Mobile Genetic Elements

ISSN: (Print) 2159-256X (Online) Journal homepage: https://www.tandfonline.com/loi/kmge20

Identification and characterization of MGEs and their insertion sites in the *gorilla* genome

Kamal Rawal, Sangey Dorji, Amit Kumar, Anwesha Ganguly & Ankit Singh Grewal

To cite this article: Kamal Rawal, Sangey Dorji, Amit Kumar, Anwesha Ganguly & Ankit Singh Grewal (2013) Identification and characterization of MGEs and their insertion sites in the *gorilla* genome, Mobile Genetic Elements, 3:4, e25675, DOI: 10.4161/mge.25675

To link to this article: https://doi.org/10.4161/mge.25675

Copyright © 2013 Landes Bioscience

View supplementary material

Published online: 10 Jul 2013.

Submit your article to this journal

Article views: 2087

View related articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=kmge20

RESEARCH PAPER

Mobile Genetic Elements 3:4, e25675; July/August 2013; © 2013 Landes Bioscience

Identification and characterization of MGEs and their insertion sites in the *gorilla* genome

Kamal Rawal,* Sangey Dorji, Amit Kumar, Anwesha Ganguly and Ankit Singh Grewal

Department of Biotechnology; Jaypee Institute of Information Technology; Noida, UP India

Keywords: mobile genetic elements, primates, LINEs, SINEs, *Alu*, L1, truncation points, physiochemical properties

Recently published gorilla genome has offered an opportunity to study human evolution through variety of approaches. Mobile genetic elements (MGEs) insert non randomly in genome through mechanisms such as retrotransposition and may cause gene inactivation, transduction, regulation of gene expression and genome expansion. Here we report that majority of gorilla genome is occupied with MGEs (> 36%) with presence of LTRs and Non-LTRs such as *Alus* and L1s. Other types of MGEs such as MIRs, retrovirus like elements ERVs and DNA transposons are also found using repeatmasker and ELAN pipeline. The distribution is similar to Humans and Macaca genome. Using DNA Scanner we also scanned preinsertion loci for number of different properties such as DNA denaturation, energy measures, potential for protein interactions and sequence based features. We also predicted preinsertion loci with > 70% accuracy using a machine learning tool called insertion site finder (ISF) based upon support vector machines.

Introduction

Gorillas (genus *Gorilla)* are the largest living primates and closest relatives of humans after chimpanzee. They are primarily herbivores, found in the African forests and important to study human origins. There are two species of gorilla- eastern lowland gorilla and western lowland gorilla. To date four surviving hominids- humans, orangutan, chimpanzee and gorilla have been sequenced. There is great deal of interest in understanding genomic differences among these organisms. Latest discoveries reveal the fact that primates like gorilla and human share common ancestry 5–8 million years ago.[1] In all three species (gorilla, human and chimpanzee), genes relating to sensory perception, hearing and brain development showed accelerated evolution and particularly so in humans and gorillas.

The mobile genetic elements were first discovered in maize plant and from that every newly sequenced genome is subjected to the discovery as well as study of these elements. Recent sequencing of gorilla genome has offered an opportunity to study distribution of these elements in single female western lowland gorilla, Kamilah (Gorilla gorilla gorilla). Previous studies in context of mobile genetic elements in gorilla has revealed the presence of *Alu*, L1, SVA, LTRs and HERV insertions in gorilla genome but the study is limited to analysis of few segments of genome.[1] Earlier studies have also reported that retrotransposons are the most abundant MGE in mammalian genomes which affects wide functional activities such as genome evolution, gene disruption and regulation.[2-4]

To date, no studies, related to the detailed analysis of presence of the mobile genetic elements in the gorilla genome, have been conducted. Here we present the detailed analysis of the occurrences of the various mobile genetic elements in the recently sequenced gorilla genome. The study reveals the presence of various categories of transposons and the retrotransposons within the genome and many interesting results have been obtained allowing the new insights and dimensions to the possibilities of study, within this genome.

Mobile genetic elements (MGEs) are fragments of DNA that can move around within the genome through retrotransposition.[2,3] The genomic hotspots are identified by DNA structure and, endonuclease (EN) nicking to that DNA sequence.[5] These insertion hot spots are characterized by presence of sequence motifs and unique patterns.[6] The present work involves the identification and study of the distribution of several MGE particularly *Alu* and L1 retrotransposons in the genome of *Gorilla gorilla* using repeatmasker and ELEFINDER.[2,3] Previously, we have used ELEFINDER to perform a genome wide analysis of the MGEs in human and macaca genome.[3] We also scanned the DNA for number of different properties such as potential for protein interactions, physicochemical properties and sequence based features using DNA SCANNER. We then used the results for computational testing of the pre-insertion loci in order to detect potential insertion sites using ISF.

*Correspondence to: Kamal Rawal; Email: kamal.rawal@jiit.ac.in
Submitted: 05/01/13; Revised: 07/08/13; Accepted: 07/09/13
Citation: Rawal K, Dorji S, Kumar A, Ganguly A, Grewal AS. Identification and characterization of MGEs and their insertion sites in the *gorilla* genome. Mobile Genetic Elements 2013; 3:e25675; http://dx.doi.org/10.4161/mge.25675

Table 1. Summary of transposable elements in gorilla genome

Chromosome no.	Alus	MIRs	LINE1	LINE2	L3/CR1	ERVL	ERVLMaLRs	ERV classI	ERVclassII	hAT-Charle	TcMar-Tigger	Unclassified
1	91871	42291	41364	26891	3063	6697	15574	7983	634	12946	6139	520
2a	38096	14943	22798	9258	1254	3280	8357	3955	226	7085	3040	186
2b	39589	15791	26554	10686	1365	4257	10258	4611	275	8033	3654	232
3	62483	29994	41408	18504	2248	6694	15493	7207	514	14484	5416	373
4	51541	22609	38952	16426	1736	7357	16305	7961	599	9508	5568	293
5	70191	27423	31568	15842	1781	4479	11247	5062	356	10542	4175	338
6	54639	20082	34897	14374	1860	5504	12690	6361	452	10123	4806	323
7	61046	17277	31824	11823	1594	4680	11114	6226	414	9072	4198	589
8	45944	20456	29423	13493	1541	4846	11889	5737	435	7747	3919	242
9	42510	20001	24463	11053	1359	3539	8655	4025	303	7403	2904	225
10	51595	19025	27478	10958	1365	3819	10141	4913	340	8233	3591	258
11	43907	25484	25990	14785	1755	3915	9339	4390	334	7388	3282	265
12	53875	21217	25459	13997	1518	4089	10637	4804	331	8414	3459	266
13	27254	9549	20480	7705	926	3359	8027	4338	225	4959	2863	147
14	33196	12630	17438	7980	953	2928	7151	3302	231	5199	2318	185
15	33979	12510	16597	7128	980	2237	5182	2505	172	5494	2201	137
16	44337	15098	13939	7880	672	2332	6541	2677	197	6227	1531	140
17	31725	9704	18919	6994	870	3340	8203	3540	250	5521	2563	183
18	23161	8703	15599	5954	942	2465	5825	2909	142	4306	2097	129
19	47025	6755	8552	4802	153	1540	2963	2744	466	2526	881	168
20	25569	11830	11180	6897	511	2253	5114	1802	82	5287	1369	154
21	11510	2974	6822	2113	228	1396	3789	1586	72	1875	859	36
22	20426	7827	5319	4023	343	837	2214	1044	101	1797	687	76
X	42894	17909	45679	12042	1651	5092	12193	6750	459	9041	3650	233

A pictorial representation can be viewed in the **Figure 1**, which clearly shows the differential distribution of the various MGEs in the genome.

Results

Elements discovery

The ~3919 Mbp (2917687013 bps) *gorilla* genome sequence was screened for transposable elements with repeatmasker software revealing 3025664 elements in the genome. These elements accounted for 36.96% of the gorilla genome.

The TEs that were identified included the major TE classes: long terminal repeat (LTR) retrotransposons, non-LTR retrotransposons and DNA transposons. The **Table 1** shows the chromosome wise distribution of different kinds of MGEs present in the gorilla genome. NonLTR transposons were the most abundant TEs in the gorilla genome, and included diverse super-families such as SINEs and LINEs. While SINEs consisted of two super-families namely, *Alus* and MIRs, three different kinds of LINE elements discovered in the genome were, such as LINE1, LINE2 and LINE3/CR1. Among the LINEs, the major part is covered by the L1s and that in the case of SINEs, the *Alus* were found to be in numbers.

LTR retrotransposons were the second most abundant TEs in the gorilla genome, and the majority of these belonged to a class of mammalian repeats derived from retrovirus like elements and it was categorized into 4 subgroups namely, ERVL, ERVL-MaLRs, ERV_class I and ERV_class II. DNA transposons are rare in the gorilla genome, and are represented by only two super-families (hAT-Charlie and TcMar-Tigger). Some of the repeat elements, that could not be classified in the above mentioned families, were also identified. However, they were quite less in number.

Some other repeat elements like, satellites, small RNAs, simple repeats and low complexity regions were also identified (see **Table 1**).

Analysis of the LINEs and SINEs

In accordance to the length of each chromosome, the analysis of the total LINEs, SINEs and LTR elements was also done. **Tables 2, 3 and 4** summarize the complete data set. The total number of SINE elements was found to be 1,464,694, LINE elements were 8, 80,335 in number and the total count of LTRs was 4,33,996 in the entire genome. The average length of the inserted element was also calculated in each case along with the length occupied within the corresponding chromosome. The average length for the SINE elements was found to be within the approximate range of 208 to 220 bp and that of LINE element was approximately 408 to 658 bp.

Genome wide coverage

Table 5 shows the total percentage of the genome sequence as covered by various MGEs. It was observed that although the total count of *Alu* within the genome is highest, yet, the maximum percentage of the genome length was covered by the LINE 1

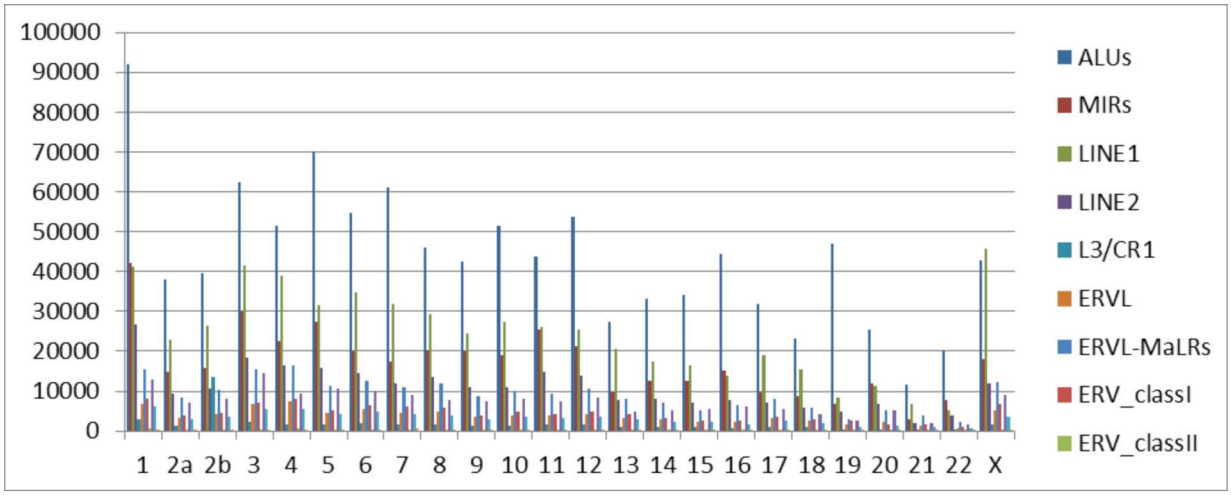


Figure 1. The chromosome wise representation of the distribution of MGEs in the gorilla genome.

Table 2. Chromosome wise summary of the SINE elements in gorilla genome

Chromosome no.	Total length (bp)	SINEs		
		Length occupied (bp)	No. of elements	Average length of element
1	229507203	28090071	134557	208.75
2a	1113551968	11251200	53188	211.53
2b	131632457	11899664	55606	213.99
3	199944510	19414804	92794	209.22
4	201139530	15782085	74411	212.09
5	165930986	20810185	97878	212.61
6	171703152	16075800	74977	214.40
7	158137892	16975146	78557	216.08
8	145327772	14118592	66630	211.89
9	121947112	13107115	62655	209.19
10	147764049	15126370	70834	213.54
11	133470886	14310615	69572	205.69
12	133360231	15996265	75292	212.45
13	97499607	7992543	36939	216.37
14	88974843	9721051	45987	211.38
15	82026568	9885170	46623	212.02
16	80971650	12725256	59528	213.76
17	94257108	9016411	41540	217.05
18	78787515	6820885	31990	213.21
19	56181278	11866327	53843	220.38
20	62603092	7807243	37473	208.34
21	35451371	3201567	14543	220.14
22	35671106	5969251	28278	211.09
X	154045127	13000221	60999	213.12

Table 3. Chromosome wise summary of the LINE elements in gorilla genome

Chromosome no.	Total length (bp)	Length occupied	No. of elements	Average length of element
1	229507203	35719713	71763	497.74
2a	1113551968	17825260	33525	531.70
2b	131632457	21599024	38868	555.70
3	199944510	33940485	62566	542.47
4	201139530	33070071	57467	575.46
5	165930986	24894380	49469	503.23
6	171703152	28474055	51529	552.58
7	158137892	24387235	45617	534.60
8	145327772	23886020	44708	534.26
9	121947112	18867313	37128	508.16
10	147764049	20895555	40065	521.54
11	133470886	21911049	42747	512.57
12	133360231	20811356	41225	504.82
13	97499607	15886349	29288	542.41
14	88974843	13905388	26549	523.76
15	82026568	12722443	24879	511.37
16	80971650	9370538	22615	414.35
17	94257108	14593897	26941	541.69
18	78787515	12068676	22639	533.09
19	56181278	5520135	13525	408.14
20	62603092	8267354	18683	442.50
21	35451371	4834686	9215	524.65
22	35671106	4145031	9728	426.09
X	154045127	39142498	59596	656.79

Table 4. Chromosome wise summary of the LTR elements in gorilla genome

Chromosome no.	Total length (bp)	LTR elements		
		Length occupied	No. of elements	Average length of element
1	229507203	16331504	31734	514.63
2a	1113551968	7938316	16285	487.46
2b	131632457	9873134	19928	495.44
3	199944510	15861956	30683	516.96
4	201139530	17614113	32882	535.67
5	165930986	10618919	21725	488.78
6	171703152	13456407	25647	524.67
7	158137892	11400774	22995	495.79
8	145327772	11703956	23430	499.52
9	121947112	8137837	16944	480.27
10	147764049	9403542	19649	478.57
11	133470886	9697311	18429	526.19
12	133360231	10222587	20364	501.99
13	97499607	8318528	16279	510.99
14	88974843	7087817	13954	507.94
15	82026568	4878339	10401	469.02
16	80971650	4955225	11971	413.93
17	94257108	7798515	15689	497.06
18	78787515	5736835	11595	494.76
19	56181278	3859077	7768	496.79
20	62603092	3862835	9450	408.76
21	35451371	3322579	6960	477.38
22	35671106	1814702	4256	426.38
X	154045127	14856258	24978	594.77

element. The total percentage of the genome as covered by *Alu* was 8.5%, but L1 covers 13.3% of the total gorilla genome. This shows that within the total area of the genome, as covered by the MGEs, the L1s form the major part. Among the non LTR elements, the L3/CR1 element was present in least numbers, and the MIRs and LINE2 covered approximately the same amount of region on the genome (~2%). The LTR elements covered ~7.4% of the total genome, out of which ERVL-MaLRs were the major constituents. We compared MGE content between chromososme no.1 and X chromosome of gorilla genome (see **Fig. 2**). We found that there are more number of L1 on X chromosome despite being smaller than chromosome 1. On the other hand *Alu* appear to be distributed more randomly i.e., number of Alus on chromosome are roughly proportionate with size of the chromosomes.

DNA SCANNER and ISF

We present results generated by DNA SCANNER on *Alu* elements insertion sites (chromosome 22) as a representative case (see **Table 6** and **Fig. 3**). In ISF module, we trained and tested using insertion site sequences of chromosome 21 and 22. The

Table 5. Summary of MGEs in gorilla genome

TE superfamily	Counts (copy no.)	Length (bp)	% of sequence covered
Non-LTR			
Alus	1048363	248964770	8.53
MIRs	412082	61542196	2.11
LINE1	582702	387476383	13.28
LINE2	261608	71320928	2.44
L3/CR1	42953	6666746	0.22
LTR elements			
ERVL	90935	44219203	1.51
ERVL-MaLRs	218901	93556837	3.20
ERV_classI	106432	71144203	2.43
ERV_classII	7610	7508372	0.26
DNA elements			
hAT-Charlie	173210	34481719	1.18
TcMar-Tigger	75170	30557703	1.04
Unclassified	5698	2800558	0.09
Total	3025664	1060239618	

accuracy of system (ability to identify positive example and reject negative examples) was found to be 77% for chromosome 21 and 76% for chromosome 22 (**Table 7**).

Methods

Retrieving genome sequences

Gorilla genome (gorGor3.1; GCA_000151905.1) was retrieved from ensembl database: ftp://ftp.ensembl.org/pub/release-66/fasta/gorilla_gorilla/dna/. The total size of genome arranged in several chromosomes is of total of ~2900 million base pairs of nucleotides.

Repeat sequence retrieval

RepBase Update (RU) is a comprehensive database of repetitive element consensus sequences.[7] Most prototypic sequences from RU are consensus sequences of large families and subfamilies of repetitive sequences. We have used sequences provided by RU for identification of transposable elements (TEs) based on their features.

RepeatMasker for screening of DNA sequences

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences based upon RU.

Whole genome distribution analysis by ELEFINDER

ELEFINDER not only identify repeats but also extracts flanking sequence at each MGE site identified as a preinsertion locus. We used it to find the insertion sites of various MGEs in the gorilla genome.[2,3] This tool finds the nature, distribution, genomic location and the site of truncation for each of the MGE and performs comparative genome analysis. It is a perl based system requiring organism name, chromosome number, element name, genome file and element file as input parameters. The

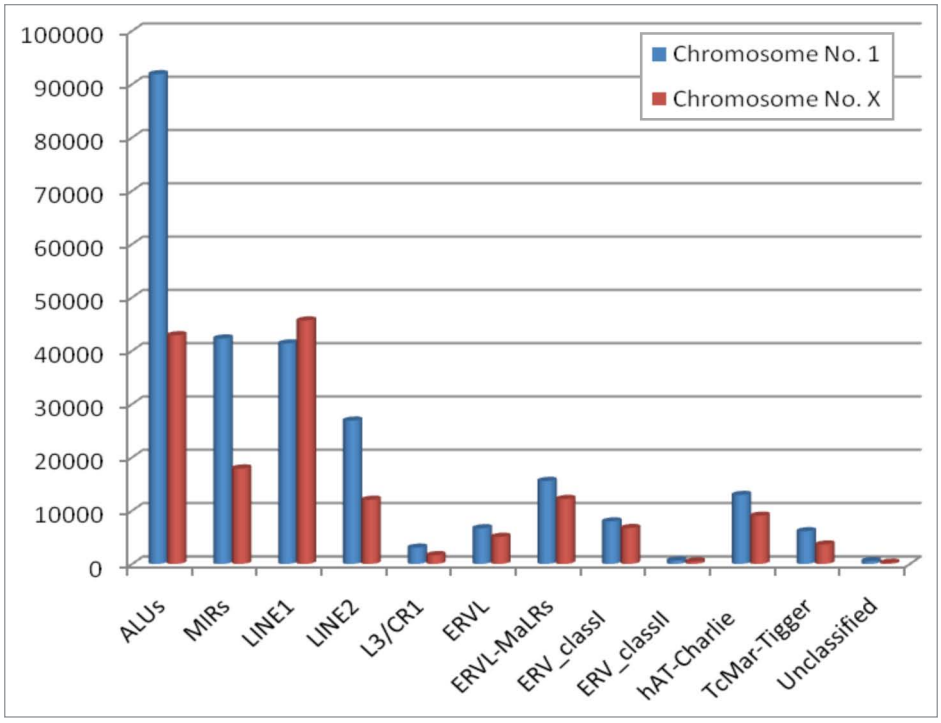


Figure 2. The MGE content comparision between chromososme 1 and X of gorilla genome.

results files are generated after performing BLAST and parsing scripts. The output files comprise the copies of MGE generated by the program categorized into 5′ truncated, 3′ truncated and both side truncated examples (see **Figs. S1–S4**).

Pairwise alignment

To understand the distribution of variety of LINEs present in REPBASE we aligned them sequentially with each other to understand their relationship at sequence level. To view this large data set we generated pairwise percentage similarity matrix (see **Tables. S3–S5**). We also used Gene cluster 3.0 and Java TreeView to view large data set in a tree format (see **Fig. S5**).

DNA SCANNER

DNA SCANNER is a tool which scans DNA using sliding window mechanism for number of different properties such as biophysical parameters, energy measures, potential for protein interactions and sequence based features such as T density, AT density etc. Sequence and physico-chemical based motifs are extracted at insertion site by using this tool.[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). It screens various physicochemical properties as described below:

(A) Structural Signals: DNA Bendability

DNA Bendability is the ability of a DNA to deform under a specific stimulus such as protein binding. A tri-nucleotide model based on DNase-I cutting frequencies predicts that DNase I binds and cuts DNA that is bent toward major groove[5,6,8]

(B) Thermodynamic Signals: Stacking Energy

Stacking energies are indicators of stability, both of a given DNA sequence and as well as protein interactions and thus plays a critical role in formation of local structures[5,9]

(C) Duplex Stability: Free Energy Signals

The relative stability of DNA duplex structure depends upon its base sequence 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 G) and melting behavior of a sequence can be predicted.[5,10]

(D) 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 dinucleotide.[5,8,11] This can be characterized by properties such as propeller twist.

(E) 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.[5,12]

Insertion Site Finder (ISF)

ISF is a machine learning tool which relies on support vector machines for learning and classification tasks. Present ISF is generic version of SVMs,[13] wherein insertion sites from any of the genome can be used. Insertion site finder identifies and predicts insertion sites of the mobile genetic elements. Earlier work involved the study of *E. histolytica* by identification of signals, thereby showing the site for the mobile genetic element to insert at a particular locus. The information provided by DNA

Table 6. DNA SCANNER output of gorilla chromosome 22 showing position and parameter values of A-rule

Position	Parameter value
0	0.275602587
1	0.275720165
2	0.277954145
3	0.279188713
4	0.278306878
5	0.279835391
6	0.282304527
7	0.282716049
8	0.28265726
9	0.282892416
10	0.282951205
11	0.28265726
12	0.283656673
13	0.283891828
14	0.28212816
15	0.282480894
16	0.28377425
17	0.284538507
18	0.284009406
19	0.283127572
20	0.282951205
21	0.284068195
22	0.287360376
23	0.289065256
24	0.288359788
25	0.289535567
26	0.291534392
27	0.292004703
28	0.291828336
29	0.292651382
30	0.29159318

Position	Parameter value
31	0.289300412
32	0.288594944
33	0.28712522
34	0.285008818
35	0.283715461
36	0.281951793
37	0.281128748
38	0.281422693
39	0.281599059
40	0.281834215
41	0.282774838
42	0.28547913
43	0.288536155
44	0.290299824
45	0.292357437
46	0.294532628
47	0.293592005
48	0.292239859
49	0.293004115
50	0.295238095
51	0.295884774
52	0.293180482
53	0.291651969
54	0.293415638
55	0.296061141
56	0.298353909
57	0.300470312
58	0.301528513
59	0.300058789
60	0.29888301
61	0.300587889

Position	Parameter value
62	0.303821282
63	0.304644327
64	0.302880658
65	0.303527337
66	0.306819518
67	0.312698413
68	0.321105232
69	0.328747795
70	0.335390947
71	0.340270429
72	0.345679012
73	0.353497942
74	0.362081129
75	0.371369782
76	0.378542034
77	0.385067607
78	0.392945326
79	0.398059965
80	0.400352734
81	0.400764256
82	0.396413874
83	0.38489124
84	0.367430923
85	0.351440329
86	0.333803645
87	0.310229277
88	0.287536743
89	0.268077601
90	0.254323021

SCANNER is used as positive and negative data sets for training. We have applied ISF on Human genome and *E. histolytica* using bayes rule by using various signals (chemical, thermodynamic and biophysical properties) to produce the score of an insertion site.[2] It gives the probability of a particular property Sj to get inserted at a particular location or insertion site Pi. It also computes sensitivity and specificity for the same. See **Figure 4** for more explanation.

We generated the positive data set labeled as Class P for insertion sites of full length copies of the given elements namely *Alu* and L1. We also created negative data set by shuffling these insertion sequences labeled as N**a**. The independent graphs were generated using DNA Scanner for the two data sets. The observed extrema for the given rule was compared in both data sets. The rules were selected in case they have shown significant value when compared with negative data set as well with background values. For instance, A rule peak was selected for chromosome 22 only when it exceeded the cut off range of 2 Std Dev from the background of A rule values. In addition, there has to be statistical difference of A rule values between positive and negative data set ($p < 0.05$).

Discussion

The gorilla genome is one of the most recently sequenced non-human primates. The successful complete mapping of the gorilla genome has given new and fresh insights to human, chimpanzee and gorilla evolution. We used ELAN

Figure 3 (Right). Various signals upstream of the insertion sites of *Alu* in chromosome 22. The y axis represents value of the property and the x-axis gives the relative position with respect to the insertion site.

Table 7. Performance of ISF in gorilla chromosome 21 and 22 for Alu element

Chromosome	Linear kernel
21	0.7779
22	0.7663

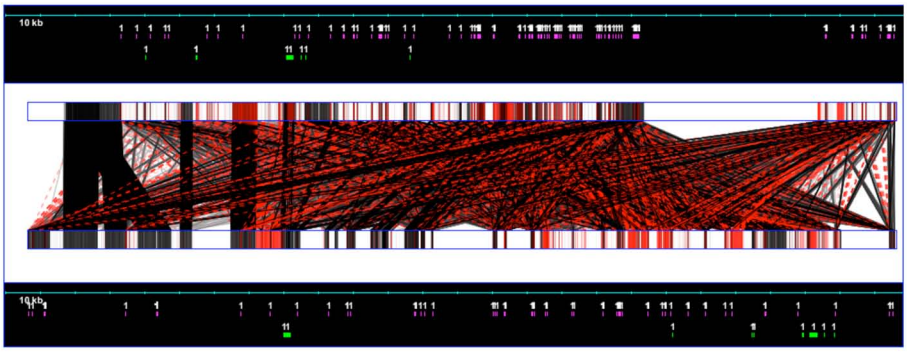
		Predicted	
		negative	Positive
actual	Negative	a	B
	Positive	c	D

Figure 4. Accuracy (AC) is the proportion of the total number of predictions that were correct: AC = (a + d) / (a + b + c + d). Recall is the proportion of positive cases that were correctly identified: R = d / (c + d). Precision is the proportion of the predicted positive cases that were correct: p = d / (b + d). Sensitivity is the ability of the system to identify actual positives: Sn = TP/TP + FN. Specificity is the ability of the system to reject negative examples: Sp = TN/FP+TN.

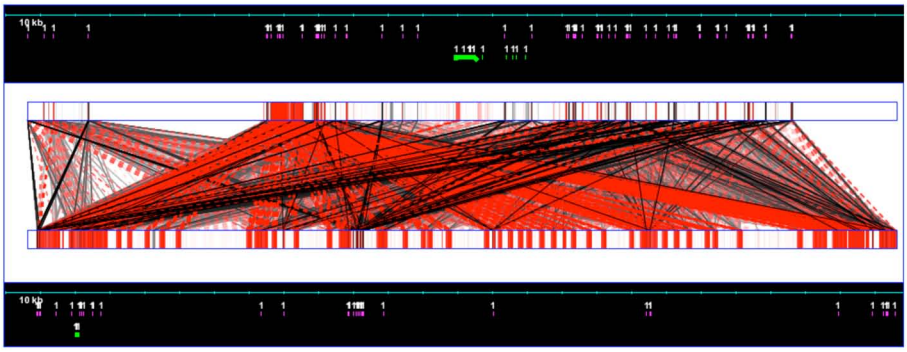
pipeline module to identify various types of mobile genetic elements in the gorilla genome and analyze physical and chemical properties as well as predict insertion sites. The results were compared with other primate genomes such as humans and macaca. It appears that gorilla MGE distribution is similar to human, macaca and mouse suggesting common mechanisms shaping spread of MGE.

To compare the distribution of Alus and L1 element in gorilla and human, we extracted 1 Mb of DNA sequence of both genomes starting from chromosome 1 (Position 1–1000000) from ensembl database. We divided these sequences into 10 Kb non overlapping segments so as to observe the distribution of Alus and L1. The genomic sequences were also aligned using pairwise blast option so as to see the effect of sequence divergence on density of Alus and L1s in the given segment (**Fig. 5**). The GATA tool (http://gata.sourceforge.net/PlotterHelp.html) was used to display the distribution of Alus and L1 in both genomes. It appears that distribution of Alus and L1s is similar even in the areas which are dissimilar at genomic level. We are trying to extend this work at whole genome level and develop a statistical model to understand this approach at multi species level.

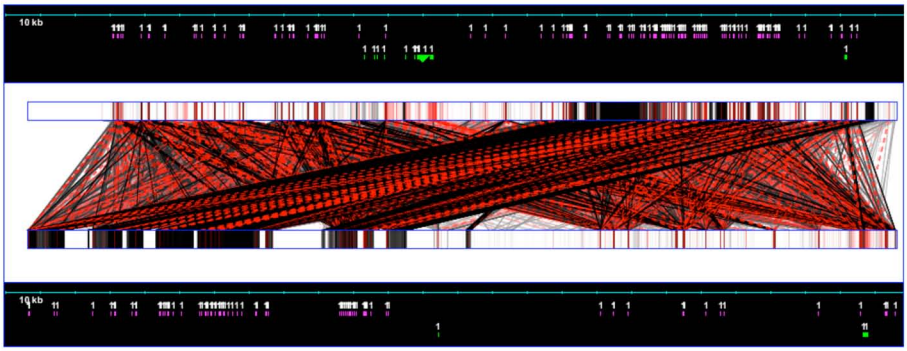
We used repeatmasker to find the different types of Alus and LINEs found in the gorilla genome. This



##########Sequence selected FROM starting position = 250001 TO ending position = 500000



##########Sequence selected FROM starting position = 500001 TO ending position = 750000



########## Sequence selected FROM starting position = 750001 TO ending position = 1000000

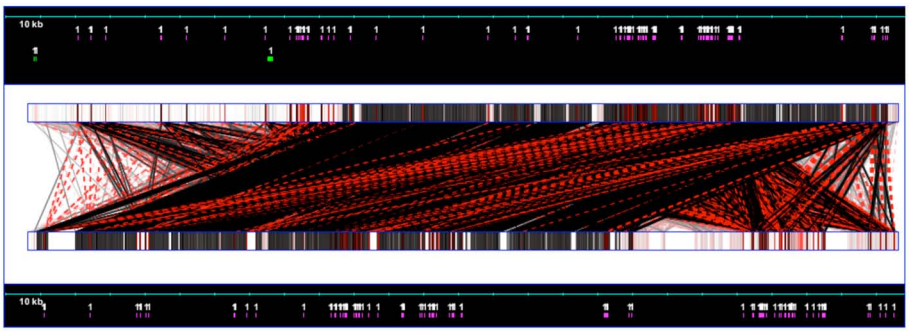


Figure 5. The boxes are plotted against horizontal representations of the input sequences with the reference sequence on top (human). The size of each box is determined by the start and stop positions in the sub-alignment. The shading of the boxes and connector line are scaled according to the sub-alignment score where solid black represents the highest score obtained, light gray the lowest. Lastly the color of the connecting line is used to indicate the sub-alignment orientation, black for +/+, red for +/–. Where windows overlap, those with the highest score are displayed on top. The dark pink lines represent Alu elements whereas parrot green represent L1 elements. The white portions in the beginning and in the last section of human (top) represent undetermined sequences (NNNNN etc). This section represents position numbers 1…250000 bp of human and gorilla chromosome 1. The GFF files were generated for Alus and L1s. The additional figures in the document show subsequent sections of chromosome 1.

was supplemented by ELEFINDER program to identify the distribution of mobile genetic elements in the gorilla genome (see **Tables S1 and S2**). DNA SCANNER was used to scan insertion sites of gorilla chromosomes for a number of different properties such as biophysical, energy, potential for protein interactions and sequence based features. Extrema present in profiles were used to predict insertions sites using machine learning systems. Potential insertion sites were detected using ISF from the result obtained by DNA SCANNER. The patterns or signals observed in gorilla genome were very similar to signals observed at insertion sites of Alus and L1 in humans.

The most common transposable element was found to be *Alu* in gorilla and its distribution is similar to previously reported distribution of *Alu* element in human genome. For all chromosomes, the *Alu* copy number roughly correlates with chromosome length. Though *Alu* element superceded all the MGEs numerically but L1s were present in exceptionally high numbers in X chromosome (see **Fig. 2**) suggesting special role of sex chromosomes for accumulation of MGE. When these results were compared with human

genome, we found characteristically similar behavior in the distribution among humans and gorilla. Different LINEs were found in different numbers in the gorilla genome hence we used various tools to identify various classes of LINEs. In future we plan to analyze MGE in context to genes showing accelerated evolution specially related to brain, sensory and speech.

References

1. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. Nature 2012; 483:169-75; PMID:22398555; http://dx.doi.org/10.1038/nature10842
2. Rawal K, Ramaswamy R. Genome-wide analysis of mobile genetic element insertion sites. Nucleic Acids Res 2011; 39:6864-78; PMID:21609951; http://dx.doi.org/10.1093/nar/gkr337
3. Rawal K, Priya A, Malik A, Bahl R, Ramaswamy R. Distribution of MGEs and their insertion sites in the Macaca mulatta genome. Mob Genet Elements 2012; 2:133-41; PMID:23061019; http://dx.doi.org/10.4161/mge.21074
4. Price AL, Eskin E, Pevzner PA. Whole-genome analysis of *Alu* repeat elements reveals complex evolutionary history. Genome Res 2004; 14:2245-52; PMID:15520288; http://dx.doi.org/10.1101/gr.2693004
5. Mandal PK, Rawal K, Ramaswamy R, Bhattacharya A, Bhattacharya S. Identification of insertion hot spots for non-LTR retrotransposons: computational and biochemical application to Entamoeba histolytica. Nucleic Acids Res 2006; 34:5752-63; PMID:17040894; http://dx.doi.org/10.1093/nar/gkl710
6. Dev BB, Malik A, Rawal K. Detecting motifs and patterns at mobile genetic element insertion site. Bioinformation 2012; 8:777-86; PMID:23055629; http://dx.doi.org/10.6026/97320630008777
7. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005; 110:462-7; PMID:16093699; http://dx.doi.org/10.1159/000084979
8. Crothers DM, Haran TE, Nadeau JG. Intrinsically bent DNA. J Biol Chem 1990; 265:7093-6; PMID:2185240
9. Delcourt SG, Blake RD. Stacking energies in DNA. J Biol Chem 1991; 266:15160-9; PMID:1869547
10. Ollis DL, White SW. Structural Basis of Protein-Nucleic Acid Interactions. Chem Rev 1987; 87:981-95; http://dx.doi.org/10.1021/cr00081a006
11. el Hassan MA, Calladine CR. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. J Mol Biol 1996; 259:95-103; PMID:8648652; http://dx.doi.org/10.1006/jmbi.1996.0304
12. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A 1998; 95:11163-8; PMID:9736707; http://dx.doi.org/10.1073/pnas.95.19.11163
13. Platt J. Fast training of support vector machines using sequential minimal optimization. Cambridge, MA: MIT Press 1999; 185-208

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/mge/article/25675