6864–6878 *Nucleic Acids Research, 2011, Vol. 39, No. 16* *Published online 23 May 2011*
doi:10.1093/nar/gkr337

Genome-wide analysis of mobile genetic element insertion sites

Kamal Rawal[1,2] and Ram Ramaswamy[1,3,*]

[1]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110 067, [2]Department of Biotechnology, Jaypee Institute of Information Technology, Noida, Uttar Pradesh and [3]School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

Received April 15, 2010; Revised July 28, 2010; Accepted April 24, 2011

ABSTRACT

Mobile genetic elements (MGEs) account for a significant fraction of eukaryotic genomes and are implicated in altered gene expression and disease. We present an efficient computational protocol for MGE insertion site analysis. ELAN, the suite of tools described here uses standard techniques to identify different MGEs and their distribution on the genome. One component, DNASCANNER analyses known insertion sites of MGEs for the presence of signals that are based on a combination of local physical and chemical properties. ISF (insertion site finder) is a machine-learning tool that incorporates information derived from DNASCANNER. ISF permits classification of a given DNA sequence as a potential insertion site or not, using a support vector machine. We have studied the genomes of *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Entamoeba histolytica* via a protocol whereby DNASCANNER is used to identify a common set of statistically important signals flanking the insertion sites in the various genomes. These are used in ISF for insertion site prediction, and the current accuracy of the tool is over 65%. We find similar signals at gene boundaries and splice sites. Together, these data are suggestive of a common insertion mechanism that operates in a variety of eukaryotes.

INTRODUCTION

A prime challenge in current genomic studies is the accurate annotation of genomes with regard to locating genes, identifying promoters, CpG islands, repeats and indeed all significant features that could have a biological consequence. There are a number of computational approaches to this long-standing problem, many of which have been translated into software tools. Well-known examples include the whole-genome annotation protocols Ensemble (1), MaGe (2), GenDBan (3) and RiceGAAS (4) for the complete annotation of both prokaryotic and eukaryotic genomes. For the identification of well-characterized entities such as protein coding genes, a number of programs (5–7) are highly successful for both eukaryotic and prokaryotic genomes.

Precisely what constitutes a significant feature on the genome can be difficult to define, especially when the corresponding function is not well understood. Mobile genetic elements (MGEs) fall in this category. They are clearly important, constituting a major proportion of most eukaryotic genomes, and it is presumed that they are responsible for the significant expansion in genome size (8), while their role in regulation is still being uncovered (9). Originally such elements were considered to be parasitic (10) but recent studies have shown MGE to be involved in gene inactivation (11), transduction (12) and regulation (9). Their role in human genetic diseases (13) is also becoming apparent. A major class of MGE arises from a retrotransposition mechanism (14) and it is presumed that these have played a major role in genome evolution since they are more numerous in higher eukaryotes in comparison to more primitive organisms.

The problem we address in this article is the identification of potential insertion sites for MGEs. Clearly this depends both on the nature of the element and on the composition of the genome and the approach that we take here is integrative, using existing tools for the identification of elements in conjunction with bioinformatics analysis of the local environment of each class of element. This then makes it possible to discover new insertion sites as well as to analyze element behavior across genomes.

We briefly discuss the context of the problem. A number of methods have been developed over the years for the identification of transposable elements (TEs), and these have been summarized by Bergman and Quesneville (15). This article also drew attention to the fact that while existing methods—that use tools that include homology, comparative genomics as well as *de novo* techniques—can be

*To whom correspondence should be addressed. Tel: +91 11 2671 7568; Fax: +91 11 2671 7586; Email: r.ramaswamy@mail.jnu.ac.in

combined into fairly successful TE detectors, the discovery and annotation of new TEs remains a challenge in bioinformatics.

Mobile elements face relatively low-selection pressure, and thus evolve quite rapidly. This has two consequences. First, MGEs are often difficult to detect due to the fact that there is a high-nucleotide sequence divergence among them. Most copies of these elements are truncated at either or both of the 3′ and 5′ ends. This can be a problem in studying older elements whose classification sometimes requires considerable subjectivity.

A second consequence of low selection pressure relates to the local environment of the element, namely the insertion site itself. While these are also evolving rapidly, what we see at the present time is the set of TEs that have managed to stay on in a given set of sites, and not necessarily all the sites where TEs attempted insertion. In other words, the locations where we see MGEs today are those that were good insertion sites and have been good 'retention' sites. How the insertion sites have evolved, and whether their evolution is related to their capacity for retention of elements are important questions, but lie somewhat outside the scope of the present work.

The distribution of retrotransposons themselves is quite variable, with EhLINEs/EhSINEs occupying <20% of the *Entamoeba* genome (16,17), while LINEs and SINEs comprise ~50% of the human genome. Further, their internal architecture and the mechanism of their insertion and proliferation within the genome is quite distinct from that of protein-coding sequences. Additional complexity arises from the fact that each genome and its respective elements present a unique case in terms of data management and analysis. Sequencing methodologies and data formats differ, ranging from unassembled contig or scaffold data in the case of *Entamoeba histolytica*, to completely assembled genomes such as the 23 chromosomes of the human genome. The heterogeneity of sequences and the huge amount of data spread across a wide variety of useful genome databases such as Ensembl (http://www.ensembl.org), FlyBase (http://www.flybase.org) and Genbank (http://www.ncbi.nlm.nih.gov/) present another bioinformatics challenge for interspecies analysis.

In this article, we describe a computational protocol, ELAN website, http://nldsps.jnu.ac.in/elan.html, which is targeted towards genome-wide retro-transposon element analysis. Given an element that has been identified either by some existing protocol (15) or by the internal module ELEFINDER that is provided within ELAN, different components permit the following analyses:

(i) Genome-wide distribution profiles of elements currently present on the genome (to any desired level of homology). This is performed by the module ELEFINDER.
(ii) Extraction of consensus insertion sites and their subsequent analysis in terms of a number of physical and chemical properties. Retro-transposon insertion sites have distinctive structural features that derive from their specific composition. Physical properties which are pertinent in this context include DNA bendability (18,19), and propeller twist (20) as well as thermodynamic features such as stacking energy (21), duplex stability (22,23) and denaturation energy (24). The program DNASCANNER analyses insertion hotspots of elements in detail and provides a set of signals or characteristics that are potentially recognized by an element for its insertion.
(iii) Prediction of potential insertion sites via ISF, an Insertion Site Finder tool that employs machine-learning techniques that use the results from the application of DNASCANNER.
(iv) Comparative genomics of MGEs via a relational database, InSiDe (Insertion Site Database) which keeps a growing list of MGEs, their locations and their insertion environment.

The software and the data generated from the present study can be accessed online at the site http://nldsps.jnu.ac.in/elan.html.

MATERIALS AND METHODS

All DNA sequences continuously undergo mutations, insertions and deletions. In this scenario, determining what constitutes an 'element' can be a computational challenge. For our purpose, an 'element' is a partial or complete DNA sequence that has many of the characteristics of a given family of TEs such as a reverse transcriptase domain, endonuclease domain, poly A tail and intact boundaries determined through target site duplications (TSDs), etc. The element itself is generally constructed as a consensus of several copies retrieved through database search and subsequent processing is needed to construct a master or ancestor gene (25). A 'fragment', namely a sub sequence of an element occurs when reverse transcription is incomplete or by indel events, post-insertional mutations and so on. We divide fragments into three classes based upon their termini: 5′ truncated–3′ intact; 5′ intact–3′ truncated; and 5′ truncated–3′ truncated (Supplementary Figure S1).

ELAN

ELAN is a series of programs that run sequentially, each component attempting to address a question of biological relevance. As shown in Figure 1, the pipeline is divided into three major parts. First, copies of a given TE are located in the genome via a BLAST N (7) search, using Perl/Bioperl (26) scripts to parse output files. Elements as well as flanking sequences are extracted and processed, mainly to identify and merge fragments. Additional programs find truncation hotspots, the distances of TEs to genes as well as to other elements. Redundancies are removed by a standard protocol.

A number of genomes such as *Homo sapiens*, *Mus musculus* and *E. histolytica* and insertion elements listed in Repbase (27) have been included in the present study and information is kept current through continuous updation on the site http://nldsps.jnu.ac.in/elan.html. ELAN has been benchmarked against RepeatMasker and sample files as well as programs are given for comparison purposes in the Supplementary Data.

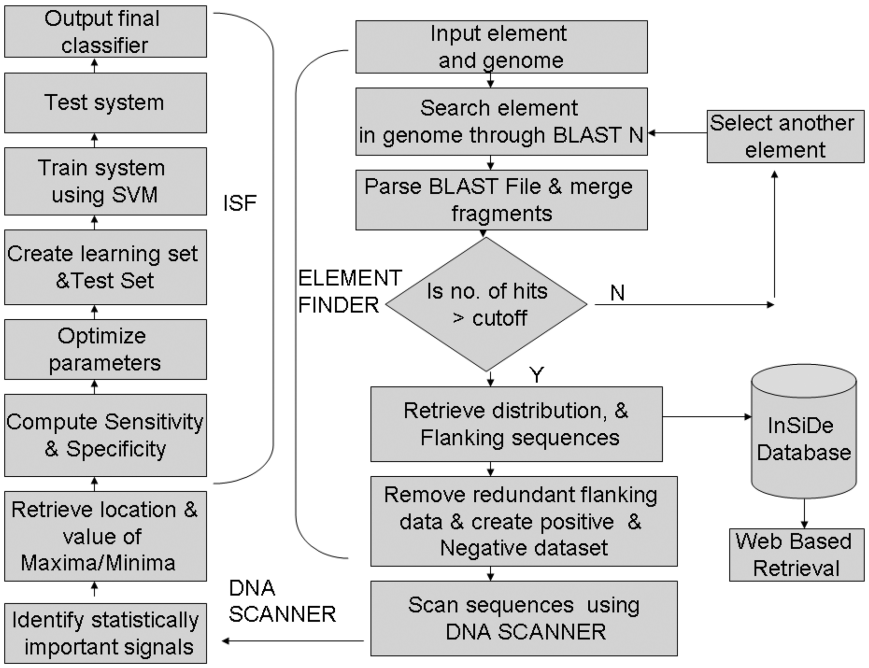


Figure 1. Schematic view of the various programs in the ELAN pipeline.

DNA SCANNER

DNA SCANNER scans genomic DNA for a number of different physicochemical properties by incorporating biophysical, thermodynamic, protein interactions and sequence based features. The algorithm is outlined in Figure 2. Based on a choice of input parameters, the program evaluates a number of properties in moving windows along the length of the query DNA sequence. Substrings of window size w are generated from the 5′-end of input DNA sequences, and further divided into words (Di/Tri nucleotides). For each of the properties (see below) parameter values T^{P} are derived and an average score S_{p} is computed as a function of position to generate a graph.

Structural signals: DNA bendability

DNA bendability is the ability of DNA to deform under a specific stimulus such as protein binding. Several models have been proposed to study relationship of sequence with structural bendability: there are both a trinucleotide model based on DNAse I cutting frequencies (18), and a dinucleotide model based on X ray crystallography of DNA oligomers, and kinetoplast DNA (28) in gel migration studies. It has been observed that highly bendable DNA contains tracts of A with 'loose periodicity' (29). The trinucleotide model (18) is based on the observations that certain enzymes such as DNAase I preferably bind and cuts DNA that is bent (or bendable) towards the major groove, and thus DNAse I cutting frequencies on naked DNA can be taken as quantitative measures of major groove compressibility and anisotropic bendability.

We have used this parameter earlier for studying the nature of pre-insertion loci (30) where the TPRT mechanism suggested that restriction endonuclease involved in retrotransposition appeared to require bent DNA for binding and nicking. Apart from this, specific bendability has been used as a feature to recognize *E. coli* promoters (31).

Thermodynamic signals: stacking energy

The relationship between sequence, structure and stacking energy has important biological consequences (32), for example, for sequence-specific interactions with proteins (33). Stacking energies are indicators of stability of the given DNA sequence as well as of protein interactions, and thus play an indirect role in formation of local structure (34,35). These include contributions from electrostatic, polarization, dispersion and repulsive forces. The total interaction energy of each stacked base pair is the sum of intra and inter-strand stacking energies (21). The stacking energies as a function of the rotational angle and separation distance between complementary pairs of all 16 dinucleotides have been used in DNA SCANNER.

Duplex stability: free energy signals

The relative stability of DNA duplex structure depends upon its base sequence (23), and more specifically upon ten different types of nearest neighbor interactions namely AA/TT; AT/TA; CA/GT; GT/CA; CT/GA; GA/CT; CG/GC; GC/CG; GG/CC. Using this information the overall stability (as a measure of ΔG) and melting behavior of a sequence can be predicted.

Propeller twist signals, bending stiffness and nucleosomal positioning

DNA must distort in order to bend around a protein: this distortion is facilitated by the deformational capacity of dinucleotides. Some are practically rigid whereas others

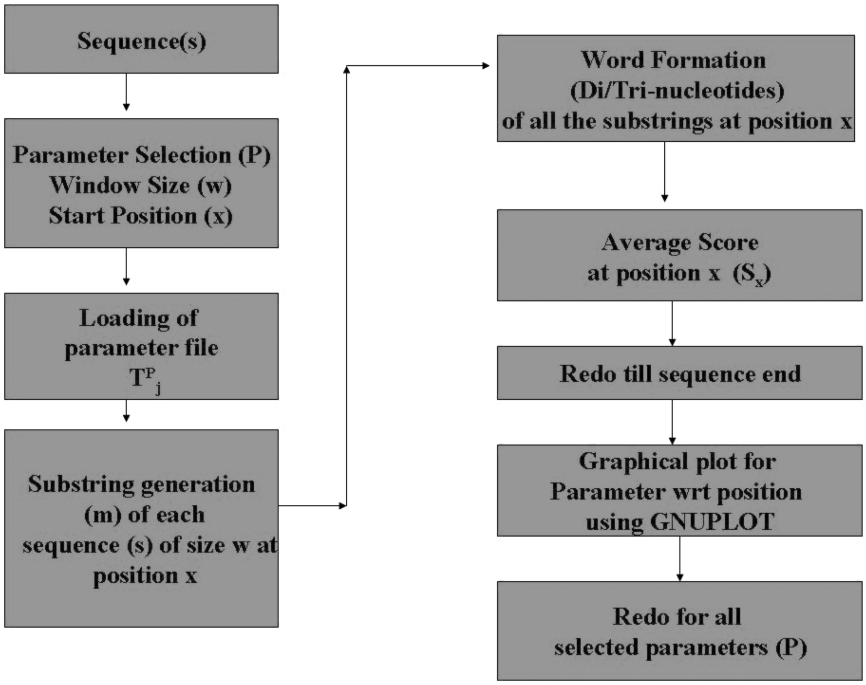


Figure 2. Flow chart depicting the sequence of procedures followed by DNA SCANNER to generate profiles for given DNA sequences. The score as function of position (x) is computed as $S_x = \sum_{j=1}^{m} T_j^P . T_j^P$ is defined as the parametric score of substring j (di/trinucleotide) derived from parameter file summed over the m substrings generated for a window of size w.

are flexible, and the propeller twist parameter measures the tendency of DNA to twist about the long axis that makes the two bases of a pair non-coplanar (20). Dinucleotides having a large propeller twist tend to be more rigid than dinucleotides with low propeller twist; thus propeller twist can also be used as a measure of DNA flexibility.

The conformational and mechanical properties of the DNA double helix vary in a sequence-dependent manner (36). Specifically, nucleosomal rotational setting and bendability is a function of underlying nucleotide sequence which contributes heavily on placement of nucleosome at specific position. Nucleosomal DNA is highly bent: histones prefer sequences which have the potential to bend for the formation of nucleosomal particles. Sivolob and Kharpunov (36) have shown that DNA sequences possessing low-bending energy correlate with potential bending and nucleosomal positioning.

Protein interaction signals

The DNA sequence carries signals specific for its potential to deform when interacting with other molecules such as proteins and also during important biochemical processes such as transcription, replication and retro-transposition. This deformability can also serve as potential long range signals for molecular recognition and conformational recognition (37), and based on information extracted from existing protein–DNA crystal complexes, empirical energy functions have been deduced. This gives the potential deformability of a given sequence and its potential to interact with proteins.

DNASCANNER has been developed in Perl/CGI-Perl and is modular in nature so that new properties can be included without changing the core code. The parametric data pertaining to the various physicochemical properties discussed above is stored in a separate flat file database; new rules can therefore be added or existing rules may be modified easily. The program is available online at http://tinyurl.com/dnascanner.

ISF

The information generated by DNASCANNER from positive and negative datasets is used by ISF to construct set of rules to classify and 'predict' insertion sites in genome.

Training and testing

Flanking sequences extracted from known insertion sites constitute pre-insertion loci which we take as the positive dataset, Class P. By shuffling these sequences we obtain a negative dataset Class N_a. This set of sequences maintains the base composition, but should not provide any suitable insertion sites. An independent negative dataset, Class N_b was constructed from randomly picked sequences in the vicinity of true insertion sites, and a third independent negative dataset, class N_c is constructed from sequences taken out of coding regions in the genome. Profiles for each property for both positive and negative datasets were generated as described above and averaged within each class. The position of the extremum (maximum or minimum) was noted for each parameter for the positive dataset (Figure 3) and was considered significant if it

exceeded 2 SD from the mean background (Table 1). The profile was also computed for the negative datasets and only properties that showed a strong differential between positive and negative sets were included in the model.

The nature and positions of an extremum differed in each of the chromosomes. Parametric files for each property were constructed and the maximum (or minimum) value and position were noted for each significant property (Table 1). This information was used as a quantitative measure to differentiate between positive and negative examples for a given chromosome (see Results section as well as Supplementary Data, tutorial.html).

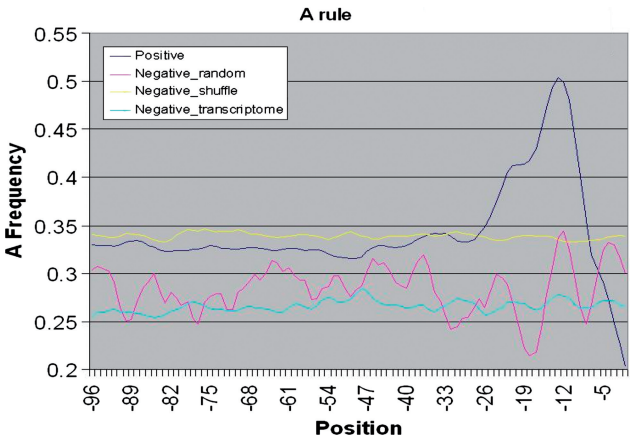


Figure 3. Adenine density upstream of insertion sites of Alu in human chromosome 2. The *y*-axis represents the value of the property under study (here this is the 'A Rule') and the *x*-axis represents the position with respect to insertion site (taken as position 0). In all cases, the properties we examine have been computed for both positive and negative datasets.

We base our scoring protocol to model insertion sites (P) or non-insertion sites (N) on Bayes' rule (38). Using the thermodynamic, biophysical and chemical properties discussed above the posterior probability or score of an insertion site, $P(P_i|S_j)$ is computed. This is the probability of classifying a DNA string P_i as an insertion site given that property S_j is true for this site. $P(P)$ and $P(N)$ are the prior probabilities of sites obtained from classes P and N, respectively; here the two classes are taken to be equiprobable. $P(S_j/P)$ denotes the conditional probability of the property being from class P calculated from training data. $P(S_j|N)$ can be computed in a similar manner, and the highest posterior is then taken as the true prediction.

Insertion sites are characterized by diverse parameters. The sensitivity and specificity based upon single property scores as determined by Bayes' rule were only ~45–55%. We therefore decided to use a combination of parameters in machine learning algorithms such as voting (39), Adaboost (40) and support vector machines (SVMs) (41). Extensive testing revealed that SVM performed better than other methods and emerged as the technique of choice in the present problem. A standard SVM was used to construct models for insertion sites of different elements, using binary scores (1 or 0).

The details of the procedure are as follows. Consider, as an example, the problem of finding Alu in Human Chromosome 22 (HC22). The training set $Z = (X, Y)$ consist of both positive and negative examples, each of which is characterized by a d-dimensional vector. Set X contains insertion site examples (labeled + 1) denoted by P and non-insertion site examples (labeled 0) as N. In order to implement SVMs, each insertion site was converted into feature vectors as described. The training set contained 595 insertion site and the same number of negative examples. Different kernels were used for learning and

Table 1. Information derived from DNASCANNER analysis of Human chromosome 2 (HC2) for Alu and L1 elements

Element	Human Chrosome 2							
	LINES (L1)				SINES (Alu)			
Number of sequences taken	209				5334			
Property	Position	Trend	Value	Mean (SD)	Position	Trend	Value	Mean (SD)
Trule	−28	*U**	0.347	0.29 (0.03)				
Arule	−14	*U*	0.530	0.36 (0.05)	−13	*U*	0.503	0.342 (0.05)
Grule					−15	*D**	0.140	0.194 (0.04)
Crule	−14	*D*	0.082	0.17 (0.03)	−13	*D*	0.102	0.173 (0.02)
Atrule	−14	*U*	0.769	0.65 (0.05)	−14	*U*	0.755	0.632 (0.05)
Bendability	−12	*D*	−0.023	−0.012 (0.003)	−12	*D*	−0.02	−0.01 (0.00)
Nucleosomal_positioning	−13	*D*	−3.887	−1.48 (0.91)	−14	*D*	−3.64	−1.14 (1.00)
b-a trimeric	−11	*U*	0.285	0.247 (0.01)	−12	*U*	0.281	0.244 (0.01)
DNA denaturation	−15	*D*	36.774	39.21 (1.06)	−14	*D*	37.03	39.55 (1.10)
Duplexstability	−14	*U**	−0.659	−0.72 (0.03)	−14	*U*	−0.66	−0.73 (0.03)
Propellartwist	−13	*D*	−7.454	−6.87 (0.23)	−13	*D*	−7.39	−6.79 (0.24)
Stabilizing energy	−12	*U*	1.827	1.72 (0.04)	−12	*U*	1.796	1.704 (0.04)
Stacking energy	−17	*U*	−3.313	−3.62 (0.12)	−14	*U*	−3.33	−3.65 (0.13)
Bending stiffness	−14	*D*	22.253	26.88 (2.11)	−14	*D*	22.95	27.44 (0.09)
Protein-induced deformability	−12	*D*	1.948	2.18 (0.09)	−12	*D*	1.949	2.186 (0.09)

The Trend column refers to whether this is a maxima (*U*) or minima (*D*).
*denotes that value of a given property (extrema) lie between mean +/- 2 Standard Deviation.

tested on an independent dataset containing the same number of negative and positive instances and compared with SVM-LIGHT (42).

Training, testing and optimization were done for each element and each chromosome separately. Parameters were obtained for each feature, element and genomic sequence (organized in a hierarchical manner as catalogued at http://nldsps.jnu.ac.in/elan.html). For instance, for a given rule (say the A rule) the following important attributes were identified:

(i) The extremum position (l) i.e. −13 bp (Table 1).
(ii) The extremum value i.e. 0.503 (50% A density).
(iii) The nature of the extremum, namely is it a minimum or maximum.
(iv) The presence of additional peak in the flanks (± 5 bp).

These values are derived from training dataset for each element on a given chromosome and optimized using ROC curves via the sensitivity and specificity values (Supplementary Figure S2).

Consider a curated set of n examples of known insertion sites and non-insertion sites and d features (or rules) on which to base this classification. Associated with each insertion site i_k is a d-dimensional vector. The SVM algorithm was trained using this dataset and the resultant model was tested with an independent dataset comprising both positive and negative examples. Given training sets there is an optimal hyperplane which separates the training data into two classes and this can be determined via standard methods (Supplementary Figure S3). Training and testing examples files are given in the Supplementary Data and on http://nldsps.jnu.ac.in/pages/isf.html.

Performance of the different rules

Sensitivity is the degree to which true examples are correctly detected whereas specificity is defined as the extent to which false instances are rejected successfully. The average of sensitivity and specificity is a measure of the overall accuracy of a specific combination of parameters used in ISF.

The performance of ISF for predicting insertion sites is evaluated chromosome-wise and the accuracy of SVMs for Alu elements in the human genome is given in Supplementary Table S1. As can be seen, this can be quite high, reaching 73% accuracy for chromosome 22. Even higher accuracy is achieved in *E. histolytica*, 80% or more for some elements. Comparable results were obtained with elements in other genomes and L1 in *H. sapiens*. In the initial stage of application of ISF, cutoff values for each parameter was calculated to be highest local maxima or minima observed at specific position in plots for each rule. These cutoffs were optimized by maximizing the overall accuracy as shown in Supplementary Figure S2.

It should be pointed out that individual rules did not perform as well as their combinations did (Supplementary Table S2) but as discussed above, this did not depend on which SVM kernels were used. The accuracy remained at the 66% level for any given rule, but combination rules worked better. A worked example is given online at the ELAN website, using the Alu element in Chromosome 22 of the human genome in the form of tutorial/screen shots along with a sample dataset (see Supplementary Data, tutorial.html).

RESULTS

In order to assess the applicability of the tool in different contexts, ELAN has been applied to over 50 genomes, including *E. histolytica*, *Caenorhabditis elegans*, *Canis lupus familiaris*, *Macaca mulata*, *M. musculus* and *H. sapiens*. Results obtained from these studies are available through the InSiDe database, as well as on the web site http://nldsps.jnu.ac.in/elan.html. Two examples are discussed here in detail, an analysis of Alu and L1 in the human genome, both well studied MGEs.

Our analysis detects ∼1.2 million copies (Supplementary Table S3) of Alus in the human genome, similar to previously reported values (43). The chromosome-wise distribution of Alu copies is shown in Supplementary Table S3, the number of Alus is roughly proportional to the chromosome length (Figure 4). There are few functional Alu copies in the human genome, most being truncated at either the 5′ or 3′ end (or both). Their distributions on the various chromosomes follow similar patterns.

As discussed in the 'Materials and Methods' section, we construct pre-insertion loci and classify them into the four groups (Supplementary Table S3): Intact on both ends, Intact on 5′, Intact on 3′ and Intact on neither end. These are then analyzed for different physicochemical properties using DNASCANNER (30).

Alu elements tend to preferentially insert in the vicinity of A-rich regions (Figure 3), where thermodynamic, structural and nucleosomal positioning parameters are also markedly different as compared to genomic average values. These deviations were found to be statistically significant: the Mann Whitney P-values were typically below 0.05 (see 'Materials and Methods' section). These patterns are observed in all chromosomes, with the largest deviation being between −12 and −15 bp from the insertion site (Figure 5). Signals are mainly seen in the 5′ intact class (with 3′ either intact or truncated) while 5′ truncated elements show signals only for some of the parameters, the location and level being inconsistent for different chromosomes (data not shown).

Each of the signals is converted into a graphical profile that typically shows the relevant quantity peaking in the vicinity of the insertion site. To validate that a given profile is significant, we compute the average and standard deviation for each parameter profile and for each class of Alu. The local extremum, that is, maximum or minimum is considered significant only if it exceeds the mean ± 2 SD (Table 1). An important indicator of the existence of a pattern is that it should not occur where Alus do not insert, and in our work we construct specific negative datasets for this purpose. Further validation is provided by randomly removing one third of the positive dataset repeatedly. The pattern persists, suggesting that the signals

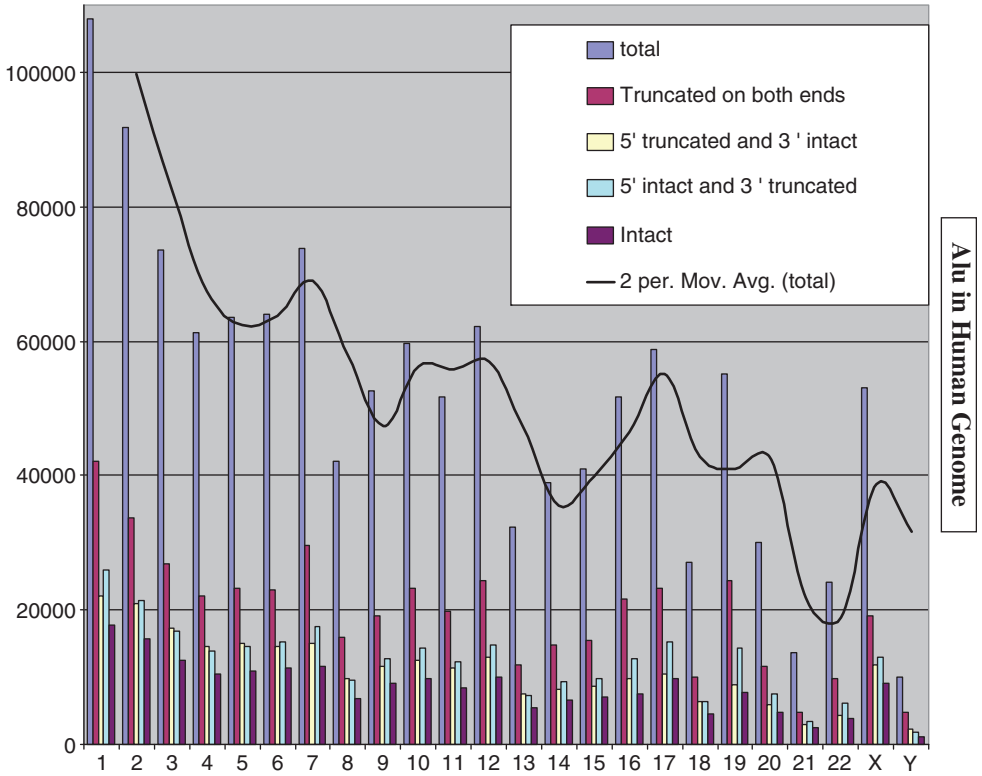


Figure 4. Distribution of Alu element across human genome. Four classes of elements (see 'Materials and Methods' section) are indicated with four different colors. The *y*-axis represents the frequency of elements found on the different chromosomes (marked along the *x*-axis).

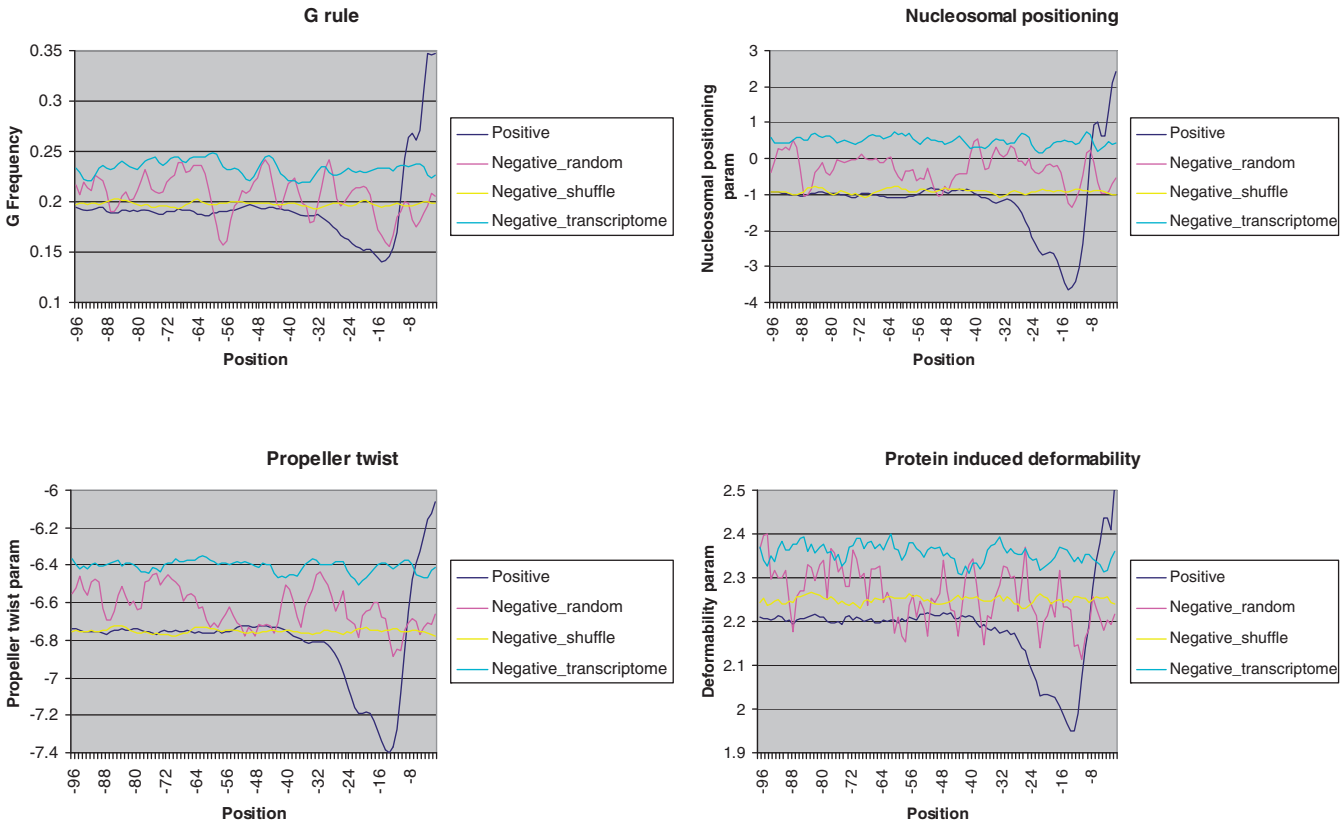


Figure 5. Various signals upstream of the insertion sites of Alu in chromosome 2. The *y* axis represents value of the property and the *x*-axis gives the relative position with respect to the insertion site (taken to be 0) (Figure 3).

are common to all members of the set. Similar signals are observed in flanking regions of Alu present in genes (Supplementary Figure S4).

The biological relevance of signals in the vicinity of insertion sites derives from the fact that some of these are structural in nature, and some pertain to the energetics. Those that we employ here include:

(i) DNA bendability which measures the mechanical flexibility around insertion sites. We find a region of low bendability (from −12 to −15 bp) followed by sharp increase in the flexibility upstream of Alu insertion sites (Supplementary Figure 5A).
(ii) Propeller twist, which also measures the flexibility of DNA via its tendency to twist about the long axis that makes the two bases of a pair non-coplanar (20). Dinucleotides having a large propeller twist tend to be more rigid than those with low-propeller twist. Since proteins encoded during TPRT seem to distort DNA, sites with specific dinucleotides that are more rigid and are followed by a lower propeller twist appear to be amenable for insertion (Figure 5).
(iii) Bending stiffness and nucleosomal positioning: nucleosomes play an important role in DNA compacting as well as in providing transcription factors access to regulatory regions. Since this is essential for activation of gene expression (44), we examined two different nucleosomal related features, the bending energy/persistence length (36) and the nucleosomal positioning profiles (36). We find that insertion sites have a low-energy region between positions −31 and −11 with significant minimum at position −14 (Supplementary Figure S5). Similar results were obtained using bending-energy profile (Supplementary Figure S5B).
(iv) Stacking energy profiles show a peak near suitable insertion sites, indicating that high energy regions are intrinsically unstable, leading to easy de-stacking or melting of DNA sequence that enables Alu insertion (Supplementary Figure S5C).
(v) Duplex stability is a measure of the relative stability of the DNA-duplex structure and is dependent on sequence (23), and more specifically upon 10 different types of nearest neighbor interactions namely AA/TT; AT/TA; CA/GT; GT/CA; CT/GA; GA/CT; CG/GC; GC/CG; GG/CC. Using this, we find that the region around the −13 position is destabilized more easily compared to controls (Supplementary Figure S5D) in suitable insertion sites.
(vi) The DNA-denaturation-energy profile (Supplementary Figure S5E) indicates that a relatively small amount of energy is required to melt DNA near insertion sites favoring retrotransposition (37 Kcal/mol at −13 bp).
(vii) Protein interaction signals, assessed via the deformability of the DNA show that a region of low deformability, followed by one of high deformability (Figure 5) facilitates retrotransposon insertion.

About 100 000 copies of L1 are found uniformly distributed in each chromosome of the human genome. Most copies (73%) are truncated on both ends, 21 852 (20%) are truncated on the 5′-end and 1938 copies (1.8%) are truncated on the 3′-end. Thus intact copies of L1 number 3663 (fewer than 4%) indicative of the extreme mutability of these elements in the face of very weak selection pressure (Table 2). The pre-insertion loci of L1 also share similar signals with Alu loci (Supplementary Figure S6).

Analysis of *E. histolytica* genome

In *E. histolytica*, an early branching unicellular eukaryote that occupies a unique evolutionary position, MGEs occupy only 11% of genome. The pre-insertion loci of EhLINE1, EhLINE2, EhSINE1 and EhSINE2 were constructed as described above. All elements show similar insertion preferences, for example, T richness at the site of insertion in contrast to A-rich regions in the human genome (for both Alu and L1 elements). A number of signals such as the propeller twist, stacking energy, bendabilty profile, free-energy profile, DNA-denaturation energy, protein-induced deformabilty and nucleosomal-related features are present in the 5′ upstream region. They are absent in negative datasets constructed specifically for the *E. histolytica* genome, namely scrambled positive dataset, sites picked from vicinity of pre-insertion loci, sites picked from prokaryotic genomes with comparable AT richness, and sites picked from within genes where elements are known to not insert.

Table 2. L1 element distribution on the different chromosomes on the human genome

Chromosome number	Total elements	Truncated on both ends	5′ Truncated and 3′ intact	5′ Intact and 3′ truncated	Intact on both ends
1	7377	5432	1557	166	222
2	8451	6239	1771	155	286
3	7766	5665	1666	154	281
4	8336	6005	1819	170	342
5	7434	5355	1650	138	291
6	6641	4849	1441	126	225
7	5406	4045	1064	103	194
8	4153	2996	886	86	185
9	4288	3188	890	68	142
10	4238	3125	891	75	147
11	4971	3575	1110	96	190
12	4468	3238	963	84	183
13	3658	2735	785	49	89
14	3188	2335	684	50	119
15	2565	1953	474	42	96
16	1654	1241	321	33	59
17	1417	1104	263	20	30
18	2544	1926	516	39	63
19	1005	787	154	25	39
20	1440	1080	296	14	50
21	1180	943	209	7	21
22	631	526	82	7	16
X	10 221	7751	1981	173	316
Y	1509	995	379	58	77
Total	104 541	77 088	21 852	1938	3663

Insertion sites of MGE in other genomes also show similar signals

We analyze genomes of other organisms for the presence of signals in regions upstream of their respective insertion elements.

In *M. musculus* genome, preinsertion loci of the B1 element were investigated (Supplementary Figure S7) and it was observed that almost all the physicochemical parameters discussed above show significant extrema between −12- and −15-bp upstream of the insertion site of intact (or full length) elements. Alu insertion sites in *M. mulata* (Supplementary Figure S8) also have same characteristics (see Supplementary Data, Master_Supplementary_file.doc).

Application of DNASCANNER and ELAN on the human genome

Some human diseases are linked to genes known to be susceptible to retrotransposon insertions (45). A plausible hypothesis is that within the coding region of such genes are signals that may be recognized as insertion sites. We applied DNASCANNER to a number of such genes. In order to investigate the optimal insertion environment within the coding region, a number of representative sequences were selected. Known sites of insertion and a flanking region of length 1000 bp were extracted from genome databases. Controls were selected in the same gene in the vicinity of insertion site as well as randomly picked regions from the same gene. We conducted detailed analyses of genes reported to be susceptible to TE insertion leading to disease (46,47), including DMD (48), CYBB (49) and APC (50) for L1, Alu and SVA insertions. Details of the identified site and orientation of insertion are given in Supplementary Table final-disease-gene-table .xls. Pre-insertion loci were constructed by removing the TE as well as the TSD. In most cases, significant signals were observed in the flanks of disease genes although these are not seen at Alu and L1 insertion sites (see Figure 6 and Supplementary Figures at http://nldsps.jnu.ac.in/TE-Vs-diseases/database/). These were classified as typical,

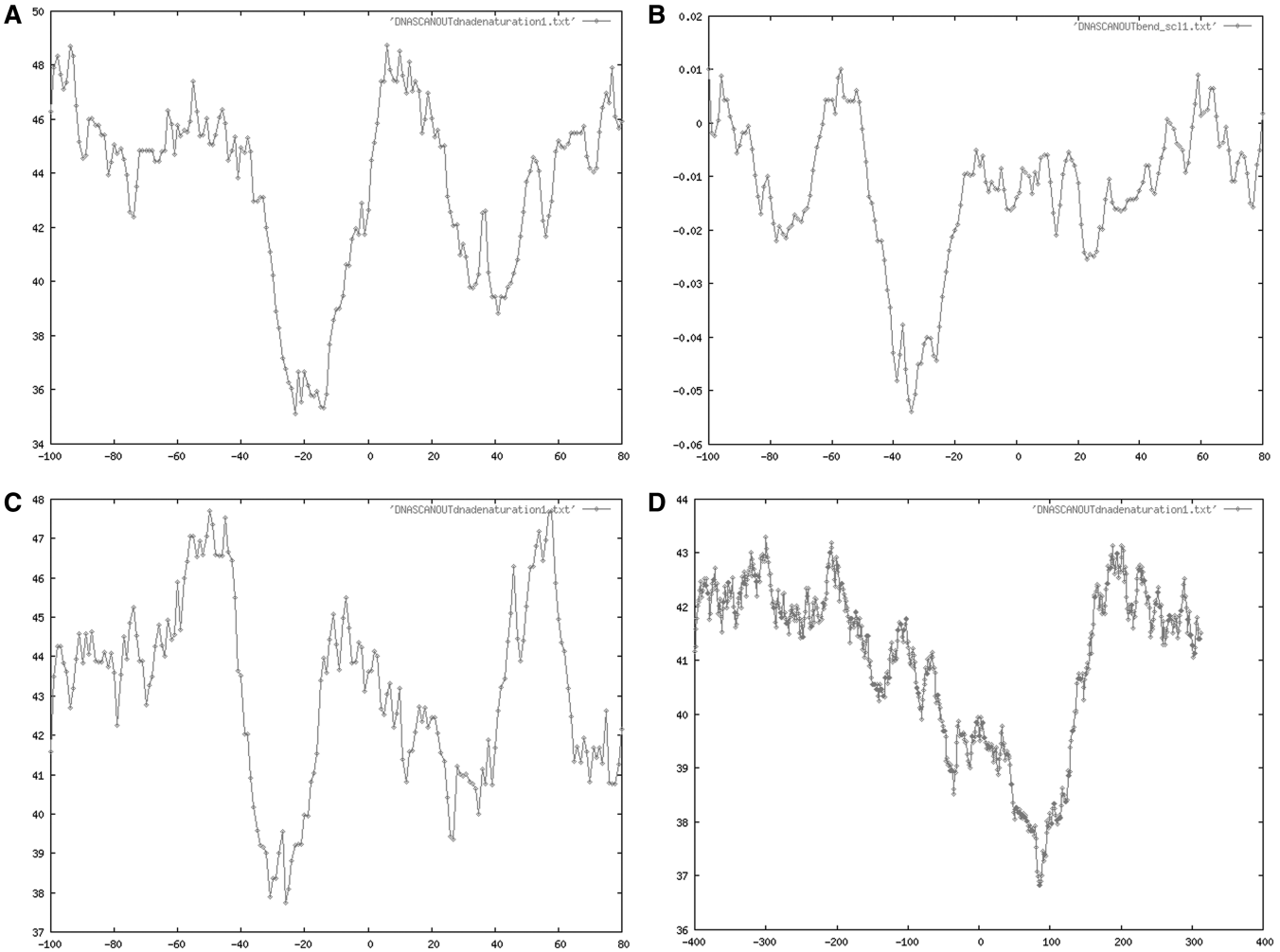


Figure 6. (**A**) DNA-denaturation profile of pre-insertion loci of ABCD1 gene reported to be disrupted by Alu element at position 0. (**B**) Bendability profile of preinsertion loci of Dystrophin gene reported to be disrupted by L1 element at position 0. (**C**) DNA-denaturation profile of pre-insertion loci of Spectrin gene reported to be disrupted by SVA element at position 0. (**D**) DNA-denaturation profile of pre-insertion loci of APC gene reported to be disrupted by Alu element at position 0. In this example the 1000 bp of sequence flanking insertion site of L1 element.

atypical or negative: typical signals were those similar to Alu and L1 insertion sites, a single statistically significant extremum 10–20-bp upstream the insertion site. Examples included gene such as SERPING1. (Supplementary Data at final-disease-gene-table.xls) The atypical had signals outside from the −10- to −20-bp region or had two extrema. The negative class did not show any significant signal. Of the 49 genes studied, 11 genes were typical, in this sense 27 genes were atypical and 11 showed no signals (see http://nldsps.jnu.ac.in/elan.html). A representative atypical case is that of DMD, one of the largest genes known, of length 2.4 Mb. An instance of L1 insertion has been reported to cause disruption in the gene at position 128 269 leading to Duchene Myotrophy disease (51,52). Flanking regions ranging from 100 bp to 5 Kb were analyzed which revealed the presence of insertion signals in the vicinity. This and other examples can be seen online at http://nldsps.jnu.ac.in/TE-Vs-diseases/database/ as well as in Supplementary Data (final-disease-gene-table.xls). Based upon our limited study of 49 genes, we assume that the absence of typical signals within exonic regions prevents wide-spread disruption of coding regions by Alu and L1s. Several copies of L1 and Alu are already present in the intronic regions of these genes; this suggests that our understanding of the role of these elements in causation of diseases by gene disruption is at present incomplete and further analysis is necessary

To investigate the role of insertion sites or signals in gene regulation, we applied DNA SCANNER to a promoter dataset derived from eukaryotic promoter database (EPD) (53). It appears that structural features of sequences within 500 bp (upstream or downstream) of the TSS of several eukaryotic genes also have characteristic signals (see Supplementary Data at the program web site) although there are significant differences in the parameters in the upstream and downstream regions. Signal extrema are mostly seen in vicinity of TSS in a variety of examples from *Xenopus laevis*, *Gallus gallus*, *M. musculus*, *Rattus norvegicus*, *Bos taurus*, as well as several plants promoters (Figure 7). Similar signals have also been seen at splice sites of *Drosophila* (Supplementary Figure S9) as well as other organisms.

Element detection within ELAN

In the present work, we use the module ELEFINDER within ELAN to detect copies of known MGEs such as Alu or L1. This was benchmarked against the standard, RepeatMasker (A.F.A. Smit *et al.*, unpublished data, www.repeatmasker.org) and also compared to other existing tools.

Our algorithm detects 99.5% of the Alu insertions in human chromosome 22 (HC22) that are located by Repeatmasker run with the following parameters: RepeatMasker -alu -s -species human -no_is hs-chr22.fa. (see Supplementary Data RM-ELAN-Ch22.xls at web site). ELAN detects 24 135 Alu copies, whereas Repeatmasker detects 24 248; the difference of 113 (0.5%) could be due to merging step of ELAN. A third program CENSOR (54) uses the Smith–Waterman nucleotide alignment to detect repeats and outputs masked genomic DNA along with a tabular summary of TE content. The online version of CENSOR (http://www.girinst.org/censor/index.php) with standard parameters found 24 237 copies of Alu, slightly more than either Repeatmasker or ELAN. A full report is given in the Supplementary Data at the web site.

ISF

The graphical results obtained from DNA SCANNER were made quantitative in the following manner. For each of the different structural or energetic features that were assessed in the vicinity of insertion sites, the location of the extremum (in base pair), its nature [minimum (D) or maximum (U)], the value at the extremum, the average and standard deviation (in order to assess the background). An example is described in Table 1 which summarizes results for Alu and L1 in human chromosome 2 (HC2).

Consider the A density. This has a maximum ($U = 0.503$) at 13-bp upstream of Alu insertion sites, with average and standard deviation of the overall profile being 0.34 and 0.05, respectively. The peak is thus significant and is included in the training. Any parameter whose values were judged to be non-significant was excluded from final classification model. A PERL script was used for automation of this procedure (http://nldsps.jnu.ac.in/dna2isf.html or tutorial.html). All Alu-insertion sites whose A-density profiles show a maximum at −13 bp were considered to be positive for A rule, and the overall sensitivity, namely the fraction of insertion sites detected by ISF was found to be 0.64. This quantity was computed for each property (see Supplementary Table S4).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}.$$

The specificity is the fraction of non-insertion sites rejected by ISF for a given rule, and as above, the results are summarized in Supplementary Table S4 for Alu insertion in human chromosome 2. For most properties these indicators lie between 0.6 and 0.7 indicating that individually they provide some level of discrimination between real insertion sites and sites that are not suitable for insertion.

Cutoff values for each parameter were calculated as the highest extremum observed at specific positions. Sensitivity and specificity at several values for a given parameter (Supplementary Figure S2) were computed and optimized (Supplementary Table S5): this removes the imbalance between sensitivity and specificity values and improves overall performance. After optimizing each parameter, these new cutoff values are used for classification. The dataset was divided equally into training and test sets (Supplemental Data, trainalu.txt and testalu.txt) and for each a property vector was constructed, basically as a column of 1's and 0's, each entry corresponding to whether a each property scored above the threshold value (1) or did not (0) (see Supplementary Data, tutorial.html).

These vectors were then used as training data for a SVM in the standard manner; see file trainalu.txt. Testing was done on an independent dataset (file testalu.txt)

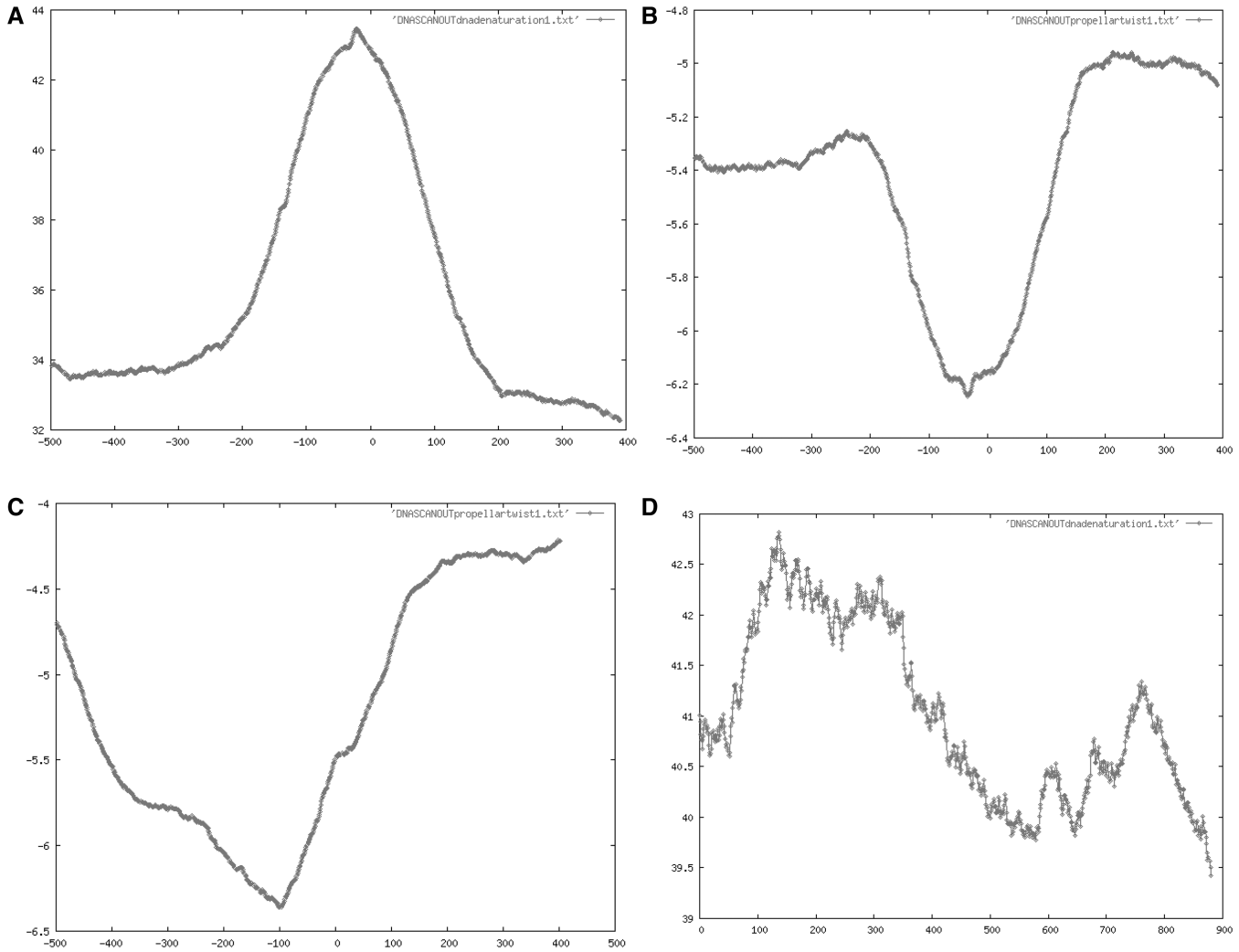


Figure 7. (**A**) The DNA-denaturation profile of DNA sequences from EPD comprising *B. taurus* promoters (−500 bp) and genes (400 bp). The +1 represent TSS of the gene. The window size was 100 bp of total length. (**B**) Propeller twist profile of same dataset. (**C**) Propeller twist profile of promoters of *Xenopus*. (**D**) DNA-denaturation profile of viral genes.

comprising of an equivalent number of positive and negative examples. We evaluated each of the parameters for a given chromosome separately to assess the performance of ISF and results for classifiers for each of the human chromosomes for Alu insertion sites are given in Supplementary Table S1. The highest accuracy was obtained from polynomial kernel in the SVM reaching 90% for human chromosome X (Supplementary Table S1), and over 80% in *E. histolytica*. Comparable results were obtained with other elements and other genomes as well.

We also investigated the effect of individual parameters and kernels on the accuracy of the SVM and the data is given in Supplementary Table S2. Although the individual rules performed poorly when used alone, the overall performance increased significantly when used in combination. Rules were then deleted systematically from the set to examine the effect on the performance under each kernel and to arrive at a minimal set of rules that gives maximal predictive accuracy (Supplementary Data).

DISCUSSION

MGEs have been termed drivers of evolution (55,56) in that they effectively expand the genome by insertion, choose optimal insertion locations, create longer intergenic and non-coding sequences—possibly also introns—and thereby promote genomic diversity. Intergenic sequences presumably have fewer mutational constraints than coding sequences, and the lack of selection pressure appears to offer more versatility in evolution.

The identification of locations on the genome where such elements can successfully insert is thus important, and in this article, we describe a set of bioinformatics tools for the analysis and prediction of putative MGE insertion sites. Such locations are an important factor for spreading of MGEs as well as genome expansion in general, and the tool that we have developed here, ELAN uses a combination of methods ranging from sequence comparison to pattern recognition, machine learning, and classification. The identification of elements, their

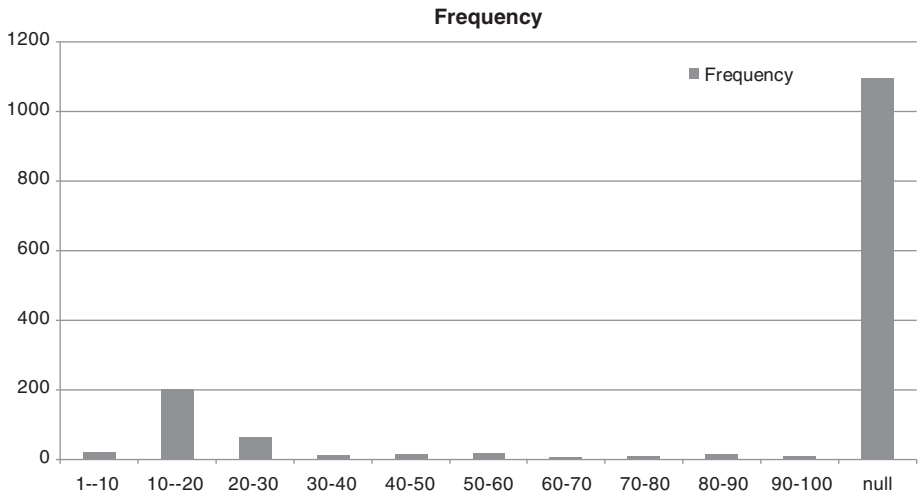
distribution and their relationship with specific genes can be analyzed in detail, and intergenomic comparisons are possible as well.

During insertion, a non-LTR element causes the target site to distort in a number of ways and requires the co-operative action of a number of proteins such as reverse transcriptase and endonuclease to break bonds, unwind the DNA, and most importantly, to nick the target site strand (14). Insertion sites therefore show significant structural patterns owing to the need for a combination of rigid and flexible regions. The potential for a given genomic sequence to have the appropriate physical properties is evaluated by two components of ELAN, the programs ISF and DNASCANNER.

The specificity and sensitivity of the technique depends on the element being analyzed and its context. Elements have a highly non-uniform distribution for instance there are ~7000 copies of L1 on the largest human chromosome I, but the X chromosome has over 10 000 copies. ELAN successfully predicts over 90% of the insertion sites of Alu using ISF on the X chromosome (Supplementary Data) whereas on the Y chromosome the prediction rate reduces to 60%. Since the present computational tool identifies insertion sites mainly through the recognition of 'typical' patterns, it would appear that the insertion sites on the X chromosome bear more fidelity to the standard or ideal insertion site. The lower predictive accuracy of ISF on the Y chromosome reflects the fact that the insertion sites may themselves have evolved (57,58). Similarly, in analyzing different age classes of Alu, namely the Y, S and J, younger elements such as Alu Y show stronger signals compared to older or truncated elements since the probability of mutation of both a given element as well its flanking sequences increases with evolutionary time.

The sequence environment of the insertion site has fairly specific motifs that depend on the element in question. The local sequence between 10- and 20-bp upstream of the site is likely the most important (although not the deciding) feature that enables insertion. Even locally, there can be considerable flexibility in sequence: studies of L1- and Alu-insertion sites report both canonical (TT/AAAA) and non-canonical (TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, TCAAGA, AAAAAA) insertion motifs for L1 and Alu (59–63). While the canonical motif, TT/AAAA occurs most frequently in the upstream region (see Figure 8, and Supplementary Data for the case of Chromosomes 22 and Y) the other motifs are also present in fair measure. Indeed the overwhelming majority of TTAAAA motifs that occur on the genome are not associated with the Alu insertion sites. At the same time, there is considerable evidence that the 10–20-bp upstream region plays a major role in the retro-transposition mechanism, and the TTAAAA motif is indeed frequently present in this region (59). Non-canonical motifs, when linked to elements, are also found in the same upstream region where the various physicochemical properties we have studied (see 'Materials and Methods' section) show maximal variation. The TT/AAAA motif is present in 13% of Alu insertion sites examples studied in Human chromosome 22 in 10–20-bp upstream of Alu insertion sites. Our results corroborate well with earlier work (59) where authors showed presence of TT/AAAA at ~15–16-bp upstream of 400 examples of Alu insertion sites. A recent study suggest longer pattern containing canonical and non-canonical motifs around Alu insertion sites (64).

Features those present on larger scales and which influence element insertion are not easily identified. Alus insert in the immediate vicinity of A-rich regions (of size ~100 bp, Figure 3) although these regions are themselves within GC regions on even larger scales (of the order of megabites). Further, Alus are enriched in gene-rich regions while L1-populate-intergenic regions. An additional problem that is only partially addressed here is whether the current distribution of MGEs reflects their insertion dynamics, or is a consequence of the retention potential of different sites. Thus other features, in addition



Distance of Alu from TTAAAA	Frequency	%
1--10	22	1.49
10--20	**202**	**13.76**
20-30	65	4.42
30-40	13	0.88
40-50	14	0.95
50-60	19	1.29
60-70	6	0.40
70-80	9	0.61
80-90	14	0.95
90-100	8	0.54
null	**1096**	**74.65**
TOTAL	**1468**	**100**

Figure 8. Most instances of the canonical TTAAAA motifs are unrelated to known Alu insertion sites. Shown here is a histogram of distances of TTAAAA to the nearest Alu insertion site, and as can be seen, >70% are >100-bp away from the nearest Alus (see http://nldsps.jnu.ac.in/elan.html for more details).

to possibly long-ranged correlations are likely to affect the distribution of MGEs.

It is well-known that MGEs tend to invade genomes in pairs of LINEs/SINEs. Examples include L1/Alu in human genome, L1/B1 in mouse and EhLINEs/EhSINEs in *E. histolytica*. Such pairs of elements share insertion site preferences in addition to sharing local sequence similarity. Alu and B1 elements belong to the same family of SINEs and the signals observed at mouse (B1) and human (Alu) insertion sites are strikingly similar (Supplementary Figures S5 and S7). This raises the possibility that the tools developed here for non-LTRs could be applied to other MGEs as well and our preliminary studies have been encouraging. For instance, the Ty1 element in *Saccharomyces cerevisiae* integrates preferentially upstream of genes that are transcribed by RNA polymerase III (65). When these regions were analyzed, a number of the signals used here for non-LTRs (the A and G rules, bending stiffness, DNA-denaturation energy) showed significant signals (see Supplementary Figures and data http://nldsps.jnu.ac.in/Ty1). The other insertion sites examined were those for P elements in *Drosophila* (66), Tn7 system (67) and *Sleeping beauty* (68), which also showed some of the same signals, although the specific insertion mechanisms in these cases can be quite different from the non LTRs. This may interpreted as though the general mechanism behind the transposase-based mobilization of DNA transposons is completely different from the reverse transcriptase-based mechanisms of retro-transposon mobilization but the biophysical properties of the genomic DNA surrounding known insertion sites are similar among all insertion sites.

The present work is an ongoing effort to develop an integrated system for analysis and *de novo* detection of MGEs. The tools available here are generic, and in principle can be adapted for use in the detection of other features as well. All data generated in these studies, as well as from our ongoing analysis of MGEs in diverse genomes has been curated in the database InSiDe that incorporates information on the distributions, insertion sites and element sequences. This is available at http://nldsps.jnu.ac.in/inside.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Professors Sudha and Alok Bhattacharya for reviewing the article, as well as for comments and suggestions.

FUNDING

Department of Biotechnology (DBT), Government of India. Funding for open access charge: Center for Excellence Fund of School of Computational and Integrative Sciences, Jawaharlal Nehru University provided by Department of Biotechnology, Government of India.

Conflict of interest statement. None declared.

REFERENCES

1. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
2. Vallenet,D., Labarre,L., Rouy,Z., Barbe,V., Bocs,S., Cruveiller,S., Lajus,A., Pascal,G., Scarpelli,C. and Medigue,C. (2006) MaGe–A microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65.
3. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. *et al.* (2003) GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
4. Sakata,K., Nagamura,Y., Numa,H., Antonio,B.A., Nagasaki,H., Idonuma,A., Watanabe,W., Shimizu,Y., Horiuchi,I., Matsumoto,T. *et al.* (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res.*, **30**, 98–102.
5. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
6. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Ungerer,M.C., Strakosh,S.C. and Zhen,Y. (2006) Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Curr. Biol.*, **16**, R872–R873.
9. Peaston,A.E., Evsikov,A.V., Graber,J.H., de Vries,W.N., Holbrook,A.E., Solter,D. and Knowles,B.B. (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell*, **7**, 597–606.
10. Orgel,L.E. and Crick,F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.
11. Tighe,P.J., Stevens,S.E., Dempsey,S., Deist,F.L., Rieux-Laucat,F. and Edger,J.D. (2002) Inactivation of the *Fas* gene by Alu insertion: retrotransposition in an intron causing splicing variation and autoimmune lymphoproliferative syndrome. *Genes Immun.*, **3(Suppl 1)**, S66–S70.
12. Szak,S.T., Pickeral,O.K., Landsman,D. and Boeke,J.D. (2003) Identifying related L1 retrotransposons by analyzing 3′ transduced sequences. *Genome Biol.*, **4**, R30.
13. Deininger,P.L. and Batzer,M.A. (1999) Alu repeats and human disease. *Mol. Genet. Metab.*, **67**, 183–193.
14. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
15. Bergman,C.M. and Quesneville,H. (2007) Discovering and detecting transposable elements in genome sequences. *Brief. Bioinformatics*, **8**, 382–392.
16. Loftus,B., Anderson,I., Davies,R., Alsmark,U.C.M., Samuelson,J., Amedeo,P., Roncaglia,P., Berriman,M., Hirt,R.P., Mann,B.J. *et al.* (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature*, **433**, 865–868.
17. Lorenzi,H., Thiagarajan,M., Haas,B., Wortman,J., Hall,N. and Caler,E. (2008) Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba species*. *BMC Genomics*, **9**, 595.
18. Brukner,I., Sanchez,R., Suck,D. and Pongor,S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: Parameters for trinucleotides. *EMBO J.*, **18**, 1812–1818.
19. Dickerson,R.E. and Chiu,T.K. (1997) Helix bending as a factor in protein/DNA recognition. *Biopolymers*, **44**, 361–403.

20. Hassan,M.A.E. and Calladine,C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.*, **259**, 95–103.
21. Ornstein,R.L., Rein,R., Breen,D.L. and MacElroy,R. (1978) An Optimized potential function for calculation of nucleic-acid interaction energies I Base stacking. *Biopolymers*, **17**, 2341–2360.
22. Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, **24**, 4501–4505.
23. Breslauer,K.J., Frank,R., Bolcker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci.*, **83**, 3746–3750.
24. Blake,R.D. (1996) Denaturation of DNA. In Meyers,R.A. (ed.), *Encyclopedia of Molecular Biology and Molecular Medicine*. Wiley VCH, New York, pp. 2–19.
25. Kapitonov,VV., Pavlicek,A. and Jurka,J. (2006) *Anthology of Human Repetitive DNA, Encyclopedia of Molecular Cell Biology and Molecular Medicine*. The Humana Press, Totowa, New Jersey.
26. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The Bioperl Toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
27. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
28. Marini,J.C., Levene,S.D., Crothers,D.M. and Englund,P.T. (1982) Bent helical structure in kinetoplast DNA. *Proc. Natl Acad. Sci.*, **79**, 7664–7668.
29. Crothers,D.M., Haran,T.E. and Nadeau,J.G. (1990) Intrinsically Bent DNA. *J. Biol. Chem.*, **265**, 7093–7095.
30. Mandal,P.K., Rawal,K., Ramaswamy,R., Bhattacharya,A. and Bhattacharya,S. (2006) Identification of insertion hot spots for non-LTR retrotransposons: computational and biochemical application to *Entamoeba histolytica*. *Nucleic Acids Res.*, **34**, 5752–5763.
31. Ozoline,O.N., Deev,A.A. and Trifonov,E.N. (1999) DNA bendability-a novel feature in E. coli promoter recognition. *J. Biomol. Struct. Dyn.*, **16**, 825–831.
32. Delcourt,S.G. and Blake,R.D. (1991) Stacking energies in DNA. *J. Biol. Chem.*, **266**, 15160–15169.
33. Ollis,D.L. and White,S.W. (1987) Structural Basis of Protein-Nucleic Acid Interactions. *Chem. Rev.*, **87**, 981–995.
34. Shakked,Z. and Rabinovich,D. (1986) The effect of the base sequence on the fine structure of the DNA double helix. *Prog. Biophys. Mol. Biol.*, **47**, 159–195.
35. Kennard,O. and Hunter,W.N. (1989) Oligonucleotide structure: a decade of results from single crystal X-ray diffraction studies. *Q. Rev. Biophys.*, **22**, 327–379.
36. Sivolob,A.V. and Kharpunov,S.N. (1995) Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *J. Mol. Biol.*, **247**, 918–931.
37. Olson,W.K., Gorin,A.A., Lu,X., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
38. Stigler,S.M. (1983) Who Discovered Bayes' Theorem? *Am. Stat.*, **37**, 290–296.
39. Parhami,B. (1994) Voting algorithms. *IEEE Trans. Reliab.*, **43**, 617–629.
40. Freund,Y. and Schapire,R.E. (1999) Short introduction to boosting. *J. Jap. Soc. Artif. Intell.*, **14**, 771–780.
41. Burges,C.J.C. (1998) A Tutorial on support vector machines for pattern recognition. *Data Min. Knowledge Discov.*, **2**, 1–47.
42. Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods – Support Vector Learning*. MIT Press, Cambridge, USA, 1998.
43. Grover,D., Mukerji,M., Bhatnagar,P., Kannan,K. and Brahmachari,S.K. (2004) Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. *Bioinformatics*, **20**, 813–817.
44. Richmond,T.J. and Widom,J. (2000) Nucleosome and chromatin structure. In Workman,J.L. and Elgin,S.C. (eds), *Chromatin Structure and Gene Expression*, 2nd edn. Oxford University Press, Oxford, UK, pp. 1–23.
45. Muratani,K., Hada,T., Yamamoto,Y., Kaneko,T., Shigeto,Y., Ohue,T., Furuyama,J. and Higashino,K. (1991) Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proc. Natl Acad. Sci.*, **88**, 11315–11319.
46. Chen,J.M., Stenson,P.D., Cooper,D.N. and Ferec,C. (2005) A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum. Genet.*, **117**, 411–427.
47. Wang,J., Song,L., Grover,D., Azrak,S., Batzer,M.A. and Liang,P. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.*, **27**, 323–329.
48. Holmes,S.E., Dombroski,B.A., Krebs,C.M., Boehm,C.D. and Kazazian,H.H. (1994) A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat. Genet.*, **7**, 143–148.
49. Meischl,C., de Boer,M., Ahlin,A. and Roos,D. (2000) A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *Eur. J. Hum. Genet.*, **8**, 697–703.
50. Miki,Y., Nishisho,I., Horii,A., Miyoshi,Y., Utsunomiya,J., Kinzler,K.W., Vogelstein,B. and Nakamura,Y. (1992) Disruption of the *APC* Gene by a Retrotransposal Insertion of L1 Sequence in a Colon Cancer. *Cancer Res.*, **52**, 643–645.
51. Narita,N., Nishio,H., Kitoh,Y., Ishikawa,Y., Ishikawa,Y., Minami,R., Nakamura,H. and Matsuo,M. (1993) Insertion of a 5′ truncated L1 element into the 3′ end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of duchenne muscular dystrophy. *J. Clin. Invest.*, **91**, 1862–1867.
52. Koenig,M., Beggs,A.H., Moyer,M., Scherpf,S., Heindrich,Q.K., Bettecken,T.T., Meng,G., Mullerj,C.R., Lindlof,M., Kaariainent,H. *et al.* (1989) The molecular basis for Duchenne versus Becker muscular dystrophy: correlation of severity with type of deletion. *Am. J. Hum. Genet.*, **45**, 498–506.
53. Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.
54. Kohany,O., Gentles,A.J., Hankus,L. and Jurka,J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7**, 474.
55. Kazazian,H.H. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
56. Kidwell,M.G. and Lisch,D.R. (2000) Perspective: Transposable Elements, parasitic DNA, and Genome Evolution. *Evolution*, **55**, 1–24.
57. Skaltesky,H., Kawaguchi,T.K., Minx,P.J., Cordum,H.S., Hillier,L., Brown,L.G., Repping,S., Pyntikova,T., Ali,J., Bieri,T. *et al.* (2003) The male specific region of the Y chromosome is a mosaic of discrete sequence classes. *Nature*, **423**, 825–837.
58. Hughes,J.F., Skaletsky,H., Pyntikova,T., Graves,T.A., van Daalen,S.K.M., Minx,P.J., Fulton,R.S., McGrath,S.D., Locke,D.P., Friedman,C. *et al.* (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature*, **10**, 1038.
59. Jurka,J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci.*, **94**, 1872–1877.
60. Feng,Q., Moran,J.V., Kazazian,H.H. and Boeke,J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905–916.
61. Morrish,T.A., Gilbert,N., Myers,J.S., Vincent,B.J., Stamato,T.D., Taccioli,G.E., Batzer,M.A. and Moran,J.V. (2002) DNA repair

mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.*, **31**, 159–165.
62. Morrish,T.A., Garcia-Perez,J.L., Stamato,T.D., Taccioli,G.E., Sekiguchi,J. and Moran,J.V. (2007) Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature*, **446**, 208–212.
63. Sen,S.K., Huang,C.T., Han,K. and Batzer,M.A. (2007) Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res.*, **35**, 3741–3751.
64. Zhang,K., Fan,W., Deininger,P., Edwards,A., Xu,Z. and Zhu,D. (2009) Breaking the computational barrier: a divide-conquer and aggregate based approach for Alu insertion site characterization. *Int. J. Comput. Biol. Drug Des.*, **2**, 302–322.
65. Brady,T.L., Schimdt,C.L. and Voytas,D.F. (2008) Targeting integration of the Saccharomyces Ty5 retrotransposon. *Methods Mol. Biol.*, **435**, 153–163.
66. Liao,G., Rehm,E.J. and Rubin,G.M. (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **28**, 3347–3351.
67. Bainton,R.J., Kubo,K.M., Feng,J. and Craig,N.L. (1993) Tn7 transposition: Target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell*, **26**, 931–943.
68. Liu,G., Geurts,A.M., Yae,K., Srinivasan,A.R., Fahrenkrug,S.C., Largaespada,D.A., Takeda,J., Horie,K., Olson,W.K. and Hackett,P.B. (2005) Target-site preferences of Sleeping Beauty transposons. *J. Mol. Biol.*, **346**, 161–173.