequence assembly which also are gaps in subclone representation. Non-random subclone representation has been observed before, and it is known that some sequences, most notably inverted and direct sequence repeats, are unstable or unclonable in certain *Escherichia coli* vector systems (7,8). Many of the uncloned regions (gaps) observed during our completion of numerous *C.elegans* cosmids occur within inverted repeat elements. These features occur quite frequently in *C.elegans*, with approximately one inverted repeat element every 5.5 kb (2). Well-conserved repeats with short spacer sequences anecdotally have been associated with these gaps. Only occasionally is an M13 subclone recovered containing even a portion of both repeat copies, and PCR amplification of these genomic regions typically is unreliable. As a result, final gap closure of these sequences requires a more time-consuming direct sequencing approach.

In order to investigate more systematically the representation of cloned genomic DNA sequences in randomly generated subclone libraries, subclone start-site distributions for two genomic clones in two sequencing vector systems have been determined. Human BAC clone 1D9, from a human chromosome 2 BAC library (9) and *C.elegans* cosmid C17C3 (10) were sonicated, and DNA fragments were subcloned into both M13 and pUC sequencing vectors. Subclones from each library were

*To whom correspondence should be addressed. Tel: +1 314 286 1460; Fax: +1 314 286 1810; Email: schissoe@watson.wustl.edu

prepared and sequenced. After sequence assembly, directed gap closure, final editing and sequence analysis, the distribution of vector-specific subclones was correlated to features identified within the genomic DNA sequence.

MATERIALS AND METHODS

Random subclone library preparation

Cosmid and BAC DNA were purified by alkaline lysis and cesium chloride banding, sonicated, and the resultant DNA fragments end-repaired, size selected and ligated to the appropriate M13 and pUC vectors as described elsewhere (3). Electrocompetent *E.coli* were transformed with each ligation and plated onto agar plates.

DNA sequence generation

Single-stranded M13 templates were purified using the ThermoMAX modified PEG/Triton protocol (4) and double-stranded pUC templates were purified using the Advanced Genetic Technologies Corporation 96-well boiling mini-prep procedure (Gaithersburg, MD). DNA templates were sequenced using fluorescent dye-primer cycle sequencing with Sequitherm DNA polymerase (11). Fluorescent sequencing reactions were electrophoresed on ABI 373A Sequencers equipped with the Stretch upgrade, and the sequence data were automatically collected and analyzed (3). DNA sequence data were automatically processed using the OTTO script, which performs quality evaluation, vector identification and removal, and initial assembly of the sequences using XBAP (12). Base-calling and sequence assembly were also performed using PHRED and PHRAP (P.Green, unpublished) in the case of C17C3. Sequence was manually edited using the XBAP interface (12). Sequence gaps were closed, the sequence was double-stranded and ambiguities were resolved (3). Sequence assemblies were verified by restriction digest fragment analysis.

DNA sequence analysis

Repeated sequences were identified within the finished sequences using TANDEM and INVERTED (R.Durbin, unpublished). Alu repeat elements in the human genomic sequence were identified using HMMFS (G.Miklem and S.Eddy, unpublished), and other human repeated sequences were identified via BLASTN (13) against a database of human repeats (15). These repeat sequences were masked prior to further analysis. Sequence repeats were displayed graphically using MIROPEATS (16) with a threshold of 30. Protein and nucleotide similarities were detected by sequence comparison to the public databases using BLASTX and BLASTN, respectively (13). Potential coding sequences were identified using GENEFINDER (P.Green, unpublished) for *C.elegans* DNA and several gene prediction programs for human DNA, including the FGENEH suite (FEXH and HEXON) (17,18), GRAIL (19), GENEFINDER, NETGENE (20) and XPOUND (21). The data generated during DNA sequence analysis was annotated using the ACEDB interface (J.Thierry-Mieg and R.Durbin, unpublished). Putative CpG islands within the human sequence were identified by cpgspans (G.Miklem, unpublished). The *C.elegans* cosmid C17C3 sequence has GenBank accession number U41279, and the human BAC 1D9 sequence has GenBank accession number U51244.

Analysis of subclone start-site distributions

Subclone start-sites were identified from the show relationships option within XBAP (12), and plotted as they occur along the sequence(s). The base pair intervals between successive subclone start-sites (start-site gaps) were calculated and the results sorted into a set of evenly distributed bins between zero bases and the maximum start-site gap size. The number of bins used for C17C3 and 1D9 were the square-root of the number of start sites. This calculation was used since it provided the appropriate number of bins (14). The observed frequency of start site intervals was compared using chi-squared analysis to the frequency expected of a Poisson distribution.

RESULTS AND DISCUSSION

Correlation of subclone start sites and under-represented DNA sequence in *C.elegans* cosmid C17C3

The start-site positions of 716 M13 subclones (373 in the 'plus' orientation, 343 in the opposite, or 'minus' orientation) and 380 pUC subclones (201 plus, 179 minus) were plotted as they occurred along the completed C17C3 insert sequence of 44 963 bp in Figure 1a. The directionality of the insert DNA within the subclones is indicated as plus or minus, assigned with respect to the arbitrarily oriented completed sequence. Overall, the slope of the pUC subclone plot was higher than for the M13 subclone plot, since fewer pUC subclones were sampled. In general, increases in the slopes of the subclone start-site plots in Figure 1a indicated areas of subclone under-representation. If the subclone start-sites were random, then their distribution would approximate a Poisson distribution (5,6). The observed distribution of subclone start-sites was compared by chi-squared analysis to that predicted by Poisson and the results are listed in Table 1. The C17C3 pUC subclone start-site distribution was more closely approximated by Poisson than that of M13. Interestingly, there was a notable difference between the distributions of the pUC minus and plus subclone start-sites.

Table 1. Chi-squared analysis of subclone start-site distribution

Clone	Subclone	Orientation	Degrees of freedom	$\Sigma\chi^2$ (obs)	$\Sigma\chi^2$ (exp)[a]
C17C3	M13	plus	18	427045.05	28.9
C17C3	M13	minus	17	4269.86	27.6
C17C3	pUC	plus	13	669.75	22.4
C17C3	pUC	minus	12	12.47	23.3
1D9	M13	plus	20	204.31	31.4
1D9	M13	minus	20	521.34	31.4
1D9	pUC	plus	13	1320.00	22.4
1D9	pUC	minus	13	540.77	22.4

[a]95% confidence value.

The average M13 subclone insert prepared by the method described above is 1450 bp and therefore, on average, a gap in subclone start-sites >1450 bp would result in a gap within the subclone DNAs representing the original clone sequence.

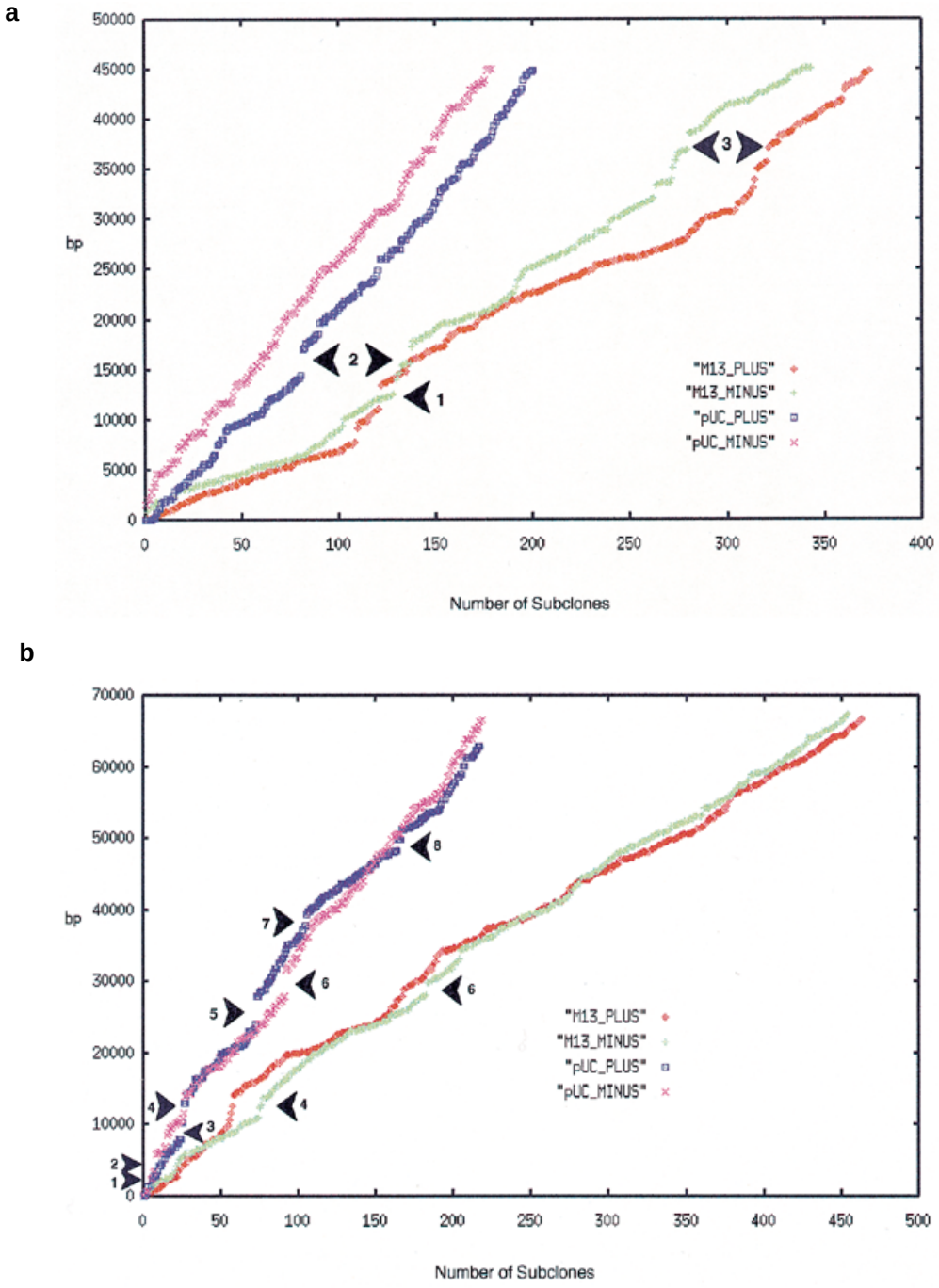

Figure 1. Subclone start-site positions. The start-site positions of sonication subclones were plotted as they occurred along (**a**) the completed *C.elegans* cosmid C17C3 sequence and (**b**) the completed human BAC 1D9 sequence. M13 and pUC subclones with inserted DNA cloned in both the plus and minus orientation are depicted. Experimentally significant regions which lack subclone start-sites are indicated by arrows.

(Start-site data for 518 forward and reverse pairs from 30 completed cosmid sequences were analyzed, and >90% of the M13 inserts were between 1200 and 1800 bp.) For that reason, distances between consecutive subclone start-sites of ≥1450 bp were identified as experimentally significant gaps in subclone representation. The gaps in subclone representation were confined to three regions for C17C3 and are listed in Table 2. To characterize the DNA sequence from these regions, the C17C3 sequence was analyzed and repeat sequences were classified by INVERTED and TANDEM are listed in Table 3a and 4a. Figure 2a displays repeated sequences as identified by MIROPEATS.

M13 plus and minus subclones were under-represented in region 1, although pUC subclone representation was not affected. This region contained an inverted repeat element centered at 12 500 bp in which the stem repeat was 52 bp in length, the similarity between the repeat copies was 100%, and the spacer sequence was 112 bp. pUC plus and M13 minus subclones were under-represented in region 2, which contained two inverted

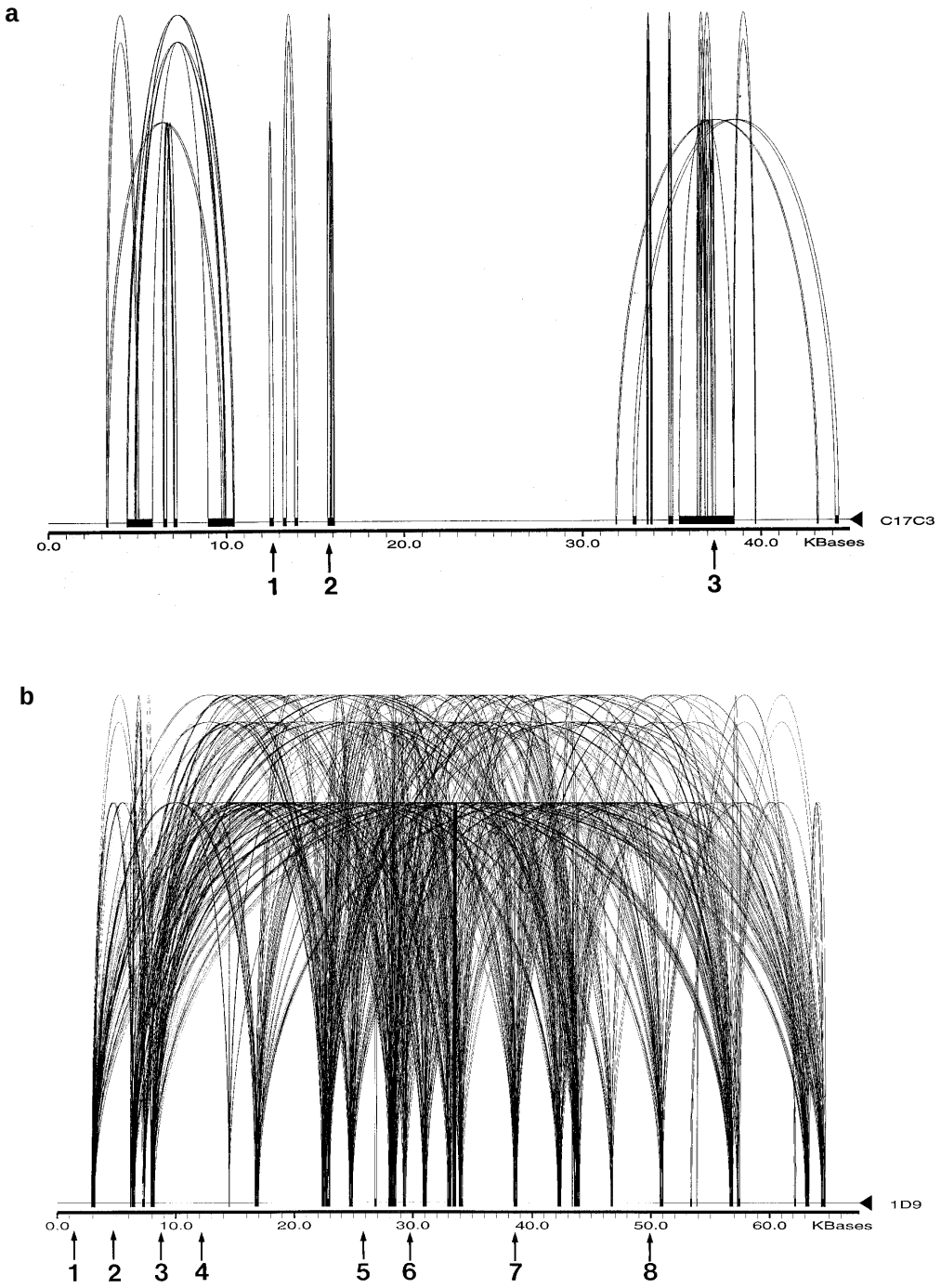

Figure 2. Repeated sequences. MIROPEATS graphical representation of sequences repeated within (**a**) the C17C3 sequence and (**b**) the 1D9 sequence. Inverted repeats were connected by an arched line extending to the top of the figure. Tandem or direct repeats were connected by an arched line of intermediate height. Regions along the sequence overlayed by a black box were regions contained within a repeat. Experimentally significant regions along the DNA sequence which lack subclone start-sites are indicated by arrows.

repeat elements centered at 15 000 and 15 900 bp and two tandem repeat regions centered at 15 800 and 16 000 bp. The first inverted repeat consisted of a 22 bp repeat with 90% sequence similarity separated by a 1051 bp spacer. The second inverted repeat element contained a stem repeat of 142 bp, separated by 93 bp, which exhibited 97% sequence similarity. Each of the two tandem repeats in region 2 contained three copies of a 22mer, and within each tandem region the highest level of sequence similarity between copies was 67 and 70%, respectively. Region 3 was under-represented by both M13 plus and minus subclones, and contained both an inverted and a tandem repeat region. The stem repeat in the inverted repeat was 1441 bp in length, the copies were separated by 173 bp, and they shared 99% sequence identity. The short tandem repeat region contained three copies of a 15mer

with 74% sequence similarity. In each region of under-representation a relevant pUC subclone occurred in at least one orientation. The under-representation in regions 1 and 3 occurred in both M13 plus and minus subclones although spanning pUC clones existed in both orientations. For region 2, both M13 and pUC subclones spanned the region since the under-representation affected only M13 minus and pUC plus subclones.

Table 2. Experimentally significant regions in the C17C3 and 1D9 sequences which lack subclone start-sites

	Region	Position	Subclone	Region
C17C3	1	11090–13964	M13-plus	11090–13428
			M13-minus	12525–13964
	2	14371–17352	pUC-plus	14371–17081
			M13-minus	15687–17352
	3	35698–38665	M13-plus	35698–37971
			M13-minus	36987–38665
1D9	1	630–2277	pUC-minus	630–2277
	2	3478–5780	pUC-minus	3478–5780
	3	7016–10209	pUC-minus	7016–8525
			pUC-plus	7940–10209
	4	10209–14133	pUC-plus	10209–12872
			pUC-minus	11621–13874
			M13-plus	12554–14133
	5	24098–27826	pUC-plus	24098–27826
	6	27921–31542	pUC-minus	27921–31542
			M13-minus	28076–29848
	7	37786–39285	pUC-plus	37786–39285
	8	48358–49799	pUC-plus	48358–49799

Each of the three regions of subclone under-representation discussed above contained inverted repeat elements which displayed high sequence similarities and an inter-repeat spacer sequence shorter than the average subclone insert length. Two of the inverted repeats implicated in subclone under-representation (from regions 1 and 3) exhibited almost perfect sequence identity (100 and 99%) and roughly the same distance between repeat copies (112 and 173 bp). However, the lengths of the stem repeat sequences were very different (52 versus 1441 bp). This suggests that the sequence similarity and inter-repeat spacer sequence length affected subclone representation more than did the length of the stem repeat sequences. Furthermore, although regions 2 and 3 did contain tandem repeat sequences in addition to inverted repeat sequences, a fourth tandem region which contained roughly the same level of sequence similarity was not associated with a region of under-representation. This suggests that the inverted repeats, rather than the tandem repeats, likely affected subclone representation in regions 2 and 3. Inverted repeat sequences which did not cause regions of experimentally significant under-representation either contained lower levels of sequence similarity or sufficiently large spacer sequences. For example, although the inverted repeat between 5101–5872 and 9738–8966 bp contained 98% sequence similarity, the repeats are separated by 3093 bp and therefore no single subclone would contain portions of both repeats.

Region 1 contained the first three 5′ exons of a gene (C17C3.10) predicted in the opposite orientation which exhibited similarity to the basic helix–loop–helix transcriptional repressor h gene. The two 3′ exons from this predicted gene were within the inverted repeat sequence from 9738–8896 bp and corresponding exons similarly were predicted within the other repeat copy from 5101 to 5872 bp. An acyl-CoA dehydrogenase gene (C17C3.12) was predicted within region 2 in the reverse-complement orientation. The inverted repeat sequences in region 3 contained two predicted genes which lacked database similarities, one in the forward orientation (C17C3.2) within one repeat copy and one in the reverse-complement orientation (C17C3.13) within the second repeat copy. Although each of these three regions of subclone under-representation contained at least some portion of a predicted gene, 14 total genes were predicted in this cosmid sequence and the remaining 10 genes were not associated with alterations in subclone start-site distribution.

Correlation of subclone start sites and under-represented DNA sequence in human BAC 1D9

The start-site positions of 919 M13 subclones (464 plus, 455 minus) and 437 pUC subclones (218 plus, 219 minus) were plotted as they occurred along the completed 1D9 insert sequence of 67 571 bp in Figure 1b. As for cosmid C17C3, the slope of the pUC subclone plot was higher than for the M13 subclone plot since fewer pUC subclones were sampled. The results of the chi-squared analysis comparison of the expected and observed distribution of subclone start-sites for 1D9 are listed in Table 1. There was not a significant difference between the pUC and M13 minus start-site distributions. However, for the pUC and M13 plus subclones, the M13 subclone distribution more closely approximated the Poisson distribution. As for C17C3, gaps in subclone start-sites >1450 bp were considered experimentally significant, and those regions of under-representation are listed in Table 2. Figure 2b depicts repeated sequences identified by MIROPEATS within the 1D9 sequence and Table 3b and 4b list the positions of those repeats as determined by INVERTED and TANDEM.

Regions 1 and 2 lacked pUC minus subclone start-sites which cannot be correlated with the occurrence of repeated sequences. However, the pUC (and M13) cloning vector contain the same *lacZ* region as the pBELOBAC11 cloning vector adjacent to the cloning site. It is possible that region 1 pUC minus subclones containing two copies of the same vector region, in a direct repeat orientation, were unstable. Region 3 was under-represented in pUC plus and minus subclone start-sites, and contained a 79 bp stem inverted repeat sequence with 78% sequence similarity and a 705 bp inter-repeat spacer sequence. These repeated sequences have similarity to the consensus Alu repeat sequence. Region 4 is immediately adjacent to region 3 and is under-represented by start-sites of all subclone types. This region contained an inverted repeat element with stem lengths of 100 bp, 73% sequence identity and a spacer sequence of 50 bp. There is a second inverted repeat within this region (12 200–12 224 and 13 944–13 920 bp). However, since the spacer sequence within this inverted repeat element (1695 bp) is greater than the average subclone insert (1450 bp), this repeat sequence would not be expected to contribute to under-representation in this region. Furthermore, this region was predicted to be a CpG island (11 169–13 675 bp with an overall GC content of 67.1%) and contained a putative exon confirmed by cDNA and BLASTX homologies. One cause of the general under-representation in this region likely resulted from technical difficulties encountered

sequencing these regions rather than clone stability issues. In support of this contention, investigation of sequences that had failed trace quality control measures identified additional subclones derived from this region. The 1D9 sequence was analyzed for other regions of low entropy sequences and for local deviations in base composition, dinucleotide frequency and trinucleotide frequency. Except for the CpG island region mentioned above, these features did not correlate to regions of subclone under-representation.

Table 3. Inverted repeat sequences[a]

Copy 1	Stem length	Copy 2	Distance between copies	Similarity (%)	Region
(a) C17C3					
3261–3344	84	5100–5017	1672	73	
5101–5872	772	9738–8966	3093	98 (1 gap)	
12453–12504	52	12668–12617	112	100	1
13217–13403	187	14044–13858	454	89	
14523–14544	22	15617–15596	1051	90	2
15717–15859	143	16095–15953	93	97	2
19083–19151	69	19824–19756	604	73	
20591–20620	30	20713–20684	63	90	
23520–23546	27	25037–25011	1464	88	
27999–28027	29	28102–28074	46	86	
35407–36847	1441	38461–37021	173	99	3
(b) 1D9					
3096–3163	68	7226–7159	3995	85	
6282–6317	36	7390–7355	1037	94	
6390–6477	88	7313–7225	747	80	
7228–7306	79	8090–8012	705	78	3
10199–10298	100	10446–10349	50	73	4
12200–12224	25	13944–13920	1695	88	
14500–14525	26	17000–16975	2449	100	
16473–16496	24	16999–16976	479	87	
22285–22586	302	28484–28183	5596	78	
22352–22585	234	29446–29214	6628	82	
26795–26832	38	26918–26881	48	84	5
27972–28113	142	28455–28314	200	76	
28157–28282	126	28577–28452	169	80	6
28501–28576	76	29263–29188	611	80	6
30894–31153	260	38748–38483	7329	84	
38467–38748	282	42447–42174	3425	85	
42231–42427	197	43920–43723	1295	80	
43766–43996	231	50992–50762	6765	74	
56581–56728	148	57421–57276	547	72	
57260–57387	128	64507–64379	6991	78	
57378–57432	55	57512–57460	27	86	
64240–64266	27	64695–64669	402	85	

[a]Calculated using INVERTED.

Table 4. Tandem repeat sequences[a]

Tandem repeat region	Copies	Similarity (%)	Region
(a) C17C3			
15754–15840	three copies of 22mer	67	2
15952–16038	three copies of 22mer	70	2
34652–34738	seven copies of 11mer	71	
36167–36225	three copies of 15mer	74	3
(b) 1D9			
12186–12224	nine copies of 4mer	88	
14462–14530	six copies of 10mer	71	
17750–17804	13 copies of 4mer	71	
21126–21208	20 copies of 4mer	71	
22622–22680	five copies of 10mer	72	
23986–24020	eight copies of 4mer	87	
33338–33620	70 copies of 4mer	80	
43322–43410	two copies of 30mer	71	
53832–53890	five copies of 10mer	72	
62454–62500	11 copies of 4mer	77	

[a]Tandem repeats calculated by TANDEM.

Regions 5 and 6 were adjacent but not overlapping, and lacked subclone start-sites in the pUC plus and in the pUC and M13 minus orientations, respectively. Region 5 contained an inverted repeat element with a 38 bp stem sequence, a 48 bp spacer sequence, and exhibited 84% sequence similarity. Region 6 contained two inverted repeat elements with stem lengths of 126 and 76 bp, and spacer sequence lengths of 169 and 611 bp, respectively. Both repeats exhibited 80% sequence similarity, and were similar to the Alu consensus sequence. The region between 27 979 and 28 576 bp contained three Alu repeats, with the flanking repeats in one orientation and the central repeat in the opposite orientation. Regions 7 and 8 lacked pUC plus subclone start-sites, although neither region could be correlated with the occurrence of repeated sequences.

There were 22 inverted repeats identified within the 1D9 sequence by INVERTED with stem sequence lengths ranging from 24 to 302 bp, with levels of sequence identity between 72 and 100%, and spacer sequences ranging from 27 to 7329 bp. Five of those inverted repeats were correlated with the lack of subclone start-sites in a defined region. Of those, three were similar to the consensus Alu sequence. Except for the inverted repeat elements which involved immediately adjacent Alu repeats in an inverted orientation, the presence of Alu sequences themselves were not correlated with subclone start-site under-representation. Additionally, the sequences with MER similarity and L1 similarity did not affect representation, but were not repeated within the 1D9 sequence.

In general, the 1D9 inverted repeat elements exhibited lower levels of sequence similarity than those noted in C17C3. It is unclear why some repeats were not correlated with under-representation. For example, the repeat from 6390–6477 and 7313–7225 bp which was not associated with subclone under-representation was very similar to the region 3 repeat from

7228–7306 and 8090–8012 bp which lacked pUC plus and minus subclones. However, the occurrence of that inverted repeat element within region 3 may have been coincidental. Several inverted repeats listed in Table 3b were not correlated with experimentally significant regions of subclone under-representation. However, each of the eight regions of subclone under-representation in 1D9 contained sequences which lacked subclone start-sites in either the pUC plus or minus orientation. This bias was indicated from the chi-squared analysis, discussed above, where the M13 subclone distribution was closer to random than the pUC subclone distribution.

CONCLUSION

Subclone start-site distributions were compared for two genomic DNA sequences represented in both M13 and pUC subclone vectors. The distributions were analyzed and correlated with DNA sequence repeats and coding features. Certain repeat sequences, most notably inverted repeat elements in the C17C3 cosmid sequence which contained short inter-repeat spacer sequences, resulted in under-representation of M13 subclones. In a few examples from the 1D9 sequence, under-representation of pUC subclones was correlated with inverted repeats. In some cases the subclone under-representation was strand specific, such that a subclone existed with a particular sequence only in one orientation. There were several regions, particularly in the 1D9 sequence, where regions lacking subclone start-sites were not correlated with the presence of inverted or tandem repeat sequences. Predicted coding sequences did not correlate with the subclone start-site distribution, although apparent under-representation occurred in a putative CpG island identified in the human genomic DNA sequence. This under-representation was due, at least in part, to difficulties obtaining quality sequence through this region. However, there was a repeat sequence identified within the predicted CpG island which also may have affected subclone representation. For the C17C3 sequence, the pUC subclone distribution more closely approximated a Poisson than the M13 subclone distribution, although this was not the case for the 1D9 sequence. Non-random representation similar to that described in this study has been observed in other *C.elegans* and human genomic clones, and in general is correlated with the occurrence of inverted repeat elements.

Due to the general success obtaining pUC subclones representing *C.elegans* repeat sequences which were under-represented in M13 subclones, we have incorporated pUC subclones into our sequencing paradigm. M13 subclones remain our initial choice for the random sequencing phase due to the inexpensive template preparation method and the robustness of the protocol, as well as the generally adequate sequence representation. Depending on the overall representation during the initial random sequencing phase, random pUC subclones may be usefully employed to recover the absent region. Since pUC subclone representation is not random as well, the complementary use of M13 and pUC subclone libraries provides a more effective approach than the singular use of either subclone type.

ACKNOWLEDGEMENTS

We gratefully acknowledge those in the *C.elegans* Genome Mapping and Sequencing Consortium, and specifically those at the Washington University Genome Sequencing Center for all aspects involved in the generation of the C17C3 and 1D9 sequence. We thank Dr E.R.Mardis for critical reading of the manuscript, and Dr M.Wendl for useful discussion. This work was supported by grant HG00958 from the NIH National Human Genome Research Institute (NHGRI). S.L.C. is supported by an NIH-NHGRI post-doctoral fellowship.

REFERENCES

1 Sulston,J., Du,Z., Thomas,K., Wilson,R., Hillier,L., Staden,R., Halloran,N., Green,P., Thierry-Mieg,J., Qiu,L., *et al.* (1992) *Nature* **356**, 37–41.
2 Wilson,R., Ainscough,R., Anderson,K., Baynes,C., Berks,M., Bonfield,J., Burton,J., Connell,M., Copsey,T., Cooper,J., *et al.* (1994) *Nature* **368**, 32–38.
3 Wilson,R.K. and Mardis,E.R. (1997) In *Genome Analysis: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, New York, NY, in press.
4 Mardis,E.R. (1994) *Nucleic Acids Res.* **22**, 2173–2175.
5 Clarke,L. and Carbon,J. (1976) *Cell* **9**, 91–101.
6 Edwards,A. and Caskey,C.T. (1991) *Methods: A Companion to Methods Enzymol.* **3**, 41–47.
7 Wyman,A.R. and Wertman,K.F. (1987) *Methods Enzymol.* **152**, 173–180.
8 Chen,E.Y. and Seeburg,P.H. (1985) *DNA* **4**, 165–170.
9 Wang,M., Chen,X., Shouse,S., Manson,J., Wu,Q., Li,R., Wrestler,J., Noya,D., Sun,Z., Korenberg,J. and Lai,E. (1994) *Genomics* **24**, 527–534.
10 Coulson,A.R., Sulston,J.E., Brenner,S. and Karn,J. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7821–7825.
11 Fulton,L.L. and Wilson,R.K. (1994) *BioTechniques* **17**, 298–301.
12 Dear,S. and Staden,R. (1991) *Nucleic Acids Res.* **19**, 3907–3911.
13 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.* **215**, 403–410.
14 Sokal,R.R. and Rohlf,F.J. (1981) *Biometry: The Principles and Practice of Statistics in Biological Research.* W.H. Freeman and Company, New York, pp. 27.
15 Jurka,J., Walichiewicz,J. and Milosavljevic,A. (1992) *J. Mol. Evol.* **35**, 286–291.
16 Parsons,J.D.(1995) *Comput. Applic. Biosci.* **11**, 615–619.
17 Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1995) In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Cambridge, UK, pp. 367–375.
18 Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) *Nucleic Acids Res.* **22**, 5156–5163.
19 Uberbacher,E.C. and Mural,R.J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 11261–11265.
20 Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J. Mol. Biol.* **220**, 49–65.
21 Thomas,A. and Skolnick,M.H. (1994) *IMA J. Math. Appl. Med. Biol.* **11**, 149–160.