Targeting of Human Retrotransposon Integration Is Directed by the Specificity of the L1 Endonuclease for Regions of Unusual DNA Structure†

Gregory J. Cost and Jef D. Boeke*

Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, 725 North Wolfe Street, Baltimore, Maryland 21205

Received August 3, 1998; Revised Manuscript Received October 30, 1998

ABSTRACT: L1 elements are polyA retrotransposons which inhabit the human genome. Recent work has defined an endonuclease (L1 EN) encoded by the L1 element required for retrotransposition. We report the sequence specificity of this nicking endonuclease and the physical basis of its DNA recognition. L1 endonuclease is specific for the unusual DNA structural features found at the TpA junction of 5′(dT_n-dA_n)·5′(dT_n-dA_n) tracts. Within the context of this sequence, substitutions which generate a pyrimidine−purine junction are tolerated, whereas purine−pyrimidine junctions greatly reduce or eliminate nicking activity. The A-tract conformation of the DNA substrate 5′ of the nicked site is required for L1 EN nicking. Chemical or physical unwinding of the DNA helix enhances L1 endonuclease activity, while disruption of the adenine mobility associated with TpA junctions reduces it. Akin to the protein−DNA interactions of DNase I, L1 endonuclease DNA recognition is likely mediated by minor groove interactions. Unlike several of its homologues, however, L1 EN exhibits no AP endonuclease activity. Finally, we speculate on the implications of the specificity of the L1 endonuclease for the parasitic relationship between retroelements and the human genome.

Despite diversity of organization, protein structure, and mechanism of replication, retrotransposons can be divided into two classes: those containing long terminal repeats (LTRs), and those without LTRs, but with transcripts containing polyadenylate tails (polyA, non-LTR). More than one-third of the human genome consists of retroelement-derived sequence (*1*). Nearly half of this total (15%) corresponds to L1 elements, polyA retrotransposons present in about 600 000 copies per haploid genome. The large majority of L1 elements are variably 5′ truncated; of those elements which are full length, all but 15−30 are retrotranspositionally inactive (*2*). L1 elements comprise a critical part of the human genome: these few active retroelements have the capacity to cause mutation, disease, genetic variation, and polymorphism; their inactive brethren remain ideal substrates for recombination and rearrangement (*3*−*5*). Retroelements have been found within gene regulatory regions and centromere heterochromatin and are likely to help define large-scale genomic structure (*6*−*9*). As L1 endonuclease guides the process of retrotransposition, the specificity of this enzyme has significantly influenced the placement of a large fraction of human DNA. Understanding the integration specificity of retrotransposons is essential for understanding the current structure and evolutionary history of the human genome. We describe here a major determinant of this specificity, the DNA sequence preferences of the L1 endonuclease domain.

Up to six kilobases long, L1 elements are typically flanked by a target site duplication of variable length. Full-length elements contain both 5′ and 3′ UTRs [the 5′ UTR contains an internal Pol II promoter (*10*)], and two nonoverlapping open reading frames (Figure 1A). ORF1 has been shown to code for an RNA binding protein specific for L1 RNA (*11*) and to form ribonucleoprotein particles with L1 RNA in vivo (*12*). ORF2 encodes endonuclease and reverse transcriptase activities required for retrotransposition (*13*−*15*). The current model of retrotransposition by polyA retrotransposons is based primarily on the biochemical work of Luan and Eickbush with the R2Bm element (Figure 1B) (*16*). In this model, an element-encoded endonuclease nicks the target DNA, generating an exposed 3′-hydroxyl which serves as a primer for reverse transcription of the element's RNA. The mechanism of second strand synthesis and nick repair is unknown. The proteins encoded by L1 elements are thought to work only in cis, that is, only on the RNA from which they were translated (*17*). Many other human retrotransposable sequences are believed to plunder L1 proteins in order to proliferate (*1*, *17*).

Previous work has identified a nicking endonuclease activity encoded in the first ∼28 kDa of the L1.2A ORF2 protein (L1 EN) (*13*). L1 endonuclease belongs to a large family of endonucleases sharing tertiary structure and conserved catalytic residues. This family can be subdivided teleologically into three classes: enzymes specific for repair of lesions in DNA, nucleases involved in retroelement replication, and digestive enzymes. Members of this family therefore include enzymes of diverse function and specificity: bovine pancreatic DNase I [a relatively nonspecific endonuclease (*18*)], ExoIII [the major apurinic/apyrimidinic (AP) endonuclease activity in *E. coli* (*19*)], human AP endonuclease, *D. melanogaster* Rrp1 [an enzyme with exonuclease, AP endonuclease, and strand transfer activities

† This work was funded in part by NIH Training Grant 5T32CA09139-24 (G.J.C.), and NIH Grant CA16519 (J.D.B.).

* Corresponding author. Telephone: 1-410-955-0398. Fax: 1-410-614-2987. Electronic Mail: jboeke@jhmi.edu.

10.1021/bi981858s CCC: $15.00
Published on Web 12/03/1998

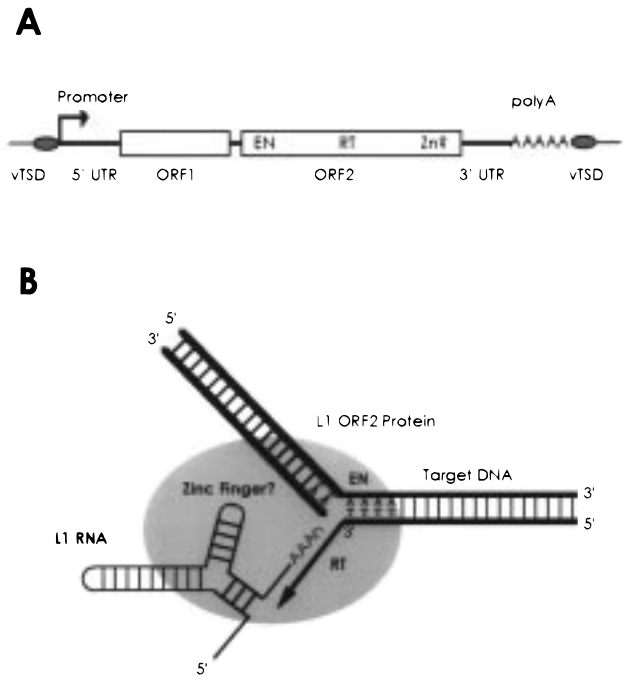

FIGURE 1: Overview of L1 element structure and model of retrotransposition. (A) Components and structure of the human L1 element. EN, endonuclease domain; RT, reverse transcriptase domain; Zn?, putative zinc finger domain; vTSD, variable target site duplication. Adapted from Feng et al. (B) Proposed mechanism of L1 retrotransposition based on the model of Luan and Eickbush. The endonuclease domain nicks the target DNA, exposing a 3′OH which is used to prime reverse trasncription of element RNA. The consensus sequence for the endonuclease domain is shown in the target DNA (see text).

(*20*)], R1Bm endonuclease [specific for a 14 bp sequence in the rDNA (*21*, *22*)], and the Tx1 element of *Xenopus laevis* [specific for another transposon (*23*)] as well as a multitude of other retrotransposon endonucleases of unknown or minimal specificity. Mutation of the conserved residues in L1 endonuclease which are known to be in the active site in similar nucleases abolishes catalytic activity, and eliminates retrotransposition in cell culture. Assays used previously to monitor L1 nicking endonuclease activity relied on conversion of the supercoiled form of a plasmid into the relaxed, open circle form (*13*). While convenient as a general measure of enzyme activity, this type of analysis offers only rudimentary resolution and is sensitive to contaminating nucleases; we have therefore developed an oligonucleotide-based system for examining the phosphodiesterase activity of the L1 endonuclease domain. Oligonucleotides are ideal substrates for such an assay because they allow precise determination of target sequences, are available in high concentration, and are easily synthesized to incorporate a variety of base modifications. Among those endonucleases which have been examined, no difference in specificity has been observed between plasmid and oligo substrates, although higher values of k_{cat} as well as K_m are commonly observed with oligonucleotides (*24*).

Analysis of L1 endo nicking sites on plasmid DNA, nicking sites inferred from new L1 insertions that occurred in cell culture, and from inspection of database sequences, yielded the crude consensus of dT_n−(nick)−dA_n (*13*). An independent study of *Alu* termini inferred that Alu elements insert into the sequence dT_2-dA_4 (*25*). Here this consensus is refined experimentally; we show that L1 EN target site selection has its basis in the recognition of the unusual structural properties of these homopolymeric sequences and the junction formed between them.

MATERIALS AND METHODS

Protein Expression and Purification. The L1 endonuclease purified by Feng et al. contains 39 extra C-terminal amino acids derived from the pET15b vector. To produce a protein with a minimal amount of non-native sequence, pQF221 was digested with *Nco*I and *Blp*I, and the sequence between these two sites was replaced by the double-stranded oligonucleotide created by annealing JB1325 (5′-CATGCATCACCACCATCATCAC-3′) and JB1326 (5′-TCAGTGATGATGGTGGTGATG-3′). This construct, pGC6, encodes the first 239 amino acids of L1.2A ORF2, 2 extraneous amino acids which are artifacts of cloning (Ala-Met), followed by 6 histidine residues. This protein has an anticipated molecular mass of 28.2 kDa and a predicted p*I* of 9.69. pGC6 was transformed into *E. coli* strain BL21(DE3), grown to an A_{600} of 0.8 in 4 L of LB medium supplemented with 100 μg/mL ampicillin, and induced for 3 h with 1 mM IPTG. All subsequent steps of the purification were performed at 0−4 °C. Cells were pelleted and frozen in liquid nitrogen, then thawed on ice, resuspended in 25 mL of sonication buffer (300 mM NaCl, 25 mM HEPES, pH 7.75) plus 1 mM PMSF, and sonicated. The mixture was clarified by centrifugation and filtration through a 0.45 μm membrane, and then gently mixed with 3 mL of nickel agarose resin (Qiagen) for 1 h. The resin was applied to a column and washed with 50 mL of sonication buffer, 50 mL of wash buffer (300 mM NaCl, 25 mM HEPES, pH 7.75, 10% glycerol), 25 mL of wash buffer plus 1 M NaCl, 50 mL of wash buffer plus 1.5 M NaCl, and finally 50 mL of wash buffer plus 40 mM imidazole. The protein was eluted in 10 mL of wash buffer plus 700 mM NaCl plus 150 mM imidazole, and vacuum-dialyzed and concentrated to 2 mL against dialysis buffer (1 M NaCl, 20 mM HEPES, pH 7.0, 20 mM EDTA, 5 mM DTT, 10% glycerol). L1 endonuclease was then gel filtered through Sephadex G-50 SuperFine resin in L1 endonuclease dialysis buffer at a rate of 12.5 min/mL. Fractions containing only L1 endonuclease (as determined by SDS−PAGE/silver staining) were pooled and concentrated by ultrafiltration (Amicon). L1 endonuclease containing an active site mutation (D205G, pGC15) was cloned, expressed, and purified identically.

Oligonucleotide and Topoisomer Substrates. In addition to the sequence shown, all oligo substrates herein contain six bases of 5′- and 3′-terminal sequence designed to facilitate annealing in the correct register and to prevent hairpin formation; the structure is therefore 5′-GCCCGG-N_n-GGCCCG-3′. For simplicity, the GC-rich terminal sequences are not displayed when describing the substrates. Oligonucleotides (Operon) were first purified as single strands by denaturing PAGE, visualized by UV shadowing, mash-eluted, and precipitated. Oligos were end-labeled with T4 kinase and [γ-^{32}P]ATP, the reaction mix was heat-inactivated, and the complementary strand was annealed by incubation at 90 °C with gradual cooling to room temperature. Annealed oligos were further purified by nondenaturing PAGE, mash-eluted, and ethanol-precipitated. In all experiments, only one strand is labeled. Oligos which underwent methylene blue

visualization during preannealing purification rather than UV shadowing yielded identical results in nicking assays, demonstrating that thymine dimer formation either is not appreciable or does not affect the nicking ability or specificity of L1 endonuclease.

Oligos with apurinic or apyrimidinic sites were generated by using uracil *N*-glycosylase (UNG, Gibco-BRL) to remove a uracil residue deliberately incorporated during synthesis. UNG reactions were stopped by the addition of a 3-fold molar excess of uracil *N*-glycosylase inhibitor, and then used directly in the nicking reaction. Oligos without uracil incorporated were treated identically. Generation of the correct AP site was demonstrated by specific cleavage by 1 unit of Exo III for 1 h in a buffer which represses most of its exonuclease activity (66 mM Tris, pH 8.0, 5 mM $CaCl_2$) (*26*).

Methylated oligonucleotides were produced by incubation of 2 pmol of double-stranded, labeled oligonucleotide with 0−80 μL of 100% DMS in buffer containing 50 mM sodium cacodylate, 0.1 mM EDTA in a final volume of 200 μL. Methylation proceeded for 10 min, at which time the reaction was stopped by addition of 40 μL of 1 M β-mercaptoethanol, 1.5 M sodium acetate, pH 7.0. The DNA was then ethanol-precipitated twice and resuspended in 10 mM Tris, pH 7.6, 1 mM EDTA.

Topoisomer distributions were made according to the method of Bowater (*27*). Approximately 200 ng of a pBS KS-topoisomer ladder covering a wide range of linking numbers was incubated for 30 min with L1 EN at a final concentration of approximately 7 μM. The reaction was then electrophoresed at 2 V/cm on a 25 cm 1% agarose gel for 20 h, and the gel was stained in 500 μg/mL ethidium bromide for 2 h.

Nicking Assay. Ingredients of the standard oligo nicking assay are as follows: 400 nM oligonucleotide, 60 mM NaCl, 50 mM HEPES, pH 7.5, 5 mM $MgCl_2$, and 3.5 μM L1 endonuclease; these conditions were optimized using the T_6A_6 substrate. Incubation was at 37 °C for 1 h, after which the reaction was stopped by addition of an equal volume of 95% formamide, 20 mM EDTA. Samples were heated to 100 °C for 2 min, then chilled on ice, and electrophoresed through denaturing 20% polyacrylamide gels. The identity of bands was assigned based on comparison to restriction endonuclease digestions of many substrates. In some cases, restriction digests or oligos synthesized to correspond to the predicted products were mixed with the reaction and co-electrophoresed in order to negate mobility differences based on salt concentration and to conclusively identify the products of the reaction. In all cases, the effect of differential salt concentration on electrophoresis was found to be negligible. Quantitation of band intensities was accomplished using a Molecular Dynamics Storm PhosphorImager and ImageQuant version 1.11 software.

Computational Analysis. Analysis of hexamer frequencies was performed with the QuickBasic Version 4.0 program Freq (source code available upon request). For analysis of genomic DNA, approximately 62 MB of human genomic DNA was used; this corresponds to the majority of human DNA which has been sequenced to date, but it is restricted in its scope to mostly chromosome seven. Analysis of samples randomly selected from different chromosomes/isochores yielded identical results. For analysis of coding sequence L1 endo target site frequency, approximately 6×10^5 nucleotides of nonredundant human cDNA sequence were randomly selected from Genbank (release 103.0). This sample of coding sequence is representative of cDNA as a whole, as the codon usage frequencies within it are quite close to those derived from inspection of the sequenced coding genome (G.J.C. and J.D.B., data not shown) (*28*). The bias in L1 endonuclease target site distribution is displayed as the ratio between the frequency of occurrence in bulk genomic DNA (F_g) and the frequency observed in coding regions (F_c). Random hexamer distribution would therefore yield a value of 1 for these ratios. The contribution of the 3−5% coding sequence contained in the genomic sequence was neglected; inspection of purely noncoding sequence would only serve to enhance rather than reduce the magnitude of the observed bias.

RESULTS

Expression, Purification, and Substrate Requirements of the L1 Endonuclease Domain. To perform biochemical analysis of the L1 endonuclease domain, the N terminal 28 kDa of L1.2A ORF2 was expressed and purified to apparent homogeneity using metal affinity and gel filtration chromatography (Figure 2A).

Prior experiments yielded clues as to the substrate requirements of L1 endonuclease. Based on these observations, we devised a model oligonucleotide substrate: 5′(dT_6-dA_6)·5′-(dT_6-dA_6) (excluding the GC clamp described under Materials and Methods), abbreviated as T_6A_6. Strong nicking activity was detected when this substrate was incubated with wild-type L1 endonuclease (Figure 2B), but not observed with identically prepared active site missense mutant protein (data not shown). Nicking of this substrate occurs primarily at the junction of the T and A tracts and within the ApA dinucleotide adjacent to the TpA. Activity on the reverse of this idealized sequence (A_6T_6) is drastically reduced in intensity at all positions and is essentially the inverse of what is seen with T_6A_6 with respect to the pattern of specificity; nicking increases with distance from the junction (Figure 2B). Thus, L1 endonuclease is capable of nicking double-stranded oligonucleotides, DNA devoid of large-scale topological constraints and free from any long-distance cis interactions.

Single-stranded DNA is thought to be generated during retrotransposition, during reverse transcription. L1 EN could in principle exhibit activity on targets which are single stranded. To examine this possibility, L1 EN was incubated with a single-stranded oligo containing a sequence known to be nicked when present in double-stranded DNA. As no nicking activity was detected on this substrate, we conclude that L1 EN has no activity on single-stranded DNA (Figure 2C).

These results are also consistent with the activity of a 3′→5′ exonuclease incapable of breaking phosphodiester bonds between thymidine residues [such as the exonuclease activity of the related *D. melanogaster* Rrp1p (*29*)]. This pattern could derive from an exonuclease which processes up the strand, cleaving bonds in succession, or could result from the sequential action of many binding/cleaving events of an enzyme limited to 3′-terminal nuclease activity. To eliminate both possibilities, a kinetic analysis was performed; the reaction was divided in two and electrophoresed under

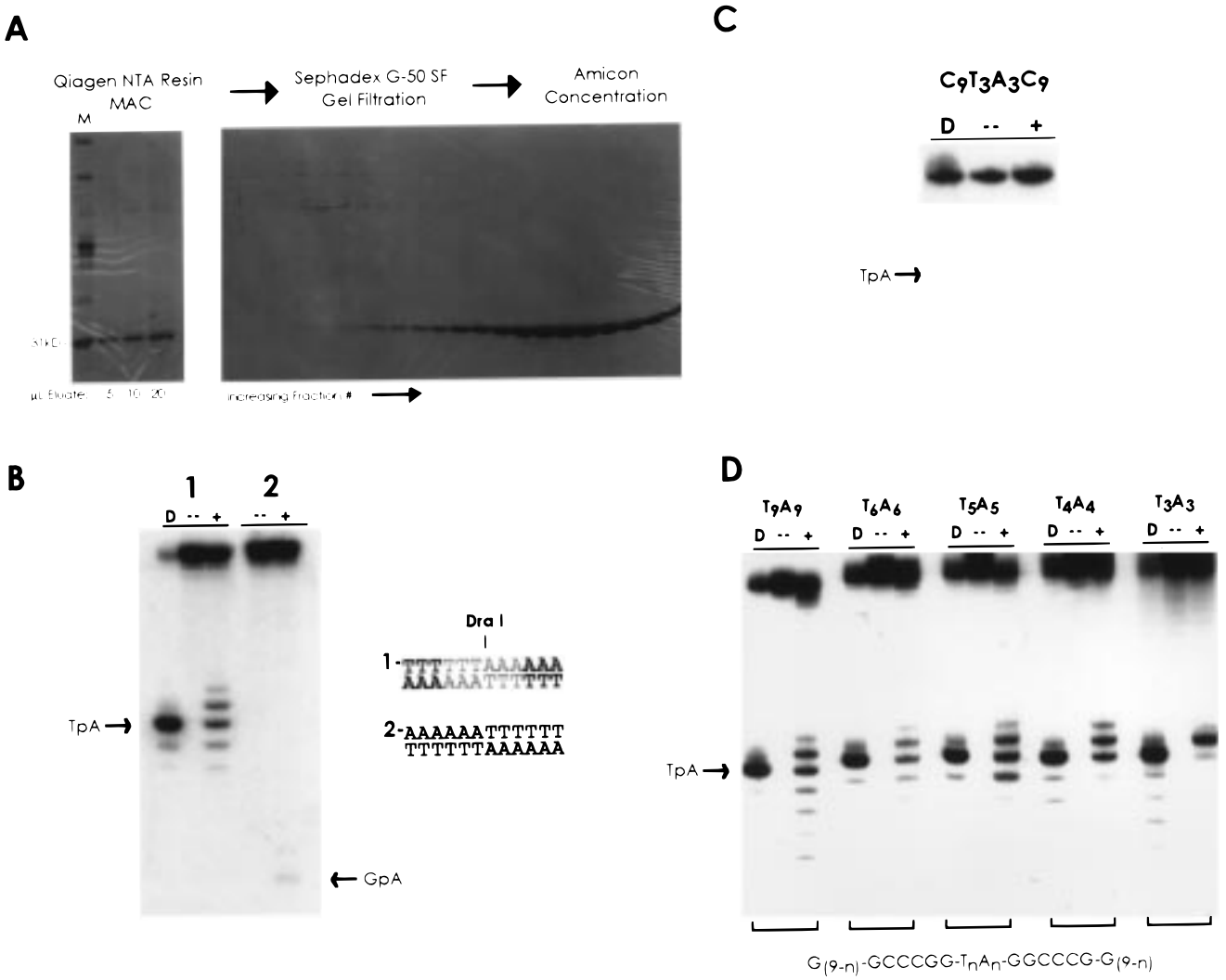

FIGURE 2: Purification, nicking activity, and minimal sequence requirements of L1 endonuclease. (A) Purification of L1 endonuclease. L1 endonuclease was purified as described under Materials and Methods and electrophoresed through 10% SDS−polyacrylamide after each step of purification. The gels were silver stained and slightly overdeveloped to emphasize any contaminants. No other bands can be seen in the lanes containing protein from the fractions which were pooled and concentrated. L1 endonuclease runs approximately 3 kDa larger than expected, a phenomenon sometimes seen with highly basic proteins. (B) L1 endonuclease can nick oligonucleotides. Oligo substrate T_6A_6 was incubated with either *Dra*I (D), buffer only (− −), or wild-type L1 endonuclease protein (+). *Dra*I cuts between the T and the A on both strands. Mutant L1 EN protein (D205G) has no activity (data not shown). (C) L1 endonuclease has no activity on single-stranded DNA. Single-stranded DNA was prepared and reacted with either *Dra*I (D), buffer alone (− −), or L1 EN (+). C_n was used to maintain the length of the oligos rather than the GC clamp described under Materials and Methods so as to avoid formation of the secondary structure seen in 2D. (D) L1 endonuclease activity on T_nA_n tracts of varying length. Oligo substrates T_9A_9, T_6A_6, T_5A_5, T_4A_4, and T_3A_3 were each incubated with either *Dra*I, buffer alone, or L1 endonuclease. Oligo length was held constant by addition of the appropriate number of G·C base pairs outside the constant six base pair 5′- and 3′-terminal sequences at both ends. Extension of the GC clamp on the ends of the substrates resulted in the formation of an increasingly stable single-stranded secondary structure which made purification (and end labeling) difficult when $n < 5$. In no other cases is this structure predicted or seen. Some contamination by aberrantly synthesized oligos can be observed, especially in the *Dra*I digests.

both denaturing and nondenaturing conditions. While the band pattern of the denaturing gel was as shown, no bands corresponding to any of the products predicted of the above two exonuclease models were observed (data not shown). The kinetics of the reaction are otherwise uninformative, with all products generated at essentially the same rate (data not shown). To rule out the possibility of an endonuclease-created nick at an A−A junction followed by limited exonuclease activity releasing a mono- or dinucleotide species undetectable by PAGE, thin-layer chromatography of nicking reactions was performed, and the separated products were analyzed. No species compatible with this hypothesis was observed (data not shown). In addition, 3′ end-labeled oligos yielded digestion patterns identical (but reversed in polarity) to those shown here (data not shown). We therefore find no evidence for exonuclease activity present in this domain.

We then sought to determine the minimal 5′(dT_n-dA_n)·5′-(dT_n-dA_n) tract capable of supporting levels of nicking activity equivalent to the T_6A_6 substrate. Oligos in which $n = 9$, 5, 4, or 3 (of constant overall length) were prepared and reacted with L1 endonuclease. Potent nicking activity was observed throughout the series (Figure 2D), declining only when $n = 3$ or 4, a reduction attributable to the encroachment by the flanking sequence on the region likely to be critical for DNA recognition by L1 endonuclease (see Figure 3B, and Discussion). Thus, the nicking activity of L1 EN is not affected by variation in the length of the T_nA_n

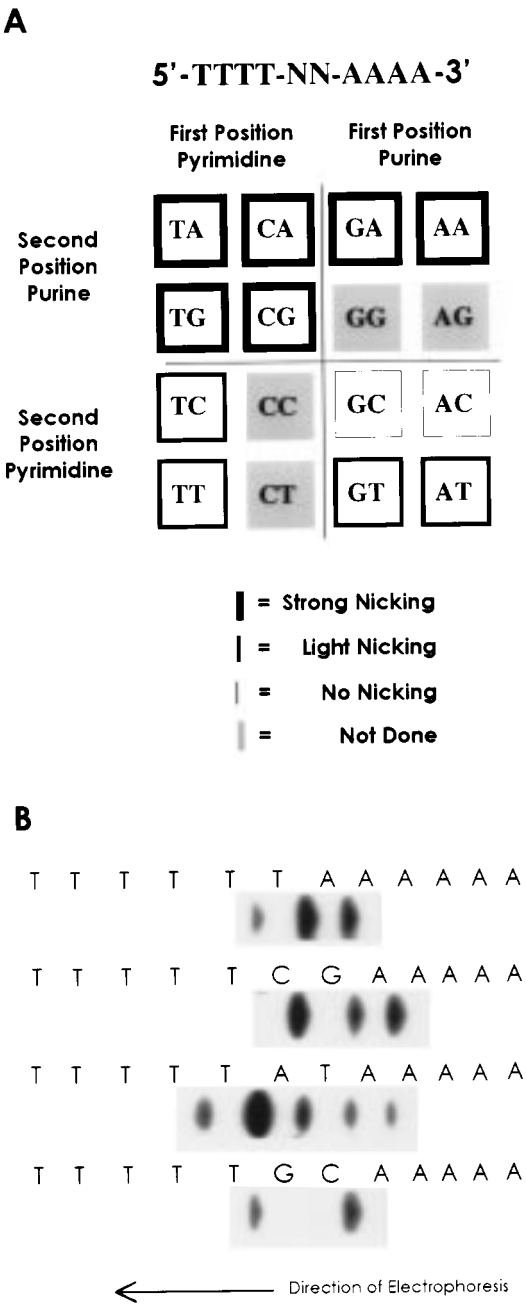

FIGURE 3: L1 endonuclease is sensitive to nucleotide composition at the junction of 5′(dT_n-dA_n)·5′(dT_n-dA_n) tracts. (A) Dinucleotide preferences. Oligonucleotides were synthesized with the general sequence $T_5N_2A_5$ where N_2 corresponds to each of the dinucleotides shown. The ability of L1 endonuclease to nick the N−N bond is indicated schematically. (B) Nicking of selected oligos from part A. The direction of electrophoresis is from right to left.

tract when $n > 4$. The T_6A_6 sequence was chosen as the basis for further investigation.

Nucleotide Sequence and DNA Structural Preferences of L1 Endonuclease. Having determined the basic features required of substrates for L1 endonuclease, we investigated more detailed sequence preferences. We asked the following question: When placed in an optimal context (replacing the TpA junction of a T_nA_n stretch), which pairs of dinucleotides allowed for nicking at the central N−N phosphodiester bond? The results, summarized in Figure 3A and shown for 4 substrates in Figure 3B, indicate that for 12 out of the 16 possible dinucleotides, nicking is generally possible within YpR dinucleotides, but not RpY dinucleotides (Y = pyrimidine; R = purine).

Scanning substitution of the T_6A_6 sequence was performed by single substitution of G·C for T·A base pairs and C·G for A·T base pairs at all possible positions. All oligos in which A·T was replaced with C·G yielded nicked products which followed the above-mentioned pattern: generation of an ApC dinucleotide severely repressed nicking at this site, and generation of a CpA dinucleotide supported nicking activity. The substitutions of G·C for T·A also gave substrates which behaved predictably with respect to dinucleotide preferences at the site of substitution, yet displayed a curious pattern (Figure 4A). While nicking adjacent to the substitution was unaffected, cleavage between the third and fourth bases after the substitution was attenuated more than 15-fold. Truncations of this sequence which move a G residue adjacent to the T tract produced similar results (Figure 2D). Such a preference implies that DNA sequence or structure at this position is important for L1 EN−DNA interaction. Further evidence of contact 5′ of the nicking site can be seen in Figure 3B. In the T_5ATA_5 substrate, L1 endonuclease is presented with two TpA junctions, one flanked by T_5 and the other by A_5, each having intact (continuous Y or R) 5′ and 3′ ends, respectively. The TpA junction having intact 5′ T_5 sequence is cleaved far more strongly than its competitor, independently illustrating the importance of favorable sequence 5′ of the nicking site.

The influence of various base substituents and substitutions on A tract bending has been examined (*30*−*32*). In general, major groove pyrimidine methyl groups are not required for bending, and minor groove amino groups (such as those of guanosine) destroy A tract structure. Most of the stability of A tracts seems to arise from base stacking interactions; purines rather than pyrimidines are therefore the dominant structural unit. In addition, any disruption of the continuity of the A tract (particularly disruption by a pyrimidine) eliminates the A tract bending. The T_nA_n sequence preferred by L1 EN is composed of two opposing A tracts, one on each strand on either side of the TpA bond. The effects seen when a G·C base pair is substituted for a T·A base pair at position −4 may be the result of similar disruption of the A tract DNA found 5′ of the nicking site. If this hypothesis is correct, then substitution of an A·T or I·C (I, inosine; lacks the minor groove amino group at position 2 which destroys A tract bending, but is otherwise identical to guanosine) base pair should have effects similar to the G·C substitution, and substitution of a C·I but not a C·G base pair would be predicted to have no effect on nicking activity. The data shown in Figure 4B confirm both of these predictions. The G·C, I·C, and A·T substitutions all show lessened nicking activity at the TpA bond and quite similar patterns of nicking overall (compare substrates 2, 3, and 4). Substitution with C·G changed the pattern of nicking activity substantially (substrate 5), but when C·I (but not I·C) was substituted for the T·A base pair at position −4, the pattern of nicking observed was identical to the unsubstituted sequence (Figure 4B, substrates 6). Thymidine methyl groups are not required for A tract bending. These methyl groups are similarly not relevant for DNA recognition and cleavage by L1 EN, as equal nicking is observed on T_6A_6 and U_6A_6 (U = uracil; T = 5-methyluracil) (Figure 4C). The activity of L1 endonuclease therefore directly parallels the ability of the sequence

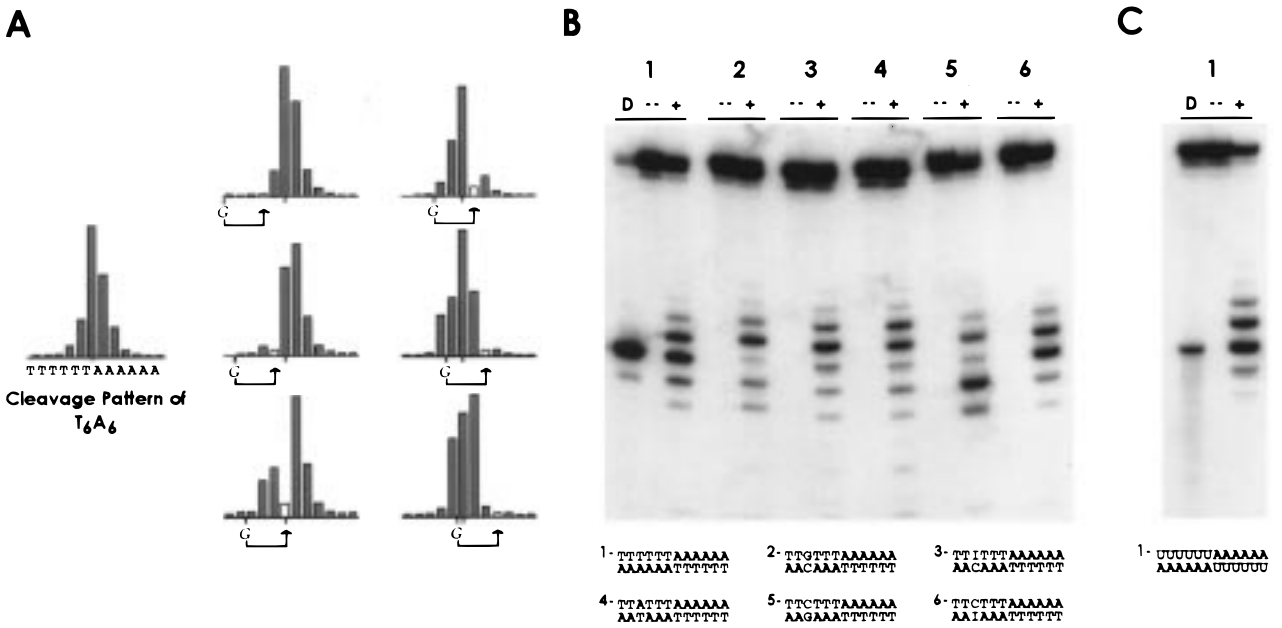

FIGURE 4: Disruption of the poly(dA) tract 5′ of the TpA junction interferes with nicking by L1 endonuclease. (A) Scanning substitution of the T_6A_6 sequence. In the general sequence T_6A_6, G·C was substituted for T·A at all six possible positions. The effect of the substitutions is manifest between the third and the fourth base 3′ of the replacement rather than at the position of the substitution itself. Nicking on the complementary strand was affected only at the position of substitution. (B) A tract structure is important for L1 EN cleavage. All possible substitutions were examined at the −4 position. Substrate 1, T_6A_6; 2, $T_2GT_3A_6$; 3, $T_2IT_3A_6$; 4, $T_2AT_3A_6$; 5, $T_2(C{\cdot}G)T_3A_6$; 6, $T_2(C{\cdot}I)T_3A_6$. D, *Dra*I; (− −), buffer only; (+) L1 endonuclease. (C) Thymidine methyl groups are not important for L1 EN cleavage. All thymidines in the T_6A_6 substrate were replaced with uracil and reacted with L1 EN.

in question to attain the canonical A tract conformation. There are two separate and distinct substrate requirements for the region 5′ of the nicking site of L1 endonuclease: (i) a Y·R base pair at position −4 (or the dinucleotide pair formed by the junction of this Y·R base pair with the flanking sequence) and (ii) an unobstructed minor groove at the −4 position. These two criteria are essentially a limited restatement of the requirements for formation of A tract DNA. Both requirements must be met for high-level nicking activity; in vivo, this translates into a strong preference for continuous T·A base pairs (an A tract on the complementary strand) in the region 5′ to the nicked phosphodiester.

Evidence That L1 Endonuclease Interacts with the DNA Minor Groove. Interference of protein−DNA interaction by covalent modification which is groove-specific or at a defined position on the DNA can provide a high-resolution view of protein contacts with the DNA surface. Other nucleolytic enzymes in the family which includes L1 endonuclease have been suggested or shown to contact DNA primarily via minor groove interactions (*33*, *34*). L1 endonuclease may therefore recognize DNA in a comparable manner. Consistent with this hypothesis, substitution of uracil in place of thymine (5-methyluracil) at all positions along the T_6A_6 substrate had no effect on the nicking ability or specificity of L1 endonuclease (Figure 4C). Additionally, L1 endonuclease is indifferent to major groove adenine methylation at positions −3 and +3 (see below). When the T_6A_6 target DNA was premethylated at adenine N3 (in the minor groove) by treatment with dimethyl sulfate (DMS), nicking by L1 EN was inhibited in a concentration-dependent manner, and was nearly completely abolished at maximal methylation (Figure 5A).

Several ligands of DNA have been described which bind exclusively in the minor groove, and preferentially in the minor groove of AT-rich sequences (*35*). When the substrate T_6A_6 was incubated with the AT-rich minor groove ligand distamycin A before reaction with L1 endonuclease, inhibition of nicking activity was observed (Figure 5B). Approximately 118 nM distamycin A was required to reduce nicking activity by 50% at the TA junction, a value quite similar to the K_d of a distamycin $A-T_5{\cdot}A_5$ complex (67 nM) as determined by circular dichroism analysis under similar buffer conditions (*36*). Distamycin A also appears to stabilize the DNA helix, making it resistant to denaturation (Figure 5B). The banding pattern was unaffected by the addition of 10 μm distamycin A after completion of the nicking reaction, demonstrating that the inhibition observed is in fact repression of nicking activity rather than protection of the nicked product from denaturation. Inhibition of nicking activity by berenil (another minor groove binding molecule) was also observed, although higher concentrations were required to achieve equivalent inhibition (data not shown). Interestingly, blockage of the TA junction region by distamycin A redistributed a small amount of L1 EN nicking to the 3′ terminus of the oligo. Nicking at this secondary region was itself eliminated at elevated distamycin A concentrations. Thus, occlusion of the minor groove blocks nicking by L1 endonuclease.

Minor Groove and TpA Structural Parameters Are Recognized by L1 Endonuclease. Poly(dA) tracts have well-defined structural features; excellent base stacking results in a large propeller twist and, importantly, a narrow minor groove (*37*−*41*). T_nA_n regions consist of convergent polyA tracts, but at the TpA dinucleotide exhibit several structural peculiarities unique to this junction. Base stacking at this position is quite poor due to steric clash of the purine rings, disrupting the continuity of the narrow poly(dA) minor groove as manifest by an abrupt widening and concomitant

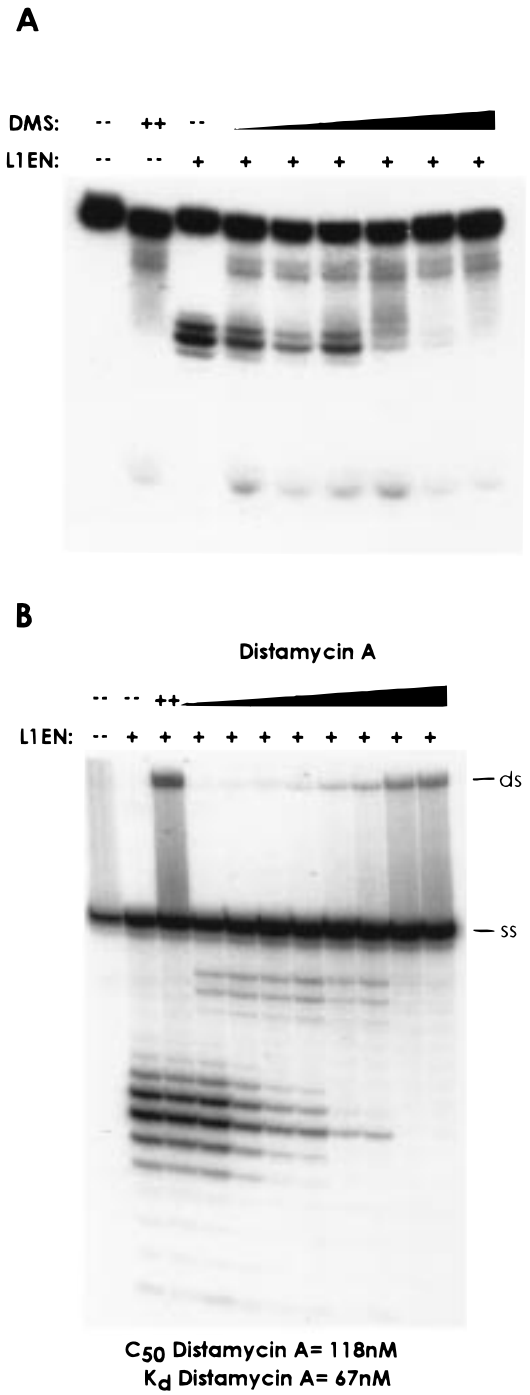

FIGURE 5: Evidence that L1 endonuclease interacts with the DNA minor groove. (A) DMS methylation of minor groove adenine N3 prevents L1 EN nicking. The T_6A_6 substrate was methylated with 0, 1, 5, 10, 20, 40, or 80 μL of dimethyl sulfate, and then exposed to L1 endonuclease (+) or buffer only (− −). DNA treated with the maximal amount of DMS (80 μL) but not L1 EN is indicated as (++). (B) Inhibition of nicking activity by distamycin A. The T_6A_6 substrate was incubated for 20 min with buffer only (− −) or 2-fold dilutions of distamycin A (Sigma) ranging from 0.078 to 10 μM and then incubated with L1 endonuclease (+). Ten micromolar distamycin A was also added after 1 h incubation with L1 endonuclease (++).

loss of an ordered array of bound water molecules (*41*−*44*). The regions which are nicked by L1 endonuclease are therefore also regions at which the width of the minor groove changes from very narrow to very wide, and then back to very narrow.

Dimethyl sulfoxide (DMSO) has been shown to dehydrate and unwind the DNA helix, thereby widening the minor groove (*45*). If minor groove width is crucial to binding/cleavage by L1 endonuclease, modulation of the minor groove width by DMSO should similarly alter the activity of L1 endonuclease. As can be seen in Figure 6A, inclusion of DMSO in the L1 nicking reaction dramatically increased nicking activity on the T_6A_6 substrate in a dose-dependent manner, to at least 7-fold over the reaction lacking DMSO in this experiment (as measured by the decrease in substrate). Higher levels of stimulation were often seen. Activity on the A_6T_6 substrate was also increased, with increasing nicking spreading inward from the termini [the region of A tracts which naturally have the widest minor groove (*42*)], creating a pattern of activity resembling the inverse of the T_6A_6 substrate, but diminished in intensity. Several other DNA dehydrating agents gave similar results, in rough proportion to their ability to unwind the DNA helix as reported by Lee et al. (data not shown), suggesting that these cosolvents affect the structural parameters of the target DNA helix rather than the L1 endonuclease itself. These data indicate that L1 endonuclease is sensitive to the groove width of DNA, and suggest either that L1 endonuclease recognizes DNA segments which are underwound and/or that DNA unwinding is the rate-limiting step in binding/cleavage events.

These data are also consistent with the hypothesis that the stimulation of nicking activity observed above, or the specificity of L1 endonuclease in general, is determined by the hydration of the DNA minor groove. In this model, displacement of water from the minor groove is the rate-limiting step for binding of L1 EN. DMSO would thus stimulate activity by removing water from the minor groove, and the general specificity of L1 endonuclease is a consequence of the relatively disordered (and therefore readily displaceable) water found naturally at the junction of T_nA_n tracts. If dehydration were the sole determinant of L1 EN specificity, disruption of the T_nA_n region by insertion of guanosine should increase activity near the site of insertion (but most likely block activity at the insertion point) via disruption of minor groove hydration by the guanosine C2 amino group, and replacement by inosine should not. As neither effect is observed (see Figure 4A,B), we conclude that a disordered or dehydrated minor groove is not sufficient for nicking by L1 endonuclease. However, as DNA hydration and structural parameters are clearly dependent variables, it is impossible to exclude from the overall stimulation of activity a contribution from dehydration alone (although in vivo, dehydration is likely to be secondary to unwinding).

Groove width is also increased when DNA is unwound by negative supercoiling (*46*). To provide a means of addressing the structural specificity of L1 endonuclease reliant on purely physical (rather than chemical) methods, the susceptibility of individual topoisomers to L1 EN was examined. When a sample containing a wide distribution of topoisomers was reacted with L1 endonuclease, plasmids were depleted from the substrate pool in proportion to (the absolute value of) their linking number (Figure 6B). More relaxed DNA was nearly untouched even after most of the other topoisomers were completely converted to the open circle form. While it is possible that the rate enhancement observed here could be due to contacts specific for supercoiled DNA, we believe it more likely that the stimulation

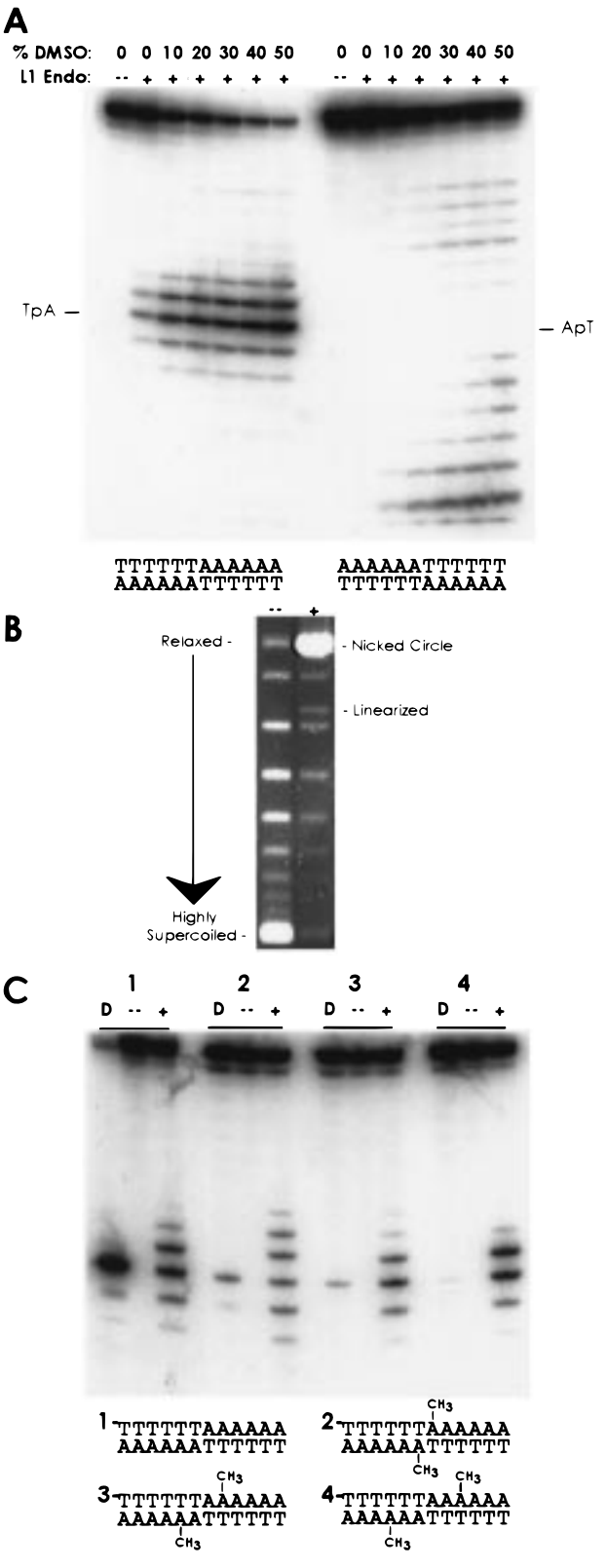

FIGURE 6: L1 endonuclease recognizes T_nA_n structural parameters. (A) DMSO enhances L1 endonuclease activity. L1 endonuclease was incubated for 10 min with either the T_6A_6 or the A_6T_6 substrate in the presence of 0−50% (v/v) dimethyl sulfoxide. (B) Highly supercoiled topoisomers are preferentially substrates for L1 EN. A distribution of topoisomers was generated and reacted with L1 EN. (C) TpA structural features are recognized by L1 endonuclease. Oligonucleotides with N^6-methyladenine incorporated were reacted with L1 EN.

results from the increasingly widened minor groove found in increasingly supercoiled plasmids.

NMR analysis of TpA junctions within T_nA_n tracts has shown that the adenines at and near the TpA step are free to sample unusual conformational states (*38*). The adenine bases at this position are thought to oscillate about the glycosidic bond some 30−50° at several thousand hertz (*38*). This mobility is a signature of the DNA structural environment found at and near TpA steps. N^6-methylation of these adenine residues destroys this conformational plasticity, eliminating the diagnostic NMR line broadening of the adenine protons at and near the TpA junction (*40*). As L1 endonuclease cleaves near this junction, recognition of this unique structural feature may play a role in substrate selection. To test this hypothesis, various methylated derivatives of the T_6A_6 substrate were synthesized, and the ability of L1 endonuclease to nick them was analyzed (Figure 6C). Methylation of the TpA step (−1 and +1) adenines diminished nicking at the TpA junction about 4-fold; methylation of the adjacent (−2 and +2) adenines had a more modest negative effect, while the nicking of the substrate with the third (−3 and +3) adenines methylated was almost identical to the unmethylated sequence. The extent of perturbation of nicking activity directly correlated with the magnitude of adenine mobility normally found (*38*) at the base modified.

It is formally possible that the moderation of nicking observed in this experiment is a result of blockage of major groove interactions near the TpA junction. In light of the multiple lines of evidence supporting minor groove interactions and because nicking activity seems to correlate with adenine mobility, we consider this explanation unlikely, but cannot exclude it.

Taken together, the preferences detailed above can be used to assemble a consensus sequence favorable for nicking by L1 endonuclease. L1 EN prefers poly(dT)·poly(dA) (A tract) DNA 5′ of its nicking site (especially 4 bp prior to the nicked phosphodiester), greatly favors nicking at YR junctions, and recognizes the peculiar conformation at the TpA junction and one base thereafter. When compiled, these data indicate that the sequence T_4A_2 would fulfill the requirements for L1 EN nicking. This sequence is indeed capable of supporting the pattern of nicking activity seen with the T_6A_6 substrate, even when imbedded in a region of 100% GC content. While the presence of the flanking 100% GC region did not affect the pattern of activity, the absolute amount of nicking of the T_4A_2 substrate was reduced when compared to T_6A_6 (data not shown). This effect of flanking sequence on activity is a property which might be expected for a nuclease which is structurally (rather than sequence) specific, as any single nucleotide influences DNA structural parameters over a several base pair window of sequence.

L1 Endonuclease Is Not an AP Endonuclease. Significant amino acid sequence similarity exists between L1 endonuclease and nucleases which have apurinic/apyrimidinic nuclease activity (*33*, *47*, *48*). Previous experiments have shown that L1 endonuclease does not prefer plasmids with AP sites to intact DNA (*13*). In the case of an enzyme which has the ability to nick intact DNA, a plasmid-based assay is an inherently poor system with which to assay this activity because potential L1 endonuclease AP activity is masked and competed by nicking at intact consensus sites also on the plasmid. L1 EN may therefore exhibit AP endonuclease

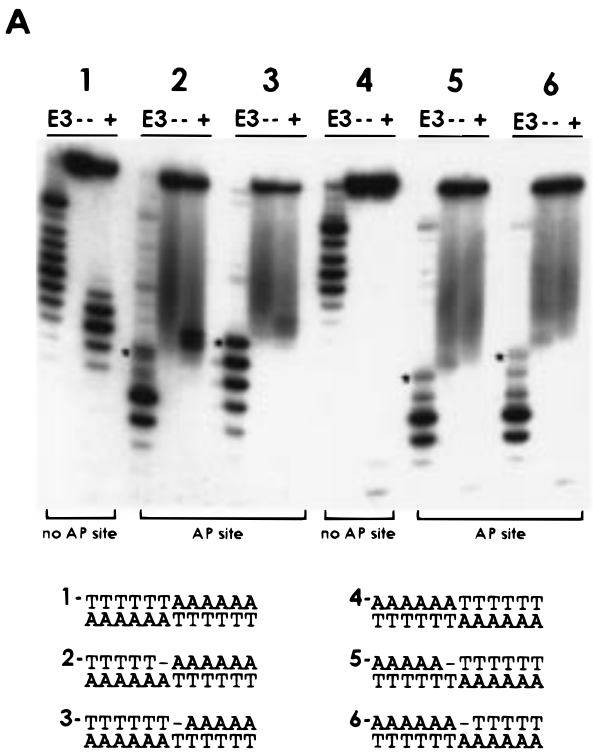

FIGURE 7: L1 endonuclease is not an AP endonuclease. (A) L1 endonuclease nicking at apurinic or apyrimidinic sites. Numbering indicates the position of the radioactive label; D, *Dra*I; E3, exonuclease III; (− −)-, buffer only; (+) L1 endonuclease. The asterisk indicates the position of the AP site and therefore the position of nicking by Exo III, and the lack of nicking by L1 EN. The exonuclease activity of exo III is not completely suppressed by Ca^{2+}, resulting in nicking at the abasic site followed by limited (2−3 nt) 3′→5′ digestion. The smear observed in the buffer only and L1 endonuclease lanes corresponds to the products of spontaneous hydrolysis of the phosphodiester backbone near the AP site during electrophoresis.

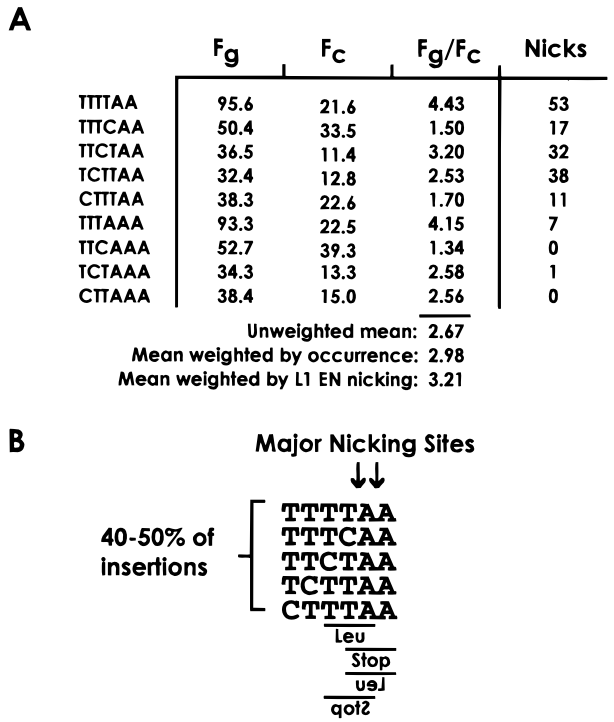

FIGURE 8: L1 target sites are nonrandomly distributed within the human genome. The frequency of the major L1 EN nicking/L1 element integration sites in whole genomic (F_g) and coding (F_c) sequence. For clarity, the frequencies are displayed as percentage $\times 10^3$. An average hexamer will occurr at a frequency of 24.4 ($1/4^6$). The values reported under "Nicks" are the absolute number of insertions recovered by Jurka in which that particular target site was used. The mean of the F_g/F_c ratios was normalized independently both by the frequency of each hexamer in the genome and by its pattern of usage by L1 EN. Note that the hexamers most overrepresented in whole genomic sequence are also those which are used the most by L1 EN. The true bias is likely to be at least 3.52-fold, the sum of the unweighted mean and the enhancements above this mean produced by both weighted averages.

activity at a rate equal to or less than its general nicking activity. To examine the ability of L1 endonuclease to recognize specific abasic sites, double-stranded oligos containing uracil were reacted with uracil DNA glycosylase to create specific apurinic or apyrimidinic sites, and then incubated with L1 endonuclease. In contrast to the strong AP endonuclease activity of *E. coli* exonuclease III, L1 endonuclease does not appreciably nick at the AP site (Figure 7). In addition, nicking between bases immediately 5′ and 3′ of the AP site is greatly reduced. Thus, L1 endonuclease not only lacks AP endonuclease activity at abasic sites but also is inhibited by the lack of bases near its cleavage site.

Repair of AP sites is a challenge faced by all organisms. Enzymes responsible for this type of DNA repair are consequently highly conserved across kingdoms. As a result, APases are often functionally equivalent and therefore interchangeable [see (*48*), for example]. Consistent with the above biochemical data, expression of L1 endonuclease in *xth* (Exo III deleted) *E. coli* and *apn1* (the *S. cerevisiae* Endo IV homologue) yeast does not complement either phenotype. (Q. Feng and J.D.B., unpublished data). In light of these results, we conclude that L1 EN is not an AP endonuclease.

L1 Target Sequences Are Nonrandomly Distributed in the Human Genome. There is a strong selective pressure on parasites to minimize damage done to the host organism. We have therefore examined the biochemical properties of L1 EN conscious of the parasitic relationship between L1 elements and the human genome. This principle of host−parasite interaction predicts that these properties will have been optimized to avoid deleterious insertions into human coding or regulatory sequence. Analysis of database sequences indicates that approximately 40−50% of *Alu* insertions occur at the TpA or the adjacent ApA junction of the sequence TTTTAA or one of the four possible T to C substitutions of this sequence (*25*). The bulk of the remainder occur primarily at the TT or AA junction, or the YA junction of purine→purine or pyrimidine→pyrimidine substitutions of this sequence (*25*). If such insertions are indeed predisposed to be innocuous, then this sequence should be found less frequently in coding sequence than in whole genomic DNA. Investigation of this hypothesis reveals that there is indeed a 3-fold greater frequency of optimal target sites in genomic DNA than in coding DNA (Figure 8A). Assuming that 95% of total DNA is noncoding, this moderate frequency bias represents a large absolute overrepresentation of L1 EN target sites in noncoding DNA. L1 elements are therefore biased toward insertion into noncoding DNA.

Conceptual translation of these sequences is depicted in Figure 8B. L1 endonuclease nicking sites in coding sequence contain TTA-Leu [nearly the most rare codon in human

coding sequence (*28*)] in the +1 frame (4/5 times) and −2 frame (3/5 times), and stop codons in frames +2 (5/5 times) and −1 (4/5 times). Thus, 16/30 reading frames corresponding to potential L1 endo nicking sites are either rare in coding DNA or already result in the termination of translation. Further investigation revealed that these sequences (dicodons) were not underrepresented within coding sequence itself relative to the frequency of their component codons (data not shown). Thus, L1 EN sites in coding sequence are likely to occur at positions where the effects of L1 element integration on gene function are not likely to be large.

DISCUSSION

The preferred DNA sequence for L1 EN activity (T_4A_2) agrees quite well with the consensus sequence for L1 integration events inferred from database sequence (*25*), suggesting that this assay faithfully mimics the behavior of L1 endonuclease in vivo. The integration consensus reported by Jurka is TTAAAA; as this was derived from inspection of integrated elements in the database, it is the reverse complement of the TTTTAA sequence which is preferred by the endonuclease domain itself (see Figure 1B). This consensus is consistent with the target-primed reverse transcription (TPRT) model of Luan and Eickbush in which the bottom strand (the strand which contains the complement of the integrated element) is nicked first. In this model, the endonuclease makes two nicks in genomic DNA. As the position of the first nick severely constrains that of the second, an exactingly specific nuclease might detrimentally limit the spectrum of favorable integration sites. Thus, the somewhat loose consensus obtained from our experiments fits well into the TPRT model derived for the R2Bm element, and is an important experimental corroboration of this model as extended to L1 element retrotransposition. In addition, the integration consensus culled from the database was built from analysis of *Alu* integration sites, with the assumption that *Alu* elements employ L1-encoded proteins for their own replication [presumed, therefore, to also be mediated by TPRT (*25*)]. Our data therefore provide the first empirical support for this hypothesis as well.

Sequences found to be good substrates for nicking by L1 endonuclease contain A tract DNA 5′ of the nicked bond. The A tract character of these sequences appears important for their recognition, as base modifications and substitutions which reduce A tract bending also reduce L1 EN nicking; L1 EN is indifferent to modifications which do not change A tract geometry. The physical peculiarities of A tract DNA result in chromatin structure abnormalities when this DNA is assembled onto nucleosomes. Long stretches of A tract DNA are refractory to wrapping on nucleosomes [(*49*) and references cited therein]. Shorter lengths of polyA containing DNA are preferentially positioned at the end positions of the core histone octamer; indeed, local placement of nucleosomes is thought to be strongly influenced by this sequence-dependent factor (*49*). The integration specificity of L1 elements can be represented by either of two models: (i) L1 integration might be actively targeted by an endonuclease specific for T_nA_n regions, or (ii) integration specificity may be passively determined by a nonspecific endonuclease recognizing the face of the histone octamer where A tract DNA is found, or recognizing DNA devoid of nucleosomes (longer A tracts). Whereas database analysis is not sufficient to distinguish between these mechanisms, the data presented here (and in Feng et al.) favor active endonuclease-mediated targeting of L1 elements. While the global choice of integration sites may be determined by the accessibility of DNA within chromatin, on a local scale the endonuclease domain is likely to be the primary determinant of the specificity of L1 element nicking and integration.

The structural significance of A tract DNA was first realized when it was observed that phased A tracts exhibit anomalous mobility during electrophoresis, a consequence of the alignment of the bends produced by individual polyA regions (*50*). It is therefore interesting to note that, whereas the structure of the A tract DNA is important for L1 EN recognition and cleavage, nicking actually occurs at segments of DNA which are straight (in the absence of L1 EN); the helix axis is straight despite the structural discontinuity associated with the TpA step (*40*, *51*). Nicking near the intrinsic bend so characteristic of the A tract is actively avoided by L1 EN; cleavage of the A_6T_6 substrate is minimal where the bend at the 3′ end of the A tract is maximal.

The effects seen on L1 endonuclease nicking of single base substitutions in the T_6A_6 oligo implicate A tract DNA structure in L1 EN substrate selection because these substitutions similarly affect A tract bending (*30*, *32*). These substitutions are capable of complete abrogation of A tract curvature, yet disrupt nicking at only a single phosphodiester bond (between the third and fourth bases 3′ to the base substitution). A possible explanation for this is that L1 EN senses DNA structure at a discrete position(s), and that the substitutions made eliminate bending by one mechanism and more subtle aspects of A tract structure by another. L1 EN nicks at the end of the A tract opposite from the bend, and A tract length is relatively unimportant for nicking (Figure 2D); bend recognition is therefore probably not crucial for L1 EN recognition, but disruption of bending may happen to correlate with (but not cause) disruption of an independent structural parameter recognized by L1 EN.

While most nicking by L1 EN occurs at the TpA bond of T_nA_n tracts, significant cleavage is also observed at the flanking dinucleotides. Two models exist which explain this cleavage pattern: (i) DNA recognition by L1 EN is quite specific, but a flexible active site permits nicking between a spectrum of nearby nucleotides; or (ii) L1 EN possesses an active site which is immobile relative to the DNA binding residues, and the DNA near the TpA junction is bound in several discrete registers. Given the structural specificity of L1 EN, the gradient distribution of DNA structural parameters such as minor groove width in A tract DNA, and the homology of L1 EN to nucleases whose structures are known, we predict that the second of these models is likely to be the correct one.

We found that minor groove width is an important factor for binding/cleavage by L1 EN. The TpA junction of T_nA_n tracts normally has a wide minor groove, and is flanked by unusually narrow minor grooves in the homopolymeric DNA to the left and right. As regions of high GC content are thought to have a wide minor groove (*18*), yet are resistant to L1 endonuclease cleavage (G.J.C. and J.D.B., unpublished data), a large minor groove width appears to be necessary but not sufficient for cleavage by L1 endonuclease. Similar suggestions correlating minor groove width with enzymatic activity have been made regarding DNase I (*18*). In the case

of L1 endonuclease, however, contacts specific for TA-rich sequences (or the lack of contacts specific for GC-rich sequences) might limit this nicking activity. Considering the preference for A tract DNA structure 5′ of the cleavage site, perhaps the specific transition of the minor groove from narrow to wide is the structural feature most pertinent for recognition. If for longer A tracts the magnitude of the minor groove width change is independent of the length of the tract, then the T_9A_9 substrate presents a relatively shallow gradient of groove width change per base pair. In contrast to all other substrates on which nicking is only observed at the first (and to a much lesser extent, the second) thymidine dinucletode, nicking of T_9A_9 (Figure 2D) between several thymidine dinucleotides is seen. This suggests that L1 EN recognizes and nicks DNA with a specific range of minor groove widths, and explains the more widely spread nicking observed on the T_9A_9 substrate compared to substrates with shorter A tracts. A tract DNA recognition is also seen in the IHF−DNA cocrystal structure: the narrow minor groove of a six base pair A tract is recognized on a structural (rather than sequence-specific) basis (*52*). Minor groove edges of bases 5′ of the nicked base interact with N74 and Y76 in the crystal structure of DNase I in complex with DNA (*34*). Similarly, the structural integrity of sequences 5′ of abasic sites was found to be required for the AP endonuclease activity of Exo III (*26*). These two residues are not well conserved between L1 endonuclease, DNase I, and Exo III (*47*). We propose that L1 endonuclease contacts DNA in an analogous but more restrictive fashion, and that such interactions are incompatible with the disruption of A tract DNA topology, particularly 5′ of the nicking site. Consistent with this, we note significant inhibitory effects of base substitutions 5′ of the nicking site and no inhibitory effect of symmetrical substitutions made 3′ of the nick site. As structural specificity to some extent requires sequence specificity, particular protein−DNA interactions which are in fact sequence-specific may occur, but are unlikely to be sufficient to mediate efficient binding/cleavage in the absence of an overall favorable structural context. As a result, L1 endonuclease is semispecific, positioned between extremely discriminating homologues such as R2Bm, R1Bm, and Tx1 endonucleases and essentially nonspecific nucleases such as DNase I. DNase I has found extensive use as a probe of DNA structure and DNA−protein interactions. L1 endonuclease might be similarly useful with regard to more specific questions of A tract DNA structure, and nucleosome or chromatin organization.

The wide minor groove at the TpA junction of T_nA_n regions is a consequence of local sequence-dependent unwinding of the helix. When the substrate was further unwound (chemically with DMSO and physically by supercoiling), L1 endonuclease activity increased. This phenomenon may have relevance in vivo, as evidence suggests that the genome is divided into nonrandom torsionally constrained and differentially supercoiled segments [(*53*) and references cited therein]. Although poorly characterized, these regions may affect L1 element targeting by providing alternately favorable or poor substrates for L1 EN. Supercoiling is also altered by transcription (*54*) and perhaps DNA replication. Although more localized, these processes may also influence L1 targeting.

Removal of the unusual structure of the TpA dinucleotide at the junction of T_nA_n tracts negatively influences nicking by L1 endonuclease. As nicking activity is diminished but not completely lost on these substrates, the TpA conformational dynamic must be significant but not essential for recognition or cleavage of DNA by L1 endonuclease. Additionally, the T_5CGA_5 substrate is nicked well at the CpG bond; perhaps the CpG dinucleotide can exhibit similar conformational fluctuations when at the center of a poly(dT)·poly(dA) tract.

The type II restriction enzyme *Dra*I recognizes the same sequence as L1 EN and nicks both strands at the TpA bond. Aside from this superficial similarity, *Dra*I appears to function quite differently from L1 endonuclease. N^6-methylation of the T_6A_6 substrate inhibited *Dra*I substantially, presumably through major groove contacts. *Dra*I is naturally inhibited by *Dra*I methylase activity at adenines +3 and −3, but shows >90% inhibition at both other positions (adenines +1 and −1, and +2 and −2). In addition, *Dra*I activity is completely inhibited by 50% DMSO (G.J.C. and J.D.B., unpublished data), and is reduced when thymidine methyl groups are removed, but is insensitive to the disruption of A tract structure outside its recognition site at position −4. Thus, L1 endonuclease and *Dra*I represent two independent and quite different solutions to similar nucleic acid recognition problems.

Not surprisingly, L1 endonuclease is not similar to *Dra*I on the protein sequence level. L1 EN is, however, similar to other non-LTR retrotransposon endonucleases. While these nucleases have many residues in common in addition to the absolutely conserved catalytic core residues, a sufficient number have diverged to preclude confident prediction of their specificities based on this work. The endonuclease domain of R1Bm, for example, is specific for a 14 bp segment of mixed sequence in the insect rDNA (*21*). If the evolutionary speculations outlined below have played a role in determining the specificity of L1 endonuclease, then perhaps similar results will emerge from investigation of these nucleases, particularly in organisms whose genome contains a high percentage of polyA sequences of retrotransposon origin.

The L1 endonuclease studied here is derived from the major subfamily of active L1 elements. Other minor subfamilies of L1 elements may have endonucleases of slightly differing specificities. The retrotransposition activity of these elements may account for some of the *Alu* and L1 insertions that do not precisely conform to the consensus L1 EN sequence. The nicking activity investigated here represents the current specificity of L1 EN. Nicking by evolutionarily more ancient versions of L1 EN may also explain some of this apparently anomalous insertion specificity.

All DNA repair enzymes which are members of the family containing L1 endonuclease display apurinic/apyrimidinic endonuclease activity. The endonuclease domain of the retrotransposon L1Tc has also been shown to possess APase activity (*48*). A priori, one might expect site-specific retrotransposon nucleases to have no need for such an activity, whereas nonspecific or semispecific retroelement endonucleases may or may not be APases (such a specificity might be beneficial by recruiting DNA repair/synthesis machinery useful for second-strand synthesis). We have found no evidence for AP endonuclease activity associated with L1

EN. It is possible that a factor critical for APase activity was absent from our assays; given the apparent absence of regions of protein structure believed to mediate APase activity, however (G.J.C. and J.D.B., unpublished data), we believe it is more likely that L1 endonuclease is simply not an AP endonuclease. Residual nicking activity was observed at the abasic site and at positions flanking the abasic site in the substrate analogous to T_6A_6. Rather than an APase activity which nicks when the absence of a base is recognized, we believe this low level of nicking results *in spite* of the absence of a base. The activity may be lessened either because of missing protein−DNA contacts or perhaps due to the significant disruption of DNA structure which accompanies abasic sites in A tract DNA (*55*).

We found that sequences which are targets for L1 endonuclease are overrepresented in noncoding regions of the human genome relative to coding regions. Much of this bias is likely a result of comparison between generally AT-rich noncoding sequence and coding DNA which has disproportionately high GC content. Indeed, A_nT_n sequences are similarly overrepresented in genomic DNA (data not shown). However, the cause of the bias itself is relatively unimportant; what is significant is that L1 EN has arrived at a specificity which allows this bias to be exploited. Even among AT-rich sequences, only the T_nA_n sequence has the potential advantages illustrated in Figure 8B. Many L1 EN hexamers are undoubtedly the remains of the 3′ polyA tails of previous L1/*Alu* insertions (which are by definition not in coding sequence). Indeed, L1 endonuclease nicks well at the 3′ end of the recently transposed tPA intron 25 *Alu* (data not shown) (*56*). Retroelements may therefore provide target sites for further insertions deleterious to neither the host nor the target element. Other retroelements are specific to repeated sequences (R2Bm, Ty1, R1Bm, DRE), and other transposons in particular (Tx1). L1 elements may therefore fit loosely into this category. While these observations may in fact only be coincidental to the evolutionary biology of L1 elements, it is intriguing to realize that few possible specificities could achieve similar results. The semi-specificity of L1 endonuclease therefore likely creates for L1 elements an optimal balance between integration efficiency and host genomic damage.

ACKNOWLEDGMENT

Thanks to Jeffrey S. Smith, David Symer, and Eleanor F. Hoff for critical evaluation of the manuscript, and to all Boeke lab members for stimulating discussion and suggestions. In addition, we thank J. Jurka and L. Hurley for providing helpful comments and suggestions.

REFERENCES

1. Smit, A. F. A. (1996) *Curr. Opin. Genet. Dev. 6*, 743−748.
2. Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., Gabriel, A., Swergold, G. D., and Kazazian, H. H., Jr. (1997) *Nat. Genet. 16*, 37−43.
3. Meuth, M. (1989) in *Mobile DNA* (Berg, D. E., and Howe, M. M., Eds.) pp 833−860, American Society of Microbiology, Washington, D.C.
4. Hutchison, C. A., III, Hardies, S. C., Loeb, D. D., Shehee, W. R., and Edgell, M. H. (1989) in *Mobile DNA* (Berg, D. E., and Howe, M. M., Eds.) pp 593−617, American Society of Microbiology, Washington, D.C.
5. Singer, M. F., Skowronski, J., Fanning, T. G., and Mongkolsuk, S. (1988) in *Eukaryotic Transposable Elements as Mutagenic Agents*, pp 71−78, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
6. Howard, B. H., Russanova, V. R., and Englander, E. W. (1995) in *The Impact of Short INterspersed Elements (SINEs) on the Host Genome* (Maria, R. J., Ed.) pp 133−141, R. G. Landes, Austin, TX.
7. Makalowski, W. (1995) in *The Impact of Short INterspersed Elements (SINEs) on the Host Genome* (Maria, R. J., Ed.) pp 81−104, R. G. Landes, Austin, TX.
8. Laurent, A. M., Puechberty, J., Prades, C., Gimenez, S., and Roizes, G. (1997) *Genomics 46*, 127−132.
9. Kapitonov, V. V., Holmquist, G. P., and Jurka, J. (1998) *Mol. Biol. Evol. 15*, 611−612.
10. Swergold, G. D. (1990) *Mol. Cell. Biol. 10*, 6718−6729.
11. Hohjoh, H., and Singer, M. F. (1997) *EMBO J. 16*, 6034−6043.
12. Hohjoh, H., and Singer, M. F. (1996) *EMBO J. 15*, 630−639.
13. Feng, Q., Moran, J., Kazazian, H., and Boeke, J. D. (1996) *Cell 87*, 905−916.
14. Mathias, S. L., Scott, A. F., Kazazian, H. H., Jr., Boeke, J. D., and Gabriel, A. (1991) *Science 254*, 1808−1810.
15. Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., and Kazazian, H. H., Jr. (1996) *Cell 87*, 917−927.
16. Luan, D. D., Korman, M. H., Jakubczak, J. L., and Eickbush, T. H. (1993) *Cell 72*, 595−605.
17. Boeke, J. D. (1997) *Nat. Genet. 16*, 6−7.
18. Drew, H. R., and Travers, A. A. (1984) *Cell 37*, 491−502.
19. Demple, B., and Harrison, L. (1994) *Annu. Rev. Biochem. 63*, 915−948.
20. Sander, M., Lowenhaupt, K., and Rich, A. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 6780−6784.
21. Feng, Q., Schumann, G., and Boeke, J. D. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 2083−2088.
22. Jakubczak, J. L., Burke, W. D., and Eickbush, T. H. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 3295−3299.
23. Garrett, J. E., Knutzon, D. S., and Carroll, D. (1989) *Mol. Cell. Biol. 9*, 3018−3027.
24. Roberts, R. J., and Halford, S. E. (1993) in *Nucleases* (Linn, S. N., Lloyd, R. S., and Roberts, R. J., Eds.) pp 35−88, Cold Spring Harbor Press, Cold Spring Harbor, NY.
25. Jurka, J. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 1872−1877.
26. Takeuchi, M., Lillis, R., Demple, B., and Takeshita, M. (1994) *J. Biol. Chem. 269*, 21907−21914.
27. Bowater, R., Aboul-Ela, F., and Lilley, D. M. (1992) *Methods Enzymol. 212*, 105−120.
28. Nakamura, Y., Gojobori, T., and Ikemura, T. (1997) *Nucleic Acids Res. 25*, 244−245.
29. Sander, M., and Benhaim, D. (1996) *Nucleic Acids Res. 24*, 3926−3933.
30. Diekmann, S., von Kitzing, E., McLaughlin, L., Ott, J., and Eckstein, F. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 8257−8261.
31. Diekmann, S., Mazzarelli, J. M., McLaughlin, L. W., von Kitzing, E., and Travers, A. A. (1992) *J. Mol. Biol. 225*, 729−738.
32. Koo, H. S., and Crothers, D. M. (1987) *Biochemistry 26*, 3745−3748.
33. Mol, C. D., Kuo, C.-F., Thayer, M. M., Cunningham, R. P., and Tainer, J. A. (1995) *Nature 374*, 381−386.
34. Sück, D., Lahm, A., and Oefner, C. (1988) *Nature 332*, 464−468.
35. Coll, M., Frederick, C. A., Wang, A. H., and Rich, A. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 8385−8389.
36. Chen, F. M., and Sha, F. (1998) *Biochemistry 37*, 11143−11151.
37. Chuprina, V. P., Lipanov, A. A., Fedoroff, O., Kim, S. G., Kintanar, A., and Reid, B. R. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 9087−9091.
38. Kennedy, M. A., Nuutero, S. T., Davis, J. T., Drobny, G. P., and Reid, B. R. (1993) *Biochemistry 32*, 8022−8035.

39. Kim, S. G., and Reid, B. R. (1992) *Biochemistry 31*, 12103−12116.
40. Lingbeck, J., Kubinec, M. G., Miller, J., Reid, B. R., Drobny, G. P., and Kennedy, M. A. (1996) *Biochemistry 35*, 719−734.
41. Yoon, C., Prive, G. G., Goodsell, D. S., and Dickerson, R. E. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 6332−6336.
42. Katahira, M., Sugeta, H., and Kyogoku, Y. (1990) *Nucleic Acids Res. 18*, 613−618.
43. Liepinsh, E., Leupin, W., and Otting, G. (1994) *Nucleic Acids Res. 22*, 2249−2254.
44. Quintana, J. R., Grzeskowiak, K., Yanagi, K., and Dickerson, R. E. (1992) *J. Mol. Biol. 225*, 379−395.
45. Lee, C. H., Mizusawa, H., and Kakefuda, T. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 2838−2842.
46. Vologodskii, A. V., and Cozzarelli, N. R. (1994) *Annu. Rev. Biophys. Biomol. Struct. 23*, 609−643.
47. Gorman, M. A., Morera, S., Rothwell, D. G., de La Fortelle, E., Mol, C. D., Tainer, J. A., Hickson, I. D., and Freemont, P. S. (1997) *EMBO J. 16*, 6548−6558.
48. Olivares, M., Alonso, C., and Lopez, M. C. (1997) *J. Biol. Chem. 272*, 25224−25228.
49. Satchwell, S. C., Drew, H. R., and Travers, A. A. (1986) *J. Mol. Biol. 191*, 659−675.
50. Marini, J. C., Levene, S. D., Crothers, D. M., and Englund, P. T. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 7664−7668.
51. Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. (1987) *Nature 330*, 221−226.
52. Rice, P. A., Yang, S., Mizuuchi, K., and Nash, H. A. (1996) *Cell 87*, 1295−1306.
53. Kramer, P. R., and Sinden, R. R. (1997) *Biochemistry 36*, 3151−3158.
54. Wang, J. C., and Lynch, A. S. (1993) *Curr. Opin. Genet. Dev. 3*, 764−768.
55. Wang, K. Y., Parker, S. A., Goljer, I., and Bolton, P. H. (1997) *Biochemistry 36*, 11629−11639.
56. Batzer, M. A., Gudi, V. A., Mena, J. C., Foltz, D. W., Herrera, R. J., and Deininger, P. L. (1991) *Nucleic Acids Res. 19*, 3619−3623.

BI981858S