Compilation of DNA strand exchange sites for non-homologous recombination in somatic cells

Andrzej K.Konopka

National Institutes of Health, National Cancer Institute, Laboratory of Mathematical Biology, FCRF, Building 469, Room 151, Frederick, MD 21701, USA

Received September 12, 1987; Revised and Accepted December 19, 1987

ABSTRACT

DNA sequences of 496 somatic cell illegitimate crossing over regions were compiled and analyzed. Sites for non-homologous recombination on linear DNAs transfected into mammalian cells (Transfected Linear DNAs; TLD) were analyzed separately from the remaining illegitimate recombination regions (IRR). Trinucleotides that are preferentially cleaved by rat liver topoisomerase I in vitro (CAT, CTY, GTY, RAT where R = purine, Y = pyrimidine) were present in the 10 base pair (bp) vicinity of the cross-over sites in 92% of IRR and 93% of TLD. Multiple repeats of these trinucleotides have been observed in 39% of IRR and 38% of TLD. Runs of five or more contiguous purines (or pyrimidines on the complementary strand) were found in 26% of IRR and 14% of TLD. Adenine-Thymine rich regions of five or more bases were found in 14% of IRR and 21% of TLD. Alternating purine-pyrimidine tracks longer than four nucleotides in length were found in 11% of IRR, though only in 4% of TLD. I discuss the possible biological significance of these findings and present an appendix containing the sequences in the 10 bp vicinity of the non-homologous recombination sites analyzed.

INTRODUCTION

Studies of recombination (crossing over) at the molecular level have relatively recent history. The most known kind of crossing over, homologous recombination, mediates genetic exchange between DNA molecules that share extensive sequence homology. Rare DNA rearrangements that require no extended sequence homology and can occur independently of the enzymes usually involved in homologous recombination have been discovered 20 years ago in bacteria (1). These so called non-homologous or illegitimate recombination events also occur in eukaryotic cells (2-81) where they cause severe genetic alterations of chromosomal fragments (inversions, deletions, transpositions, duplications, amplifications etc.).

In this study, phosphodiester bonds broken in the course of recombination will be referred to as non-homologous cross-over points. Frequently, it is impossible to establish the exact non-homologous cross-over points since both parents share small stretches of sequence.

In these cases, the sequence ambiguities define the limits within which crossing over occurred and will be referred to as non-homologous cross-over sites. Any given cross-over point (or site), together with flanking 5' and 3' sequences, will be referred to as the non-homologous cross-over region.

One sequence feature of non-homologous cross-over regions, observed in initial studies (33,36,38), is the presence of adenine-thymine rich DNA. Several other studies have suggested that sequences either at, or very close to, the cross-over sites are important in these recombination events. This suggestion originated, in part, from the observation that trinucleotides that are cleavage sites for topoisomerase I in vitro (83,84) are associated with certain non-homologous cross-over sites (5,43,49,79). It has also been shown that cross-over sites used during SV40 excision from the 14B cell line correspond to topoisomerase I cleavage sites in vitro (85).

Another feature that has been suggested to play a role in illegitimate recombination are short regions of partial homology, flanking the cross-over sites, that are shared by both parents (36, 38, 50). However, arguments have been raised against the involvement of patchy homologies in non-homologous recombination (5, 82).

All the studies mentioned above have been performed on a very limited number of sequences and therefore could not be conclusive. In the absence of a direct assay for sequence requirements of illegitimate crossing over, I decided to compile sequences known to have been used during non-homologous recombination in somatic cells, and to search them for common motifs.

METHODS

Sequences analyzed. The sequences compiled and analyzed in this study are derived from diverse forms of non-homologous recombination in somatic cells and represent an extensive survey of available literature (2-81). Among the examples cited are rearrangements associated with integration events into host chromosomes by Polyoma virus (2-4,79), SV40 (5-10), Hepatitis B virus (11,12) and Adenovirus (10,13-18); as well as viral excision (5,19) and proviral deletion events (20,21). I have also compiled structural variants (generated during passage in cells) of viruses such as polyoma (22-30), SV40 (31-40), Epstein-Barr (41), Cowpox (42), Minute virus of mice (43) and Ad2-SV40 hybrids (44-47). Other examples of illegitimate

recombination regions present in our data base include rearrangements formed during transfections of linear (48-54) and circular (49,55,56) viral containing DNAs into somatic cells in culture. Non-viral rearrangements include deletions in the B2-macroglobulin (57), collagen (58,59), LDL receptor (60,61), APRT (62,63), p53 (64) and myelin genes (65). Furthermore, various deletions in the human β -globin (66-71) and human α -globin gene clusters (72-74) are presented, as is a deletion that resulted in a C γ to C δ immunoglobulin class switch (75). Finally, our data base includes rearrangements associated with transposition of an enhancer (76), gene amplification events (77,78), non-homologous recombinations occurring during formation of transgenic mice (79), a duplication within the heptaglobin Hp2 gene (80) and fusion between the human tpr locus on chromosome 1 and the met locus on chromosome 7 (81).

Ninety-eight of the parental sequences compiled in this study are the result of recombination events that occurred during transfections of linear DNAs (48-54) and I have analyzed them separately from the remaining 398 parental sequences. This step was taken in view of the observation that many non-homologous recombination events involving transfected linear DNA reportedly arise via blunt-end ligations (48,90).

Not included in this study are examples of oncogene translocations into immunoglobulin switch regions. These rearrangements are thought to involve an endonuclease that is normally associated with immunoglobulin site specific recombination (86). Also omitted are studies of the rearrangements that formed acutely transforming retroviruses. These events might have occurred in the context of proviral integration and an endonuclease, encoded by pol gene of retroviruses, has been shown to be required for viral integration (87-89).

Alignment. A total of 496 cross-over regions was analyzed. The exact phosphodiester bonds broken during the various recombination events (cross-over point) could not be determined in 297 of these sequences for two reasons. First, in certain cases only one of the two parental sequences were known. Second, in those cases where both parents were known, a short region of homology at the cross-over sites often precluded identification of the cross-over points. The length of this region of uncertainty (cross-over site) did not exceed 4 nucleotides in over 95% of parental sequences from our collection.

Sequences from both strands of the parental cross-over regions were aligned in the following way. If there was no uncertainty in the location

of the cross-over point, the nucleotide preceding the break was assigned position -1; the nucleotide after the break was assigned +1. If there was an uncertainty, the nucleotide just preceding the uncertainty was assigned position -1 and the first nucleotide in the uncertain region +1.

Search for various sequence motifs. The aligned sequences were searched for the occurrence of the following motifs:

1) Trinucleotides that are topoisomerase I cleavage sites in vitro. Trinucleotides used in this study (CAT, CTY, GTY, RAT; R = purine, Y = pyrimidine) include only those rat liver topoisomerase I cleavage sites that were cleaved greater than 50% of the time upon disruption of cycles of nicking and closing by detergent (83). Furthermore, since wheat germ topoisomerase I generated virtually the same consensus sequence as rat liver topoisomerase I (83) I have assumed that the rat liver topoisomerase I consensus sequence is representative of all higher eukaryotic type I topoisomerases.

1a) Topoisomerase I sites (Topo I*) having the 3' most base in the window from -1 to the end of uncertainty region. These sites could be directly involved in topoisomerase I mediated illegitimate recombination.

1b) Multiple Topoisomerase I cleavage sites.

2) Runs of five or more bases consisting solely of adenine and thymine.

3) Runs of alternating purines and pyrimidines not shorter than five bases in length. Such runs are known to be very underrepresented in nucleic acids sequences data banks (91, 92, 94). Nevertheless, they have been reported to promote homologous recombination in somatic cells (93).

4) Runs of five or more purines (or pyrimidines).

5) Combination of 1) with 2), 3) and 4).

6) Combination of 1a) with 2), 3) and 4).

With the exception of 1a) and 6), all searches were performed within positions -5 to -1 and from +1 to r+5 (r is the length of the uncertainty) for every sequence in the collection. For a given motif the total number of matches F was determined. To evaluate statistical significance of F we generated fifty collections of random sequences that preserved the dinucleotide composition of the actual sequences studied. In the collections of random sequences the distribution of frequencies of a given motif was approximately gaussian.

Therefore, I used the z score $z = (F - F_e) / \sigma$ (where F is the actual frequency, F_e is the mean frequency from the searches of 50 collections of random sequences and σ is the corresponding standard deviation) to evaluate

the statistical significance of frequencies F . I assumed a level of significance 0.002. Whenever z was greater than the critical (at this level of significance) value 3.08, F could be considered significantly different from the expected value F_e .

I also determined the actual and expected frequencies of occurrence of the different topoisomerase I cleavage sites at every position from -5 to + 5. By convention, the occurrence of a trinucleotide with its 5' most nucleotide at position -3 and ending at position -1 was scored at position -1. Similarly, a triplet extending from position -2 to +1 was scored at position +1, etc.. I then calculated values of the z -score: $z(p) = (f(p) - fe(p))/\sigma(p)$; where $f(p)$ and $fe(p)$ stand for the actual and expected frequencies of any given trinucleotide at a given position p and $\sigma(p)$ is the standard deviation of the distribution $D(fe(p), \sigma(p))$. I assumed that the occurrence of a particular trinucleotide at position p is non-significant if $z(p)$ was lower than +1.96 and significant if $z(p)$ was greater than +1.96 (critical value of the z score at the level of significance 0.05).

RESULTS AND DISCUSSION

The results of searching for various sequence motifs in the collections of IRR and TLD are presented in Table 1. Sequences in the 10 bp vicinity of the cross-over sites studied are presented in the Appendix.

It can be seen from Table 1 and the Appendix that in both IRR and TLD triplets corresponding to topoisomerase I cleavage sites are present in over 92% of the examples within the 10 bp vicinity of the cross-over sites. It can also be seen that in 38% of IRR and 32% of TLD the 3' most nucleotide of the potential topoisomerase I cleavage sites (topo I*) occurs in the region from -1 to the end of uncertainty (i.e. at phosphodiester bonds that may have been directly involved in the recombination events). It is apparent from Table 1 that relative to the collection of random sequences, topo I and topo I* sites are significantly overrepresented in the cross - over regions studied. Another observation which appears from Table 1 and the Appendix is multiple occurrence of topoisomerase I cleavage sites. In statistically significant portions of IRR (39%) and TLD (38%) these sites occurred two or more times in the vicinity of cross-over point.

The cross-over regions containing either purine (pyrimidine) runs or sequences rich in adenine and thymine are significantly overrepresented and are always associated with a potential topoisomerase I site (see Table 1).

MOTIF	ILLEGITIMATE RECOMB.			TRANSFECTED LINEAR DNAs		
	REGIONS (398 sequences)			(98 sequences)		
	NB.of seq.		z score	NB. of seq.		z score
	act	exp		act	exp	
Topo I	368	196.12	11.49 s	91	46.27	8.63 s
Topo I*	154	64.34	9.25 s	32	14.13	6.81 s
2 or more Topo I	156	53.45	10.26 s	38	16.56	5.50 s
Rn	102	23.28	12.17 s	14	5.18	4.02 s
Rn and Topo I	102	11.47	25.23 s	14	3.04	6.35 s
Rn and Topo I*	33	4.06	12.34 s	7	1.39	6.00 s
Y/R	43	19.16	5.51 s	4	3.11	0.58 ns
Y/R and Topo I	17	10.01	2.21 ns	4	2.15	1.41 ns
Y/R and Topo I*	8	4.32	2.00 ns	1	1.03	0.00 ns
A(T) rich	54	22.63	6.28 s	21	8.38	4.59 s
A(T) and Topo I	53	12.54	9.48 s	21	4.30	8.50 s
A(T) and Topo I*	23	4.23	7.73 s	8	2.03	4.24 s

Table 1. Results of searching for selected sequence motifs in the 10 bp vicinity of the cross-over sites (or cross-over points). The z-score has been used to decide whether overrepresentation of a given motif is significant. The symbol s appearing after the numerical value of the z-score means that the actual frequency significantly differs from what is expected in a random case. (The symbol ns means that the difference between the actual and the expected frequency is not significant at the level of significance 0.2%; see Methods section for more details regarding the evaluation of significance). Topo I - at least one topo I site in the region. Topo I* - at least one topo I site having its 3' most base within positions -1 to the end of uncertainty. Rn - n subsequent purines (n greater or equal 5). Y/R - tracks of alternating purines and pyrimidines not shorter than 5 bases in length. A(T) - tracks of 5 or more bases consisting solely of adenines or thymines.

The biological significance of overrepresentation of motifs rich in A and T is not clear since such sequences are overrepresented in many entries selected randomly from a data bank (Konopka, Trifonov and Maizel; in preparation). However, runs of purines (pyrimidines) might be important for four reasons. First, they can assume unusual conformations that are unable to bind histones to form nucleosomes (95, 96). Second, homopurine (homopyrimidine) tracks are found at S1 hypersensitive sites and reportedly can adopt altered (non-B) conformations both *in vitro* and *in vivo* (97 - 100). Third, they are present in centromeres of chromosomes (96) and in almost all known intron sequences (101). Fourth, runs of purines

STRING	Positions of significant occurrence ¹	
	IRR	TLD
AAT	+4	-8,-7,+9
CAT	+1	-3,+11
CTC	-1	-6,-5,-1,+4,+7
CTT	+3	+6
GAT	+7*	+8*
GTC	+9	+7,+10
GTT	-1,+8	+1,+5

Table 2. Positions of significant occurrence (relative to crossover points) of trinucleotides that are preferential rat liver topoisomerase I cleavage sites in vitro. IRR; illegitimate recombination regions. TLD; transfected linear DNAs.

(pyrimidines) can promote recombination events that occur during formation of core regions of hypervariable minisatellite DNAs (102 - 103).

Inspection of Table 1 shows that alternating repeats of purines and pyrimidines are significantly overrepresented in IRR but not in TLD. It can also be seen in Table 1 that the occurrence of alternating purine - pyrimidine tracks is generally not associated with the presence of topoisomerase I cleavage trinucleotides. Two suggestions can be made on the basis of these observations. First, the pathways of non-homologous recombination that involve TLD might be different than those involving IRR. Second, unusual DNA structures can be involved in those instances of illegitimate recombination which are topoisomerase I independent.

The suggestion that many topoisomerase I independent (or indirectly dependent) pathways of illegitimate recombination are likely to exist also appears from the analysis of occurrence of topoisomerase I cleavage sites in different positions relative to the cross-over point (Table 2). It can be seen from this Table 2 that AAT, CAT, CTC, CTT and GTT are significantly overrepresented between positions -1 and +4 (i.e are Topo I* sites) in IRR, whereas in TLD only CTC and GTT are overrepresented in this region. If topoisomerase I would break the DNA and catalyze the linkage of the end to which it is attached to another free end (the other end can be generated by topoisomerase I cleavage or by any other mechanism), we should expect that topoisomerase I cleavage sites will be present within the few bases 5' to the actual crossover point. Results from Table 2 suggest that such a mechanism would not be possible in many instances of illegitimate recombination (particularly TLD). However, topoisomerase I mediated

breakage of DNA followed by either trimming or extension of the non-topoisomerase coupled acceptor ends (85) could still be possible and might enable both direct and indirect roles for this enzyme in non-homologous recombination.

The following conclusions can be drawn from this study:

1) In vitro topoisomerase I cleavage sites are present in a vast majority of the in vivo non-homologous recombination regions studied. This suggests that topoisomerase I may play a role in many somatic cell non-homologous recombination events.

2) The association of topo I sites with runs of purines (or pyrimidines) and A/T rich regions suggests that unusual DNA structures can be involved in many instances of illegitimate recombination.

3) Relative to the position of the cross-over sites (or cross-over points), many potential topoisomerase I cleavage sites are positioned in ways that do not suggest direct roles of this enzyme in breakage and rejoining events.

4) More than one pathway for illegitimate recombination in somatic cells is likely to exist.

Obviously, determination of the mechanism(s) for non-homologous recombination in somatic cells will require further studies. I believe that the compilation presented in the Appendix will be helpful in these future studies. The data base of sequences compiled in this study will be available under request.

ACKNOWLEDGEMENTS

I am indebted to Drs. J.V. Maizel and R.J. Fisher for helpful discussions and to Mr. G.W. Smythers for his expert assistance. This work was done in strict collaboration with Dr. P.A. Bullock whom I would like to thank for providing published sequences of illegitimate recombination regions and for discussions during preparation of the manuscript.

APPENDIX

SEQUENCE	CHARACTERISTICS	REFERENCE	SEQUENCE	CHARACTERISTICS	REFERENCE
tgaaa aggga acttt TCcct	high R, Topo I		2 ctaac aaaaa gattg ttttt	high A-T, Topo I	45
acgaa gaggt tgcTT Cccca	high R, Topo I		3 GTTga GTCgc caact Cagcg	Topo I	45
ttCTT cagaa aagaa gtcctt	high Y, Topo I*		3 ctgCA TAAAT gacgt attta	high Y/R and A-T, Topo I	45
atatac acaaa taTAG tgttt	Topo I		3 cccgc gagct gggcg CTCga	Topo I*	45
caaaa cgacg gttTT GCTGc	Topo I		3 ggtGT Tgggc ccaca acccg	Topo I	45
aaCAT ggGT Caaaa ttgTA Ccca gtttt	Topo I*		4 agcgc caCTC tcgcg gtgag	Topo I	45
ttatg ggggt attat aaTAC ccca TAAta	high A-T, Topo I*		4 ttaaa aaAAT aattt tttta	high A-T, Topo I	45
aCATg CTCc cccaa tgTAC gagg ggggt	high R/Y and Y, Topo I*		4 CTTaG TTTat gaatg aaata	high A-T, Topo I	45
agTct CTCc TCagg tcaga gagg agtcc	high Y, Topo I*		4 AATgg aaGAT tTACc TTCTa	high R, Topo I*	45
ttaca gac gagcc aatgt CTG CTCgg	Topo I*		4 tcaaa ATccc agttt TAGgg	high A-T, Topo I*	45
gaCTT gaC ATttt CTGaa CTG TAAaa	high A-T, Topo I*		4 attct a CTcct TAAga t gagga	high Y, Topo I	46
ataCA Tca CTTca tatgt AGt gaagt	high R/Y, Topo I*		4 atgga a GATgg TACct T CTACC	high R, Topo I	46
AATcC Tca AATgg tTAGg agt tTACC	Topo I*		5 CTTac TTTat gaatg aaata	high A-T, Topo I	46
ggaga aC TTTga cCTCT TG aaact	high R, Topo I		5 ttaaa aaAAT aattt tttta	high A-T, Topo I	46
cccag aC ATTgg gggtc TG TAAcc	Topo I*		5 gccac aa agaca cgggt tT Tctgt	high R, Topo I*	47
CATCA TCT Ttcag gTAGT AGa aagtc	high Y, Topo I*		5 CATaa AA TgAAT gtatt tT ACtta	high A-T, Topo I*	47
tCATG Tct gGATc agTAC aga cCTAG	Topo I*		5 ttgtA A Tttac aacat T Aaatg	high A-T, Topo I*	47
caATC TT GATcc gtTAG aa cTAGg	Topo I*		5 tGTTG A Tactg acaac t atgac	Topo I	47
agAAT tt CATtc tctTA Aa gTAAg	Topo I*		5 acAAT aaATt tgtta ttTAA	high A-T, Topo I*	47
acctg caA Tgcaa tggac gtT ACgtt	Topo I*		5 atagg ctgcc tatcc gacgg		47
ttCTC caA ATctg aagag gtt TAGac	high Y, Topo I*		5 acggt gcc gagga tgcca cgg CTCct	high R, Topo I*	47
gcCTT aa ctaga cggaa TT Gatct	Topo I*		5 aaagg gcc TCgtg tTTCc cgg agcac	high R, Topo I	47
caggg aa gcctg gtccc TT Cggac	high R, Topo I*		5 atcGA T tgCTT TAGct A Acgaa	Topo I*	47
gCATa gaa GTTaa cgtat cTT Caatt	Topo I*		5 tttCT T agacg aaaga a tCTGc	high Y, Topo I*	47
tgAAT gaa ccaca actTA CTT Ggtgt	Topo I*		5 gcctg agtgt cggaC TCaca	Topo I	48
GTTtG TTTta caaac aaaat	high A-T, Topo I		6 tacta actgc atgat TGacg	Topo I	48
actGT Tgaag tgaca actTC	Topo I		6 gGATg aagca ccTAC TTCgt	Topo I*	48
caGTC actaa gtcag tgatt	Topo I*		7 ttAAT GTCat aatTA Cagta	high A-T, Topo I*	48

SEQUENCE	CHARACTERISTICS	REFERENCE	SEQUENCE	CHARACTERISTICS	REFERENCE
cCTCa AATgg ggagt tTACc	Topo I	7	ttata gGTTa aatat ccaat	high A-T, Topo I	48
acaca gcagt tgtgt cgtca	high R/Y	8	gcctg agtgt cggaC TCaca	Topo I	48
ttctg GTTtt aagac caaaa	Topo I	8	cgggg aAA7g gcccc ttTAC	high R, Topo I	48
aGTTt aGTct tcaaa tcaga	Topo I	8	atcag CAT ttctc TAGtc gTA Aaagg	high Y, Topo I*	49
ttCTT tttgt aagaa aaaca	high Y, Topo I*	8	gaCTC a aaaaa CTGag t ttttt	high A-T, Topo I*	49
ttagt gactg aatca CTGac	Topo I*	8	caacC T Ctggag gTTGg a gaCTC	high Y, Topo I	49
cCTCa AATgg ggagt tTACc	Topo I	8	tgcaa g tgcca acgTT C acggt	Topo I	49
gtgtg tgtgt gtgtg cacac acaca cacac	high R/Y	9	tCATg aacag agTAC TTGtc	Topo I*	49
gacCT TgtGT CGTca CTGga acaca gcagt	high R/Y, Topo I*	9	aAATg G ATCTC ttTAC c TAGag	Topo I*	49
CA TCATc gT AGTAG	Topo I	10	atggg g cggag TACcc c gcCTC	Topo I	49
atcCT Tcgcg TAGga agcgc	Topo I	11	aGTTg C ATccc tcaac g TAGgg	Topo I*	49
atCTC caCTTcC TCccc TAGag gtgaagg agggg	high Y, Topo I*	12	atgCA T gCTTt TACgT A Cgaaa	high R/Y, Topo I*	49
aGTTt gtgac tcaaa caCTG	Topo I	13	aagCA Ta tgcag TTCgt aT ACgtg	high R/Y, Topo I*	49
cccta aaacc gggat tTTGg	high A-T, Topo I	13	gaGTT ag gggcg CTCaa tc cccgc	Topo I*	49
tatta tctgc aTAAT AGacg	high A-T, Topo I	14	tCATT a AATct agTAA t tTAGa	high A-T, Topo I	49
cGATa gcagt gctat cgtca	Topo I	14	aggcc t atgct tccgg a TACga	Topo I*	49
acgcC ATttt tgcgg TAAAA	high A-T, Topo I*	15	atgac t GTctt TACTg a cagaa	Topo I	49
cgccc TCtcc gcggg agagg	high Y, Topo I	15	cccGT Ca gggcg gggca gt cccgc	Topo I*	49
ttaaA ATgaa aattt TACTt	high A-T, Topo I*	16	AATtg g aagTT tTAAc c TTCaa	Topo I*	49
CATAA Tatac gtatt atatg	high Y/R and A-T, Topo I	17	tacca CAT ctatt atggt gTA GaTAA	Topo I*	49
atatt atGAT taTAA TACTa	high A-T, Topo I*	17	ttcaa a CTTtt aagTT T Gaaaa	high Y, Topo I	49
ggcga gCTTg ccgCT Cgaac	Topo I	17	tctat t gaCAT agaTA A CTGta	Topo I*	49
cagac tattt gtCTG aTAAa	high A-T, Topo I	17	CTTct g gtGAT gaaga c cacta	high Y, Topo I	49
agCTT at CTCTc tcgaa TA Gagag	high Y, Topo I*	17	tAATc gGTCT atTAG ccaga	Topo I	49
gaagg gccta cTTCc cggtat	high R, Topo I	18	gaaga g cCATa cTTCT c ggtat	high R, Topo I	49
gtgtg tAATt cacac atTAA	high A-T and Y/R, Topo I	18	cgccc g gcCTC gcggg c cggag	Topo I	49
atata tAATa tatat attat	high A-T, Topo I	18	ccgca c tgaag ggcgt g acTTC	Topo I	49
aaacc c tgagc tTTGg g aCTCg	Topo I	19	gacCT T tgGTC CTGga a accag	Topo I*	49

ccaCT C CATtt ggtga g gTAAa	Topo I*	19	cacgg ta aGTCc gtgcc aT TCagg	Topo I*	49
aaaa g aaaa g ttTTC ttTTC	high R, Topo I	20	aacaa ag accaa TTGTT TC tggtt	Topo I	49
ggagc ctggg cCTCg gaccc	Topo I	20	ctGTT ttgac gacaa aACTG	Topo I*	
AATgc aag gaGTT tTACg TTC CTCaa	high R, Topo I*	21	tttGT Cag CTCtg aaaca gtc gagac	Topo I*	49
tttat aag taggc aaata TTC atccg	high A-T, Topo I*	21	aaggg t agaca TTCcc a tCTGt	high R, Topo I	49
tctGT T atgag agaca a TACTc	Topo I*	22	ccacc CA Tccgc ggtgg gT AGgcg	Topo I*	49
GTTgc t gagcc caacg a CTCgg	Topo I*	22	tatat g cctgc ataTA C ggcacg	high A-T and Y/R, Topo I	49
agcga cc aagac tcgCT Gg TTCTg	Topo I*	22	cagcc gcacg gtcgg cgtgc	high Y/R	50
CTCat cc cccccc gagTA Gg gggggg	high Y, Topo I	22	ggcgc ctaca ccgcg gatgt		50
aaqta ccaaa TTCat ggttt	Topo I	23	gcggc ggcgc cgccg ccgcg		50
aaqac a cCTTt TTCTg t ggaaa	high Y, Topo I	23	gGATa tGTTc cctaT ACAag	high R/Y, Topo I	50
tctgc a gAATg agcgc t cTAC	Topo I	23	tcccc GTCgg agggg cagcc	high Y, Topo I	50
ctagt a AATgt gatca t tTACa	Topo I	23	ccccc GTCgt ggggg cagca	high Y, Topo I	50
gacaA A Tgtga GTGtt T ACact	Topo I*	23	ggaga AATac cCTCt ttatg	high R, Topo I	50
GTTtt gAATg caaaa ctTAC	Topo I	23	AATGt TgAAT tTACa actTA	Topo I	50
caagC A Tgccta gTTCg T ACggat	high Y/R, Topo I*	23	GATat aggcg ctata tccgc	high A-T, Topo I	50
tgtgc a cagcg acacg t gtcgc	high Y/R	23	ttagg GTTg aatcc caaac	Topo I	50
tagta AATgt atcat tTACa	high A-T, Topo I	24	ttaAA Ttgct aattT AACga	high A-T, Topo I	50
aaqCA Tgcct TTCgT ACgga	high Y/R, Topo I	24	aCTCa cGTTa tgagt gcaat	Topo I	50
agaag gc taCAT tcTTC cg atgta	high R, Topo I	25	gtgaa g gaacc cacTT C cTTGg	high R, Topo I	51
agaaa gc aaCAT tcTtT Cg TTGta	high R, Topo I*	25	tatag G TTAAT atatc c aatta	high A-T, Topo I	51
CTTgt G TTgct gaaca c aacga	high Y/R, Topo I	26	gtGTT GATgc cacaa cTACg	Topo I*	52
accaa g accag tggTT C tggtc	Topo I	26	tgtgt tgaca accac aCTGt	Topo I	52
aaCTT G TCaaa TTGaa c agttt	Topo I*	27	ctgca aacaa gacgt TTGtt	Topo I*	52
tttca g cCTCa aaagt c ggagt	Topo I	27	taCTT ctgtg atgaa gacac	Topo I*	52
aaCTT G TCaaa TTGaa c agttt	Topo I*	27	tGTTG A Tgcaa acaac T ACgtt	Topo I*	52
ccgcC T Cattt ggcgg a gTAAa	Topo I	27	aaggt a AATat TTCca t ttata	high A-T, Topo I	52
CTTgg aggcg gaacC TCcgc	high R, Topo I	27	ataca g gaaag tatgt c ctTTC	high R, Topo I	52
cacCA TCATg gtggT AGTAC	Topo I	27	atccc g tGTTg TAGgg c acaac	Topo I	52

SEQUENCE	CHARACTERISTICS	REFERENCE	SEQUENCE	CHARACTERISTICS	REFERENCE
attgt agagg TAAca tCTCc	high R, Topo I		28 ctgtg CT Tactg gacac ga atgac	Topo I	52
gaggc gctgt CTCc gacac	Topo I		28 aacta ct GATTc TTGat ga cTAAg	Topo I	52
aggCT T ccaga tccga a ggtct	Topo I*		29 ttctc gtGT Cgccc aaagg caca gcggg	Topo I	52
ctaac t atgac gaTTG a TACTg	Topo I*		29 tGTTt gtgt atttt acaaa caca TAAAA	high R/Y and A-T, Topo I*	52
gCTTc cag aggca cgaag gTC TCCgt	Topo I*		29 ggcaC TTTt cgggg ccgtg aaaa gcccc	high Y, Topo I*	52
caaGT CaG TTAag gTTCa gtc aaTTC	Topo I*		29 taaaA Tttt taagt atttT AAAa aTTCa	high A-T, Topo I*	52
gcCTC c aAATa cggag g tttat	high A-T, Topo I*		29 tatga aGATg aTACT TCTAC	Topo I	53
actaa c tgacc tgaTT G aCTGg	Topo I		29 taggg TTtat atccc aaata	high A-T, Topo I	53
gcaag CA Tgcct cgTTC gT ACgga	high Y/R, Topo I*		29 caggc CATgg gtccg gTACc	Topo I	53
AATGT Ca aGTCa tTACA gT TCagt	Topo I*		29 tgtGT TtGTT acaca aacaa	Topo I	53
aaqca g gaagt TTCgt c cTTCa	high R, Topo I		29 gtgtg gc gccgg cacga cg cggcc		53
ggagg g cCTCc cCTCc c ggagg	high R and Y, Topo I		29 GTCaC gc caggc cagt g gcgtc	Topo I	53
caaaa agc CTCTc gtttT TCg gagag	high R and Y, Topo I		29 tgctg cgcc ggctg acgac gcgg ccgac		53
acaac agc tGTTt tgTTG tcg acaaa	Topo I		29 tGTCa cgcc aggcc acagt gcgg tccgg	Topo I	53
agaaa a caacc tcctt T GTTGg	high R, Topo I		29 aggcC ATggt tccgg TACca	Topo I*	53
aaaggc a GTCac TTCcg t cagt g	Topo I		29 cggct GTCac gccga cagt g	Topo I	53
gAATA tCTCc cttat AGagg	high Y and A-T, Topo I		29 CATgg t gctgc gTACc a cgacg	Topo I	53
tcccg aCTCt aggcC TGaga	Topo I		29 tccGT T aaggt aggca a TTCca	Topo I*	53
taagc acagg aTTCg tgtcc	Topo I		29 aGTT cCTC ttcta tcaa ggag aaagt	high Y, Topo I*	53
aggaa gtgac tccTT CaCTG	high R, Topo I		29 ccagg cCTC cGTTa ggtcc ggag gcaat	Topo I*	53
GATta gctga cTAAc cgact	Topo I		30 CATC Tttgc gTAG aaacg	Topo I	54
aagag ga aaccc TTCtc ct TTGgg	high R, Topo I*		30 tgcaa agCTT acgtT TCgaa	Topo I	54
GTttt ga gCTTg caaaa CT Cgaac	Topo I*		30 CTTtg caaa gCTTt gaac gtTT Cgaaa	Topo I*	54
aaaaa gc acagc ttttT Cg tgtcg	high R and A-T, Topo I		30 gcCTC caaa aaagc cggag gttt TTTCg	high A-T, Topo I*	54
tcgct gc cCTTg agcga cg ggaac	high R, Topo I		30 agCTT ttctg tcgaa aaacg	high A-T, Topo I*	54
tttca gagga aaagt CTCct	high R, Topo I*		30 cCTCc aaaaa ggagg ttttt	high A-T and Y, Topo I	54
attat gag gaggg TAATA Ctc CTCcc	high R and A-T, Topo I*		30 ttttg caaaa gccta aaaac gtttt Cggat	high R, Topo I*	54
GTCag gag tgact cagtc CTC aCTGa	high R, Topo I*		30 gcCTC caaaa aagcc cggag gtttt TTCgg	high A-T, Topo I*	54

caaCA TaCTT gTTGt atgaa	high Y/R, Topo I	30	aaaag cct aggcc ttTTC gga tccgg	high R, Topo I	54
gccct agAAT cggga tctta	Topo I	30	agcCT CCT Cacta tcgga gga gtgat	high Y, Topo I*	54
ggagt gaCTC cCTCa CTGag	Topo I*	30	aaaAT taGTC tttTA Atcag	high A-T, Topo I*	54
tgctg act AATtg acgaC TGA tTAAc	Topo I	31	gAATg ggcgg ctTAC ccgcc	Topo I	54
cccta act gacac gggaT TGA CTGtg	Topo I*	31	taGTC agcca atcag tccgt	Topo I*	54
ctaac c aaGTT gaTTG g TTCaa	Topo I*	31	AATgg gcgga tTACc cgcct	Topo I	54
tgctg c gccgg acgac g cggcc	high Y/R	31	cggag AATgg gcCTC tTACc	high R, Topo I	54
tgaAA Taacc acttt aTTGg	high A-T, Topo I	32	gaact gggcg cTTGa cccgc	Topo I	54
aggCT Cc ccagc tccga gg ggtcg	high R, Topo I*	32	gggc gga actgg cccg cct TGacc	Topo I*	54
cccta a CTcCg gggaT T Gaggg	Topo I	33	tgggc gga GTTag acccg cCT Caatc	Topo I*	54
ttggc a aaagg aaacc t tTTCc	high R, Topo I	33	AATgg g cggaa tTACc c gcctt	Topo I	54
acaga gagga tgtCT Ctcct	high R, Topo I	33	tgggc g gaGTT acccg c CTcAa	Topo I*	54
gGTTa tttca ccaaT AAagt	high A-T, Topo I	33	tgtgt gtgtg acaca cacac	high Y/R	55
atgaa gaaaa TACTT Ctittt	high R, Topo I	33	gtaac CATta caTTG gTAAat	Topo I	55
atttc a ggcca TAAag t ccggt	Topo I	34	tctat CTCTa agaTA Gagat	Topo I	55
aaaaa a cagct ttttt T Gtcga	high A-T, Topo I	34	cagtg aaaa gtcac ttttt	high A-T	55
ggtgc tg cgccg ccacg ac gcggc	high Y/R	34	tCTCt c tatct agaga g aTAGa	high Y, Topo I	55
ttttt gt GTTtg aaaaa ca caaac	high A-T, Topo I	34	aacaa c AATtg TTGTT G tTAAc	Topo I	55
tCTTt cc gcCTC agaaa gg cggag	high Y, Topo I	34	CTCTa tCTCt gagat AGaga	high Y, Topo I	55
tcacg cc aggcc agtgc gg tccgg		34	gCATt gCATt cgTAA cgTAA	Topo I	55
aggcC AT ggtgc tccgg TA Ccag	Topo I*	34	tgtgt gtgtg acaca cacac	high Y/R	55
ggaag AT actgt ccTTC ta tgaca	high R, Topo I*	34	CTCtc tatct gagag aTAGa	high Y, Topo I	55
atttc a ggcca TAAag t ccggt	Topo I	34	taaag c AATag attTC g ttatc	Topo I	56
ccaga a aGTTa ggtct T TCAat	high R, Topo I*	34	cccac c tgtgt gggtg g acaca		56
taacc a aGTTc aTTGg T TCaag	Topo I*	34	agaac aactg tcTTG TTGac	Topo I*	56
tGTTg a cggga acaaC T Gccct	Topo I	34	GTTGA Tgcaa caact ACgtt	Topo I	56
taacc aaGTT aTTGg TTCaa	Topo I*	34	tggct a gAAtt accga t ctTAA	Topo I	56
gaagA Tactg cTTCT atgac	Topo I	34	agcga a agcgg tcgct T TCgcc	high R, Topo I*	56
gaagg tac ctaac cTTCc atg gaTTG	Topo I	34	gagtg g GTTAc CTCac c caatg	Topo I	56

SEQUENCE	CHARACTERISTICS	REFERENCE	SEQUENCE	CHARACTERISTICS	REFERENCE
gaaGA Tac tGTTg cTTCT atg acaac	Topo I*		34 agTTA g cagac tcaat c gtCTG	Topo I	56
agaag gt accta tcTTC ca tggat	high R, Topo I		34 cct g agtgt gga C TCaca	Topo I*	56
atact GT Tgacg tatga ca aCTGc	Topo I		34 aaaAT g cgcga tttTA C ggcgt	Topo I*	56
cCTCc GTTaa ggagg caatt	high R, Topo I		34 gtggg TTGAT tgcta caccc aacTA Acgat	Topo I*	56
aggGT Ttatg tccca aaTAC	Topo I		34 actac tGAtT cTAAT tgatg acTAA gatta	Topo I*	56
atttc agggc TAAag tccgg	Topo I		34 tcaag gtggT agtTC cacca	Topo I	57
caaag CTCgt gtTTC gagca	Topo I		34 ttatg aaCTC aaTAC TTGag	Topo I*	57
GTTat ttc aggcc caaTA Aag tccgg	high A-T, Topo I		34 gagcc acaaa CTCgg tgttt	Topo I	58
gggaC TTC ccaga ccCTG aag ggtct	high Y, Topo I*		34 ccgtg tggcc ggcac accgg	high Y/R	58
tcaga gTTTa agtCT Ccaat	high R, Topo I		34 gagcc acaaa CTCgg tgttt	Topo I	59
ggggT aaactg cccca TTGac	Topo I*		34 tgcgt gtggc acgca caccg	high Y/R	59
aaGTT cC TCttt TTCaa gg agaaa	high Y, Topo I*		34 tgccc gcCTT acggg cggaa	Topo I	60
atagg cc aGATg tatcc gg tcTAC	Topo I		34 cctgt aGTCC ggaca tcagg	Topo I	60
cCTCt t tcaga ggaga a agtct	high Y, Topo I		34 ttcag t gccaa aagtc a cggtt		61
agggg t GTTgg tcccc a caacc	high R, Topo I		34 tgtAA T CTCag acatT A GagtC	Topo I*	61
CTCtt tcaga gagaa agtct	high Y, Topo I		34 ttCAT ccta acagg aagTA GgaT Tgtcc	Topo I*	62
taaca ggagg ATTGt cCTCc	high R, Topo I		34 acagg ccta gaCTC tgtcc ggat CTGag	Topo I*	62
gCTTt gcaaa cgaaa cgttt	Topo I		35 acagg cc tagac tgtcc gg atCTG	Topo I	62
aagga AATtg TTCct tTAAC	high R and A-T, Topo I		35 agaCT Cc agggc tCTGA gg tcccc	Topo I*	62
attgc cC ATttg TAAcg gg TAAac	Topo I*		35 CTCC T TGTTC gagg a acaag	high Y, Topo I	62
aagaa cc agctg TTCTT Gg tcgac	Topo I		35 gccc C Cctgc cgggg a ggacg	high Y, Topo I	62
gtaaa AT atgct catTT ta TACga	high A-T, Topo I*		35 aagcc caagt TTCgg gTTCa	Topo I	63
gtgtG AT CTTgg cacac TA Gaacc	Topo I*		35 aagct gctgccccAA Tagcc TTCga cgacggggtt atcgg	Topo I	63
tCTTt ccgcc agaaa ggcgg	high R, Topo I		35 cAATt gctgcc ATact gtTAA cgacggg tatga	Topo I	63
gGTCT tgaaa ccaga acttt	Topo I		35 agagG TTAag TCatc tCTCc aaTTC agTAG	Topo I*	63
tgccA ATtcc acggt TAAgg	Topo I*		35 CTCaC TTAag ggcag gagtg aaTTC ccgtc	Topo I*	63
tttcc g cCTCa aaagg c ggaqt	high Y, Topo I		35 ggtgg agc tgacc ccacC TCg aCTGg	Topo I	63
aaagg G ATatt tTTCC c taTAA	high R, Topo I		35 gacac agc aGATc CTGtg tcg tcTAG	Topo I	63

ggcCT Ccaaa ccgga ggttt	Topo I	36	gtgtg agc ctggt cacac TCg gacca	Topo I	63
acaga gagga tgtCT Ctcct	high R, Topo I	36	agaga agc tagga tCTCT TCg atcct	high R, Topo I	63
caaaa gccta gttTT Cggat	Topo I	36	ccaga gA ATTcg ggtCT Ct TAAgc	high R, Topo I*	63
aactg ggcgg TTGac ccgcc	Topo I	36	CTCag ga gctgg gagtc CT Cgacc	Topo I*	63
agGTC ttgaa tccag aactt	Topo I*	36	aaggt aacgg TTCca TTGcc	Topo I*	64
tccaa aaaag aggtt ttTTC	high A-T, Topo I	36	ggccC TTAag ccggg aATTc	Topo I	64
agaaa gCTTc tctTT Cgaag	high R, Topo I	36	ttggc tGTTg aaccg acaac	Topo I	65
acCTT aCTTc tgga tgaag	Topo I*	37	gcaca caaga cgtgt gTTCt	high R/Y, Topo I	65
taaaa Ttttt atctt AAaaa	high A-T, Topo I	37	gctgg a gtaga cgacC T Catct	Topo I	66
ctact gCTTc gatga cgaag	Topo I	37	tcaga a ccctg agtct T Gggac	Topo I	66
attgg acaaa TAAcC TGttt	Topo I	37	tgaga cc cagtg aCTCt gg gtcac	Topo I	67
gctac tgc TTcga cgatg acg aagct	Topo I	37	cCTTg cc aaagt ggaac gg tTTCa	Topo I	67
tcGAT tgc TTagt agcTA Acg aatca	Topo I*	37	tggtta tct ggagg accaT AGa cCTCc	high R, Topo I	67
attgc ta ctgct TAAcg at gacga	Topo I	37	cAATt tct attaa gtTAA aga TAAtt	high A-T, Topo I*	67
tgCTT ta gAATg acgaa at ctTAC	Topo I*	37	gtaac tGATg cATTG acTAC	Topo I	68
tggGT TG ATTgc accga ac TAAcg	Topo I*	37	aaaag GATTc ttTTC cTAAg	high R, Topo I	68
gctac tg ctact cgatg ac gatga		37	gtaGA Ta ggggg catct at ccccc	high R, Topo I*	69
gcctg g ggact cgga c cCTGA	high R, Topo I	38	atccc ag caCTT TAGgg tc gtgaa	Topo I	70
gtatà g cagtg catat c gtcac		38	agaaa aG TCtgc tcttT TC agacy	high R, Topo I	70
tgaag gggaa acTTC cccct	high R, Topo I	38	agggg aAAAT tccct tttta	high R and A-T, Topo I	70
aacac agcag TTGtg tcgtc	high Y/R, Topo I	38	tCATg gcaag agTAC cgTTC	Topo I	70
ttaaa ctact aatTT Gatga	high A-T, Topo I	38	gtgtg tgtgt cacac acaca	high R/Y	71
aggCT Cccca tccga ggggt	high Y, Topo I	38	gtGTT tgtgt cacaa acaca	Topo I*	71
GATGA Tgagg cTACT ACtcc	Topo I	38	aGTC CTT CTCTc tcag gaa gagag	high Y, Topo I*	71
gAATg cAATt ctTAC gtTAA	Topo I	38	tCTCt CTT tCTCt agaga gaa agaga	high Y, Topo I*	71
ttaaa a AAATg aattt t ttTAC	high R, Topo I	38	cca gggct ggt cccga		71
tCTCc a gcagc agagg t cgtcg	high Y, Topo I	38	gctgg gCATa cgacc cgtat	Topo I	71
tggtG TTtg gACTT acacc aaac CTGaa	Topo I*	38	CTTtt CTTta gaaaa gaaat	high Y, Topo I	71

SEQUENCE	CHARACTERISTICS	REFERENCE	SEQUENCE	CHARACTERISTICS	REFERENCE
attgC Tttg CTTct TAAcg aaac gaaga	Topo I*		38 tCTTt attag agaaa TAAtc	high Y and A-T, Topo I*	71
cCTTa AATcc ggaaat tTAGg	high A-T, Topo I		39 ttggC TTTta tgcca aacgg aaaaT ACggt	high A-T, Topo I*	71
cAATg gcctg gtTAC cggac	Topo I		39 AATat atttt aCATA ttata TAAaa tgstat	high A-T, Topo I*	71
tccctg g tGTTg aggac c acaac	Topo I		39 tG TCatagAATAA TtCTT ac agtatcttatT AAgaa	high A-T and Y, Topo I*	71
tGATt g ctact actAA c gatga	Topo I		39 ttttt tCATagAATAA Ttctg aaaaa agtatcttatT AAgac	high A-T, Topo I*	71
cctgg tGTTg ggacc acaac	Topo I		39 gtgtg GA Tgtga cacac cT ACact	Topo I*	71
agCTC taagg tcgag aTTCc	Topo I*		39 attgt ga gtGTT TAAca CT Cacia	Topo I*	71
tccctg gt GTTga aggac ca caact	Topo I		39 ccAAT ttcta ggtTA Aagat	high A-T, Topo I*	71
aacAA Tg gcctg TTGtt AC cggac	Topo I*		39 taaag GTTcc atTTC caagg	high R, Topo I	71
gagtg tg caaag CTCac ac gtTTC	high R/Y, Topo I		39 actac taa actgg tgatg att TGacc	Topo I*	71
ctAAT tG Tttgt gatTA Ac aaaaa	Topo I*		39 ctgcc tAA Taaaa gacgg att atttt	high A-T, Topo I	71
tctgc t aactg agacg a TTGac	Topo I*		39 ttttc tgaac aaaag acTTG	high Y, Topo I	71
attGT T tgtgt TAAca a acaca	Topo I*		39 ctgcc tAATA gacgg attat	high A-T, Topo I	71
actgC ATatg tgacg taTAC	high V/R, Topo I		39 tctat taaag agaTA AtTTC	high A-T, Topo I	71
ttGTT tgtgt aacaa acaca	Topo I*		39 aactg gggga TTGac cccct	high R, Topo I	71
tgCAT a tgCTT acgta T ACgaa	high V/R, Topo I*		39 agcct aggca tcgga tccgt		72
GATga a gCATg cTACT T CgTAC	Topo I		39 ttaag ggcca aaTTC ccggt	high R, Topo I	72
AATtg gacaa tTAac CTGtt	Topo I*		39 gGTTC tag cccct ccaag atc gggga	high Y, Topo I	73
gtgta tAATg cacat attAC	high R/Y, Topo I		39 ccagc taC TCggg ggtcg atg agccc	Topo I	73
aaaaag CTTt gcaaa ttTTC gaaa cgttt	Topo I*		40 gccta ggc aaCAT cggat ccg TTGta	Topo I*	73
ctacc cccag ggtgg gatgg gggtc ccacc			41 ttaag ggc cagcg aaTTC ccg gtgcc	Topo I	73
tccca cccag actag agggt gggTC TGatc	Topo I*		41 agctg gggct tcgac cccga		73
caAAT at CTTaa gttta TA Gaatt	high A-T, Topo I*		42 atagt gaaca tatca cTTGt	Topo I	73
ataaa AT atacg tattt ta tatgc	high A-T and Y/R, Topo I*		42 CTCac aaCAT gagtg TTGTA	Topo I*	73
tCATA ta ctGAT agtat at gacTA	high A-T and Y/R, Topo I		42 ttctg tagct aagac atcga		73
atttt ta tccca TAAaa aT AGggt	high A-T, Topo I*		42 ttaag ggcca aaTTC ccggt	high R, Topo I	74
tctAA T attaa agatt a TAAtt	high A-T, Topo I*		42 CTCat aa agCTT gagta tT TCgaa	high A-T, Topo I*	75
attcG T Ttacc TAAgc a aaTAG	high A-T, Topo I		42 aaaaag AA Ttata ttTTC tT AAtat	high A-T, Topo I*	75

ggATc CTT tggct ccTAG gaa accga	high Y, Topo I*	42 ctGTT ccacc gacaa ggtgg	Topo I*	76
CTTtt CTT CATta gaaaa gaa gTAAt	high Y, Topo I*	42 ggacc a cctgg cCTGg t ggacc	Topo I	77
cagtg act Ctggc gtcaC TGA gaccg	Topo I	43 tGTTA A TCTTg acaat T AGaac	high A-T, Topo I*	77
ctgac act GATgc gaCTG tga cTACg	Topo I	43 acaac ATc cagca tgTTG TAG gtcgt	Topo I*	78
aaCAT ttG TCcag TTGTA Aac aggtc	Topo I	43 atCAT atC Tttat TAGta TAG aaata	high A-T, Topo I*	78
caaCA Ttg ccaac gTTGT AAC ggTTG	Topo I*	43 tctga CTCag agaCT Gagtc	Topo I	78
GTTt t aaaca caaa a tTTGt	high A-T, Topo I	43 agtaa aaccc tcatt TTGgg	high A-T, Topo I*	78
tttta t tcAAT aaaaT A Agtta	high A-T, Topo I	43 cgCTT gccgg gcgaa cggcc	Topo I*	79
caaC TTTtcac tgcct gTTG aaaagtg acgga	high Y, Topo I*	43 GTcTt accgg cagaa tggcc	Topo I	79
ctgCA TTTtcac accca gacgT AAAagtg tgggt	Topo I*	43 ccccg ac GTTgg ggggC TG caacc	Topo I	79
acagt aacc actgc tgtca TTGg tgacg	Topo I*	43 cgcct ac AATcc gcgga tg tTAGg	Topo I	79
acaca aacc tagac tgtgt TTGg atCTG	Topo I*	43 ttGTC ttatc aacag aaTAG	Topo I*	79
tttCT Ta gcagt aaaaa at cgtca	high Y, Topo I*	43 ataag gcgca taTTC cgcgt	Topo I	79
CTTtc ta gAATt gaaag at ctTAA	high Y, Topo I	43 gcCTT tcCATtggc TTcta cggaa aggTAAccg aagat	high Y, Topo I*	80
gacta t accAA CTGat a tgggt	Topo I	43 tacaG TTCaccagc caggg atgtc aagtggctg gtccc	Topo I*	80
cctat t gggac ggaTA A ccCTG	Topo I	43 aaaaa aaa agGTT ttttt ttt TCcaa	high R, Topo I*	81
acAAT gAATt tgtTA CtTAA	Topo I*	44 agccc aaa CATtg tcggg tTT GTAAc	Topo I*	81
atagg ctgcc tatcc gacgg		44		

Appendix. Sequences in the vicinity of non - homologous recombination sites and the most characteristic motifs found in them. Spaces interrupting the sequences indicate the cross-over sites (or cross-over points). If there is only one space, it represents the phosphodiester bond that was broken during non-homologous cross-over (cross-over point). Blocks of nucleotides situated between two spaces indicate the limits of uncertainty (cross-over site) in the position of the cross-over point. Trinucleotides that are preferential topo I cleavage sites *in vitro* (rat liver) are written in capital letters. In every case, the polarity of upper strand is 5' to 3'. Abbreviations: 'high' indicates five or more contiguous bases. Y and R stand for pyrimidines and purines respectively. Topo I indicates at least one topo I site within positions -5 to r + 5, where r is number of bases in the region of uncertainty. Topo I* indicates at least one topo I site having its 3' most base within positions -1 to r. Y/R and R/Y represent alternating purines and pyrimidines. A-T indicates region consisting solely of adenines and thymines.

REFERENCES

1. Franklin, N.C., (1967), Genetics, 55, 699 - 707.
2. Bourgaux, P., Sylla, B.S. and Chartrand, P., (1982), Virology, 122,84-97.
3. Hayday, A., Ruley, H.E. and Fried, M., (1982), J. of Virology, pp. 67-77.

4. Ruley, H.E. and Fried, M., (1983), *Nature*, 304, 181-184.
5. Bullock, P., Forrester, W. and Botchan, M., (1984), *J. Mol. Biol.*, 174, 55-84.
6. Hasson, J.F., Mougneau, E., Cuzin, F. and Yaniv, M., (1984), *J. Mol. Biol.*, 177, 53-68.
7. Botchan, M., Stringer, J., Mitchison, T. and Sambrook, J., (1980), *Cell*, 20, 143-152.
8. Stringer, J.R., (1981), *J. of Virology*, 38, 671-679.
9. Stringer, J.R., (1982), *Nature*, 296, 363-366.
10. Sambrook, J., Greene, R., Stringer, J., Mitchison, T., Hu, S.-L., and Botchan, M., (1979), *Cold Spring Harbor Symp. Quant. Biol.*, 44, 569-584.
11. Koshy, R., Koch, S., Freytag von Loringhoven, A., Kahmann, R., Murray, K. and Hofschneider, P.H., (1983), *Cell*, 34, 215-223.
12. Dejean, A., Bougueleret, L., Grzeschik, K.H. and Tiollais, P., (1986), *Nature*, 322, 70-72.
13. Visser, L., Reemst, A.C.M.B., van Mansfeld, A.D.M. and Rozijn, T.H., (1982), *Nucl. Acids Res.*, 10, 2189-2198.
14. Westin, G., Visser, L., Zabielski, J., van Mansfeld, A.D.M., Pettersson, U. and Rozijn, T.H., (1982), *Gene*, 17, 263-270.
15. Stabel, S. and Doerfler, W., (1982), *Nucl. Acids Res.*, 10, No. 24, 8007-8023.
16. Deuring, R., Winterhoff, U., Tamanoi, F., Stabel, S. and Doerfler, W., (1981), *Nature*, 293, 81-84.
17. Gahlmann, R. and Doerfler, W., (1983), *Nucl. Acids Res.*, 1, No. 21, 7347-7361.
18. Doerfler, W., Gahlmann, R., Stabel, S., Deuring, R., Lichtenberg, U., Schulz, M., Eick, D. and Leisten, R., (1983), *Current Topics in Microbiology and Immunology*, 109, 193-228.
19. Sylla, B.S., Huberdeau, D., Bourgaux-Ramoisy, D. and Bourgaux, P., (1984), *Cell*, 37, 661-667.
20. Maruyama, K. and Oda, K., (1984), *J. of Virology*, 49, 479-489.
21. Bullock, P., (1984), Ph.D. Thesis, Univ. of California, Berkeley.
22. Smolar, N. and Griffin, B.E., (1981), *J. of Virology*, 38, 958-967.
23. Hattori, J., Carmichael, G.G. and Benjamin, T.L., (1979), *Cell*, 16, 505-513.
24. Soeda, E. and Griffin, B.E., (1978), *Nature*, 276, 294-298.
25. Carmichael, G.G. and Benjamin, T.L., (1980), *J. of Biol. Chem.*, 255, 230-235.
26. Bendig, M.M., Thomas, T. and Folk, W.R., (1980), *J. of Virology*, 33, 1215-1220.
27. Bendig, M.M., Thomas, T. and Folk, W.R., (1980), *Cell*, 20, 401-409.
28. Ding, D., Jones, M.D., Leigh-Brown, A. and Griffin, B.E., (1982), *EMBO J.*, 1, 461-466.
29. Davies, B., Passananti, C., and Fried, M. (1987) Unpublished.
30. Norbury, C. and Fried, M. (1987) Unpublished.
31. Van Heuverswyn, H. and Fiers, W., (1979), *Eur. J. Biochem.*, 100, 51-60.
32. Wakamiya, T., McCutchan, T., Rosenberg, M. and Singer, M., (1979), *J. of Biol. Chem.*, 254, 3584-3591.
33. McCutchan, T., Singer, M. and Rosenberg, M., (1979), *J. of Biol. Chem.*, 254, 3592-3597.
34. Johnson, A.D., Barkan, A. and Mertz, J.E., (1982), *Virology*, 123, 464-469.
35. Sheflin, L., Celeste, A. and Woodworth-Gutai, M., (1983), *J. of Biol. Chem.*, 258, 14315-14321.
36. Woodworth-Gutai, M. and Nathans, D., (1978), *J. Mol. Biol.* 126, 275-288.
37. Self, I., Khoury, G. and Dhar, R., (1980), *Nucl. Acids Res.*, 8, 2225 - 2240.
38. Woodworth-Gutai, M., (1981), *Virology*, 109, 344-352.

39. Thimmappaya, B. and Shenk, T., (1979), *J. of Virology*, 30, 668 - 673.
40. Dhruva, B.R., Shenk, T. and Subramanian, K.N., (1980), *Proc. Natl. Acad. Sci. USA*, 77, 4514-4518.
41. Jeang, K.T. and Hayward, S.D., (1983), *J. of Virology*, 48, 135-148.
42. Pickup, D.J., Ink, B.S., Parsons, B.L., Hu, W. and Joklik, W.K., (1984), *Proc. Natl. Acad. Sci. USA*, 81, 6817-6821.
43. Hogan, A. and Faust, E.A., (1986), *Mol. Cell. Biol.*, 6, 3005-3009.
44. Zain, B.S. and Roberts, R.J., (1978), *J. Mol. Biol.*, 120, 13-31.
45. Ling, L.E., Manos, M.M. and Gluzman, Y., (1982), *Nucl. Acids Res.*, 10, 8099-8112.
46. Westphal, H., (1981), *J. of Virology*, 40, 526-532.
47. Ahmed, C.M.I., Lee, E. and Zain, B.S., (1984), *Virology*, 133, 233-237.
48. Roth, D.B., Porter, T.N. and Wilson, J.H., (1985), *Mol. Cell. Biol.*, 5, 2599-2607.
49. Winocour, E., Chittlaru, T., Tsutsui, K., Ben-Levy, R. and Shaul, Y., (1987) *Cancer Cells*, Cold Spring Harbor Press, 509-516.
50. Anderson, R.A., Kato, S. and Camerini-Otero, R.D., (1984), *Proc. Natl. Acad. Sci. USA*, 81, 206-210.
51. Wilson, J.H., Berget, P.B. and Pipas, J.M., (1982), *Mol. Cell. Biol.*, 2, 1258-1269.
52. Chang, X.B. and Wilson, J.H., (1986), *J. of Virology*, 58, 393-401.
53. Woodworth-Gutai, M., (1981), *Virology*, 109, 353-365.
54. DiMaio, D. and Nathans, D., (1980), *J. Mol. Biol.*, 140, 129-142.
55. Stringer, J.R., (1985), *Mol. Cell. Biol.*, 5, 1247-1259.
56. Subramani, S. and Berg, P., (1983), *Mol. Cell. Biol.*, 3, 1040-1052.
57. Parnes, J.R., Sizer, K.C., Seidman, J.G., Stallings, V. and Hyman, R., (1986), *EMBO J.*, 5, 103-111.
58. Chu, M.L., Gargiulo, V., Williams, C.J. and Ramirez, F., (1985), *J. Biol. Chem.*, 260, 691-694.
59. Barsh, G.S., Roush, C.L., Bonadio, J., Byers, P.H. and Gelinas, R.E., (1985), *Proc. Natl. Acad. Sci. USA*, 82, 2870-2874.
60. Lehrman, M.A., Schneider, W.J., Sudhof, T.C., Brown, M.S., Goldstein, J.L. and Russell, D.W., (1984), *Science*, 227, 140-146.
61. Lehrman, M.A., Russell, D.W., Goldstein, J.L. and Brown, M.S., (1986), *Proc. Natl. Acad. Sci. USA*, 83, 3679-3683.
62. Nalbantoglu, J., Phear, G. and Meuth, M., (1987), *Mol. Cell. Biol.*, 7, 1445-1449.
63. Nalbantoglu, J., Hartley, D., Phear, G. and Meuth, M., (1986), *EMBO J.*, 5, 1199-1204.
64. Rovinski, B., Munroe, D., Peacock, J., Mowat, M., Bernstein, A. and Benchimol, S., (1987), *Mol. Cell. Biol.*, 7, 847-853.
65. Molineaux, S.M., Engh, H., De Ferra, F., Hudson, L. and Lazzarini, R.A., (1986), *Proc. Natl. Acad. Sci. USA*, 83, 7542-7546.
66. Mager, D.L., Henthorn, P.S. and Smithies, O., (1985), *Nucl. Acids Res.*, 13, 6559-6575.
67. Jennings, M.W., Jones, R.W., Wood, W.G. and Weatherall, D.J., (1985), *Nucl. Acids Res.*, 13, 2897-2906.
68. Spritz, R.A. and Orkin, S.H., (1982), *Nucl. Acids Res.*, 10, 8025-8029.
69. Jagadeeswaran, P., Tuan, D., Forget, B.G. and Weissman, S.M., (1982), *Nature*, 296, 469-470.
70. Vanin, E.F., Henthorn, P.S., Kioussis, D., Grosveid, F. and Smithies, O., (1983), *Cell*, 35, 701-709.
71. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.R., Forget, B.G., Weissman, S.M., Slightom, J.L., Blech, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J., (1980), *Cell*, 21, 653-668.

72. Nicholls, R.D., Higgs, D.R., Clegg, J.B. and Weatherall, D.J., (1985), *Blood*, 65, 1434-1438.
73. Nicholls, R.D., Fischel-Ghodsian, N. and Higgs, D.R., (1987), *Cell*, 49, 369-378.
74. Orkin, S.H. and Michelson, A., (1980), *Nature*, 286, 538-540.
75. Gilliam, A.C., Shen, A., Richards, J.E., Blattner, F.R., Mushinski, J.F. and Tucker, P.W., (1984), *Proc. Natl. Acad. Sci. USA*, 81, 4164-4168.
76. Corcoran, L.M., Cory, S. and Adams, J.M., (1985), *Cell*, 40, 71-79.
77. Nalbantoglu, J. and Meuth, M., (1986), *Nucl. Acids Res.*, 14, 8361-8371.
78. Pasananti, C. Davies, B., Ford, M. and Fried, M., (1987), *EMBO J.*, 6, 1697-1703.
79. Wilkie, T.M. and Palmiter, R.D., (1987), *Mol. Cell. Biol.*, 7, 1646-1655.
80. Maeda, N., Yang, F., Barnett, D.R., Bowman, B.H. and Smithies, O., (1984), *Nature*, 309, 131-135.
81. Dean, M., Park, M. and Vande Woude, G.F., (1987), *Mol. Cell. Biol.*, 7, 921-924.
82. Savageau, M.A., Metter, R., and Brockman, W.W., (1983), *Nucl. Acids Res.* 11, 6559 -6570.
83. Been, M.D., Burgess, R.R., and Champoux, J.J., (1984), *Nucl. Acids Res.* 12, 3097.
84. Edwards, K.A., Halligan, B.D., Davis, J.L., Nivera, N.L., and Liu, L.F., (1982), *Nucl. Acids Res.*, 10, 2565.
85. Bullock, P., Champoux, J.J., and Botchan, M., (1985), *Science*, 230, 954-958.
86. Hope, T.J., Aguilera, R.J., Minie, M.E., and Sakano, H., (1986), *Science*, 231, 1141 - 1146.
87. Grandgentt, D.P., Vora, A.C. and Schiff, R.D., (1978), *J. of Virology*, 89, 119-132.
88. Kopchick, J.J., Harless, J., Geisser, B.S., Killam, R., Hewitt, R.R., and Arlinghaus, R.B., (1981), *J. Virology*, 37, 274-283.
89. Schwartzberg, P., Colicelli, J., and Goff, S.P., (1984), *Cell*, 37, 1043-1052.
90. Wake, C.T., Gudewicz, T., Porter, T., White, W., and Wilson, J.H., (1984), *Mol. Cell. Biol.*, 4, 387-398.
91. Konopka, A.K., Reiter, J., Jung, M., Zarling, D.A., and Jovin, T.M., (1985), *Nucl. Acids Res.*, 13, 1683 - 1701.
92. Nordheim, A. and Rich, A., (1983), *Nature*, 303, 674 - 679.
93. Bullock, P., Miller, J., and Botchan, M., (1986), *Mol. Cell Biol.* 6, 3948 - 3953.
94. Trifonov, E.N., Konopka, A.K., and Jovin, T.M., (1985), *FEBS Lett.*, 185, 197 - 202.
95. Arnott, S., Chandrasekarn, R., Hall, I.H., and Puigjaner, L.C., (1983), *Nucl. Acids Res.*, 11, 4141 - 4155.
96. Carbon, J., (1984), *Cell*, 37, 351 - 353.
97. Htun, H., and Dahlberg, J.E., (1987), *V-th Conv. Bimol. Stereodynam.*, 080.
98. Moser, H.E., and Dervan, P.B., (1987), *Science*, 238, 645 - 650.
99. Schon, E., Evans, T., Welsh, J., and Efstratiadis, A., (1983), *Cell*, 35, 837 - 848.
100. Hoffman-Liebermann, B., Liebermann, D., Troutt, A., Kedes, L.E., and Cohen, S., (1986), *Mol. and Cell. Biol.*, 6, 3632 - 3642.
101. Konopka, A.K., Smythers, G.W., Owens, J., and Maizel, J.V., (1987), *Gene Anal. Techn.*, 4, 63 - 74.
102. Jeffreys, A.J., Wilson, V. and Thein, S.L., (1985), *Nature*, 314, 67 - 70.
103. Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., White, R., (1987), *Science*, 235, 1616 - 1622.