

Supplementary Material: Rates and Patterns of Great Ape Retrotransposition

Fereydoun Hormozdiari¹, Miriam K. Konkel², Javier Prado-Martinez⁴, Giorgia Chiatante³, Irene Hernando Herraes⁴, Jerilyn A. Walker², Ben Nelson¹, Can Alkan⁶, Peter H. Sudmant¹, John Huddleston¹, Claudia R. Catacchio³, Arthur Ko¹, Maika Malig¹, Carl Baker¹, Tomas Marques-Bonet^{4,5}, Mario Ventura³, Mark A. Batzer², and Evan E. Eichler^{1,7}

1. Department of Genome Sciences, University of Washington, Seattle, WA, USA
2. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA
3. University of Bari, Bari, Italy
4. Institut de Biologia Evolutiva, (UPF-CSIC) Barcelona, Spain
5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
6. Department of Computer Engineering, Bilkent University, Ankara, Turkey
7. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Section 1 – Methods: Paired-end mapping and detection of mobile element insertions (MEIs)

Paired-end mapping for MEI detection

Reads were mapped to the human reference genome Build 36 (hg18) using the BWA¹ alignment software and relaxed edit distance parameters (*-n 0.01*). Read pairing was performed with the same tool limiting the maximum number of occurrences of a read for pairing to 1000 (*-o 1000*). Discordant reads were extracted from the resulting BAM files and mapped again to Build 36 and to a consensus set of mobile elements (RepeatMasker) using the mrsFAST² read aligner and a maximum hamming distance of five mismatches per read. The consensus list of repetitive elements included 30 different *Alu* element sequences and 60 different L1 element sequences (**Table S1**). We note that our method also discovers truncated insertions with more than 5% sequence divergence, thus we can discover L1 and *Alu* subtypes not explicitly included in our list (e.g., L1Pt).

Table S1. Consensus *Alu* and L1 elements used for mapping.

<i>Alu</i> elements	L1 elements		
AluJb	L1HS	L1Med	L1PA6_3end
AluSc	L1P_orf2	L1MB5	L1PA8_3end
AluSg	L1HS_3end	L1MB7	L1PB1_3end
AluSp	L1PA3_3end	L1MB8	L1PBa1_5end
AluSq	L1PA4_3end	L1ME1	L1PBa_5end
AluSx	L1PA5_3end	L1ME2	L1PBb_5end
AluSz	L1PA7_3end	L1ME3	
AluY	L1PA8A_3end	L1ME3A	
AluYa5	L1PA10_3end	L1ME3B	
AluYa8	L1PA11_3end	L1ME4a	
AluYb8	L1PA12_3end	L1MC1	
AluYb9	L1PA13_3end	L1MC2	
AluYc1	L1PA15_3end	L1MC3	
AluYc2	L1PA17_3end	L1MD1	
AluYd2	L1PB2_3end	L1MD2	
AluYd3	L1PB3_3end	L1MD3	
AluYd3a1	L1MA1_3end	L1MC4a	
AluYd8	L1M2	L1MC5	
AluYe5	L1M4	HAL1	
AluYa1	L1M3c	HAL1b	
AluYa4	L1M4b	L1	
AluYb3a1	L1M4c	L1M1_5end	
AluYb3a2	L1MCa	L1MA2_3end	
AluYf1	L1MCb	L1MA3_3end	
AluYg6	L1MC4	L1PA14_3end	
AluYh9	L1MDa	L1PA14_5end	
AluYi6	L1MDb	L1PA15-16_5end	
AluYbc3a	L1Mea	L1PA16_3end	
AluYe2	L1MEb	L1PA17_5end	
AluYf2	L1MEc	L1PA2_3end	

Paired-end-based calls of MEIs

We used two different approaches for calling MEIs in great ape genomes. One is for calling the elements that already exist in the human reference genome and the other is for the ones that do not exist in the human reference genome.

MEIs not in the human reference genome: We modified our tool VariationHunter³⁻⁵ to detect MEIs and, for each sample report, all the paired-end read mappings supporting a particular MEI (this approach was previously successful for predicting *Alu* insertions in sequenced human genomes). We then merged all these mappings into five different groups (one for each species: chimpanzee, human, orangutan, bonobo, and gorilla) and ran the VariationHunter algorithm on each of these five merged sets reporting any possible MEIs supported by at least two reads. Rare calls were not penalized. We next merged the five call sets into a binary matrix indicating the existence of each MEI (1 present, 0 absent) for each sample. The criteria to merge any two calls independently discovered in different species were based on the distance between their predicted insertion loci: MEI predictions within 150 bp (i.e., almost half of the clone size for most samples) of each other and in the same orientation were merged into a single call. We filtered any calls falling inside or within 100 bp from an annotated mobile element in hg18 of the same type^{4,6}. While this filtering step reduced the genomic space in which we can call MEIs by 10-15%, it has a significant effect on improving the false discovery rate (FDR). This filtering step is a standard approach used in all similar previous studies for MEI discovery using high-throughput sequencing technologies^{4,6}. Finally, all calls with fewer than five supporting reads in all samples were filtered.

MEIs present in the human reference genome: We ran our tool VariationHunter^{3,7} for deletion discovery on all great ape samples and considered deletions of *AluY* or L1 elements. We next genotyped all those potential deletions in all samples using the BWA mappings (see above). In any sample if a deletion signature spanning the mobile element was observed (i.e., two or more discordant paired-end reads spanning the element which indicated the deletion of the mobile element), we assumed that mobile element does not exist in that sample. We required all loci to have either an insertion or deletion signature in all individuals; otherwise, the locus was filtered.

Insert size distribution of paired-end reads

The two major features affecting the power of read-pair methods to call structural variation are the average insertion size between read pairs and the sequence coverage of the sample. We did the MEI calling on 93 samples (including 10 humans); however, for most of the analysis we only included the 72 samples with the best homogeneous-observed insert-size distribution and coverage (**Table S2**).

Table S2. Individuals assessed for MEIs. Individuals with poor insert-size distributions or low coverage were excluded from this analysis.

Species	Individual Name	Coverage	Fragment Size	Read Len
Pan troglodytes ellioti	Damian	22.9x	240	51-100
	Paquita	10.4x	206	51
	Basho	11.2x	228	51
	Banyo	7.4x	224	51
	Kopongo	9.4x	210	51
	Akwaya-Jean	25.9x	237	51-100
Pan troglodytes verus	Jimmie	31.7x	477 and 494	100
	Donald	21.7x	212 and 390	100
	Clint	39.3x	294	100
	Bosco	17.8x	210 and 389	100
Pan troglodytes schweinfurthii	Kidongo	49.8x	478 and 495	100
	Bwambale	48.3x	475 and 494	100
	Nakuu	46.4x	491 and 508	100
	Eddy	31.5x	498 and 520	100
Pan troglodytes troglodytes	Doris	39.4x	450 and 480	100
	Vaillant	35.0x	486 and 504	100
	Julie	29.7x	438 and 503	100
	Clara	25.6x	458 and 503	100
Pan paniscus	Bono	32.1x	502 and 519	100
	Hermien	44.9x	459 and 471	100
	Dzeeta	48.1x	469 and 484	100
	Hortenes	41.7x	483 and 500	100
	Pongo	39.2x	465 and 485	100
	Kumbuka	36.3x	489 and 507	100
	Chipita	29.6x	487 and 509	100
	Natalie	33.1x	460 and 476	100
	Desmond	46.5x	450 and 466	100
	Kosana	43.1x	474 and 489	100
Human	Karitiana	15x	270	94
	French	16.7x	290	94
	Papuan	14.7x	265	94
	Dai	14.9x	260	94
	Han	16.6x	275	94
	Sardinian	15.9x	275	94
	Madenka	16.1x	270	94
	Mbuti	14.4x	270	94
	Yoruba	17.3x	285	94
	San	16.4x	265	94
Pongo abelii	Vicki	37.1x	464 and 478	100
	Suma	38.2x	473 and 492	100

	Rochelle	31.7x	464 and 481	100
	Kiki	34.1x	473 and 510	100
	Dunja	41.1x	474 and 493	100
	Buschi	35.2x	469 and 489	100
Pongo pygmaeus	Tilda	37.7x	468 and 482	100
	Nonja	32.6x	464 and 479	100
	Napoleon	36.8x	473 and 488	100
	Sari	32.3x	475 and 491	100
	Lotti	29.2x	448 and 474	100
	Kajan	31.4x	472 and 506	100
	Temmy	29.2x	514 and 541	100
		Mkubwa	18.9x	216 and 391
G. beringei	Kaisi	34.6x	482 and 498	100
Gorilla gorilla diehli	Nyango	23.8x	451	100
Gorilla gorilla gorilla	Carolyn	12.7x	436	100
	Choomba	25.4x	460	100
	Coco	29.7x	468 and 490	100
	Delphi	40.1x	461 and 511	100
	Dian	35.2x	489 and 527	100
	Dolly	17.5x	438	100
	Anthal	23.4x	434	100
	Helen	15.6x	401	100
	Banjo	31.3x	474 and 492	100
	Amani	39.2x	486 and 505	100
	Katie	22x	445	100
	Katie_KB4986	15.7x	423	100
	Kolo	31x	479 and 497	100
	Akiba_Beri	22.8x	442	100
	Mimi	32.1x	462 and 479	100
	Paki	14.7x	471	100
	Sandra	27.7x	467	100
	Vila	13.2x	431	100

Section 2 – MEI statistics

Using the 93 samples, we predicted more than 50,000 loci of *Alu* and L1 insertions in great ape evolution. In **Tables S3, S4 and S5** a summary of *Alu* and L1 insertions is provided.

Table S3. Statistics of *Alu* insertions on 93 great ape genomes.

Species	Samples	Novel <i>Alu</i> Insertion Statistics (vs. hg18)			Gene Intersecting			AIM	
		Total # Ins	Avg # Ins	Total # Ins (Seg Dup filter)	CDS	UTR-3'	UTR-5'	Fix***	Poly (>0.5)****
<i>Pan</i>	35	11157	3237	10046	1	59	4	451	1014
<i>P. paniscus</i>	12	4229	3002	3922	0	21	0	784	1361
<i>Pan troglodytes</i>	23	8715	3360	7721	1	47	4	307	708
<i>P. t. troglodytes</i>	4	5570	3551	5207	1	30	1	6	46
<i>P. t. ellioti</i>	8	5341	3418	4685	1	25	3	11	87
<i>P. t. verus</i>	5	4129	3057	3753	1	20	1	43	132
<i>P. t. schweinfurthii</i>	6	5607	3408	4962	1	31	2	5	79
<i>Homo Sapiens</i>	10	2932*	764**	2739	2	14	0	16	299
<i>Gorilla</i>	35	8809	4604	8223	0	61	0	367	2770
<i>G. g. gorilla</i>	32	8445	4598	7888	0	52	0	0	171
<i>G. b. graueri</i>	2	5382	4755	5079	0	30	0	189	189
<i>G. g. diehli</i>	1	4491	4491	4269	0	29	0	16	16
<i>Pongo</i>	13	1739	1201	1635	1	7	0	321	1029
<i>P. abelii</i>	6	1666	1263	1569	1	7	0	3	25
<i>P. pygmaeus</i>	7	1571	1147	1478	1	7	0	4	32

* There are 7,041 total *Alu* insertions in human samples that also exist on hg18.

** There are 6,392 Avg. *Alu* insertions in human samples that also exist on hg18.

*** Fix AIMs (Ancestry-Informative Marker): Insertions seen in all samples of a species or subspecies and no other species or subspecies.

**** Fix Poly AIMs (Ancestry-Informative Marker): Insertions seen in >0.5 samples of a species or subspecies and no other species and subspecies.

Table S4. Statistics of L1 insertions on 93 great ape genomes.

Species	Samples	Novel L1 Insertion Statistics (vs. hg18)			Gene Intersecting			AIMs (>200 bp)		
		Total # Ins	Avg # Ins	Total # Ins (Seg filter)	Ins (Seg Dup)	CDS	UTR-3'	UTR-5'	Fix ***	Poly (>0.5)***
<i>Pan</i>	35	7215	2162	6314		0	6	0	280	750
<i>P. paniscus</i>	12	3067	2095	2773		0	3	0	181	413
<i>Pan troglodytes</i>	23	6066	2198	5234		0	5	0	53	180
<i>P. t. troglodytes</i>	4	3326	2262	3062		0	4	0	2	17
<i>P. t. ellioti</i>	8	3951	2291	3370		0	3	0	2	87
<i>P. t. verus</i>	5	2958	1975	2571		0	3	0	17	65
<i>P. t. schweinfurthii</i>	6	3697	2215	3164		0	5	0	2	27
<i>Homo Sapiens</i>	10	448*	121**	340		1	0	0	3	36
<i>Gorilla</i>	35	5059	2159	4602		2	11	1	89	1105
<i>G. g. gorilla</i>	32	4686	2147	4281		2	10	1	1	44
<i>G. b. graueri</i>	2	2748	2294	2563		2	7	1	135	135
<i>G. g. diehli</i>	1	2295	2295	2163		2	7	1	18	18
<i>Pongo</i>	13	13410	7771	12893		1	13	0	2809	5948
<i>P. abelii</i>	6	11378	7848	10951		1	13	0	64	419
<i>P. pygmaeus</i>	7	10797	7706	10383		1	10	0	98	580

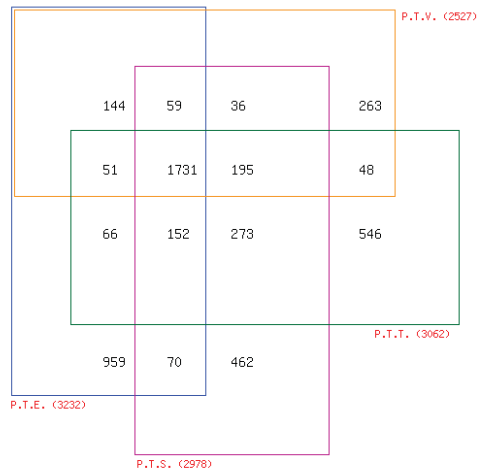
* There are 1,488 total L1 insertions in human samples that also exist on hg18.

** There are 1,397 Avg. L1 insertions in human samples that also exist on hg18.

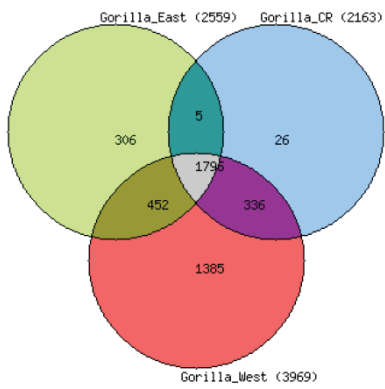
*** Fix AIMs (Ancestry-Informative Markers): Insertions seen in all samples of a species or subspecies and no other species and subspecies.

**** Poly AIMs (Ancestry-Informative Marker): Insertions seen in >0.5 samples of a species or subspecies and no other species and subspecies.

L1 insertion overlap in chimpanzee



L1 insertion overlap in gorilla



L1 insertion overlap in orangutan

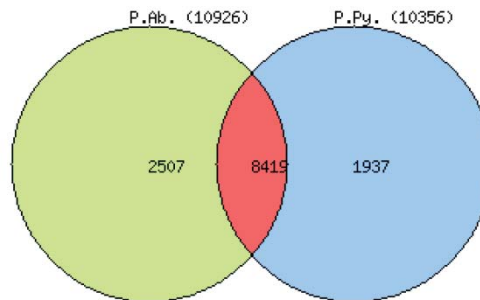


Fig. S2. The overlap between L1 insertion predictions for each subspecies of chimpanzee, gorilla, and orangutan for the top 72 samples.

Section 3 – Validation and comparison with great ape reference genomes

Comparison of MEIs represented in the panTro3

To determine the quality of our MEI predictions and genotypes, we performed a comparison of our predictions to the assembled genome of a chimpanzee. We note that the Western chimpanzee individual Clint, who was used to build the majority of the chimpanzee reference genome⁸, was also among the samples we sequenced and assessed for MEIs. Among the predicted Clint *Alu* insertions that we could LiftOver (UCSC genome browser) to the panTro3 reference genome 89.9% (2,346/2,610) validated. For L1 insertions 91.3% (1,598/1,749) validated. The remaining ~10% of the predicted insertions are likely the result of heterozygous mobile elements in the reference genome in addition to false calls by our method (**Fig. S3**). We also assessed our individual chimpanzee subspecies MEI calls against panTro3 (**Fig. S4**). As expected, we find the highest concordance between Western chimpanzee calls and panTro3. More frequent calls were more likely to be observed in panTro3 (**Fig. S5**).

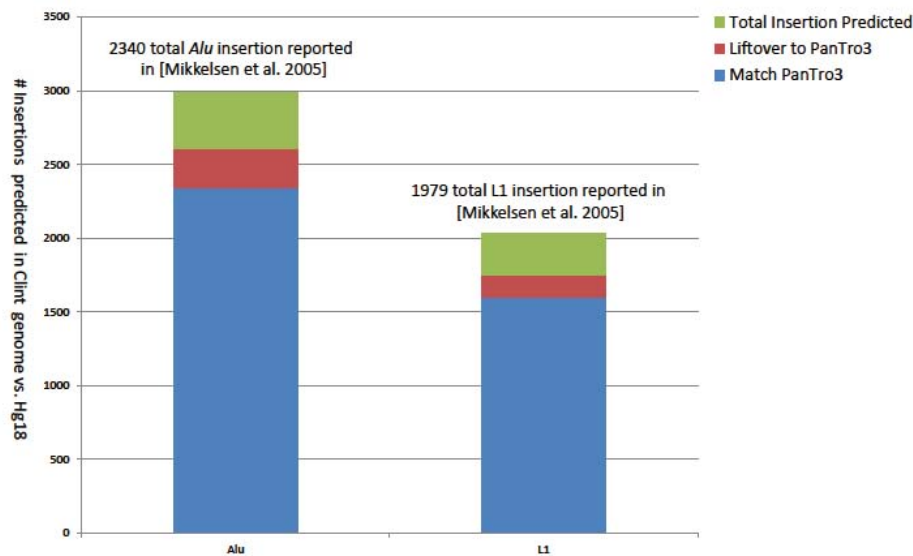


Fig. S3. The total number of *Alu* and L1 insertions detected against Build 36 for the individual Clint is plotted in green. The number of loci that were successfully lifted over the panTro3 is plotted in red and the number of those that validated is plotted in blue.

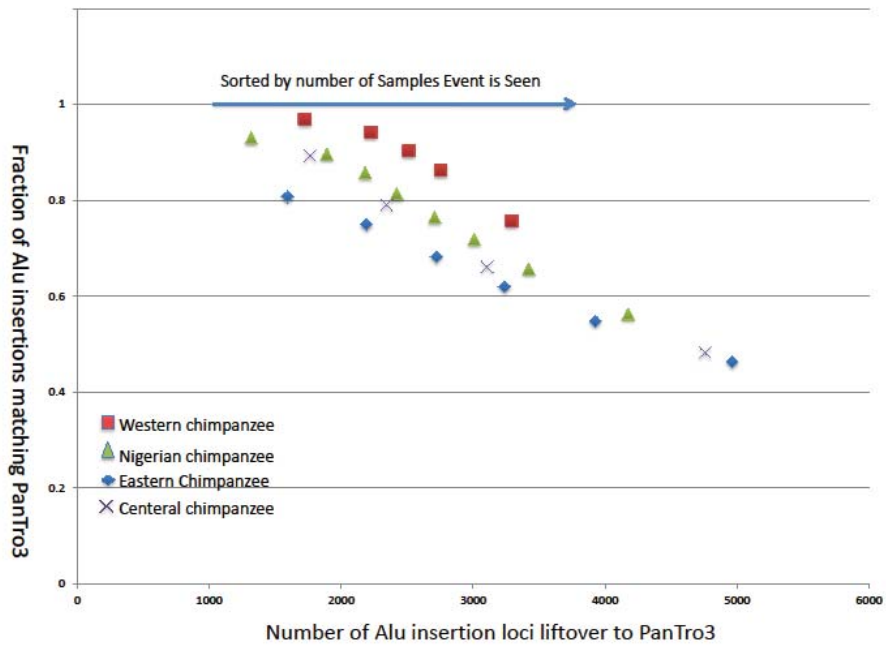


Fig. S4. The ratio of *Alu* insertion loci validated against panTro3 is plotted in decreasing order of frequency of the event. Events at higher frequency (further to the left) have a higher validation rate.

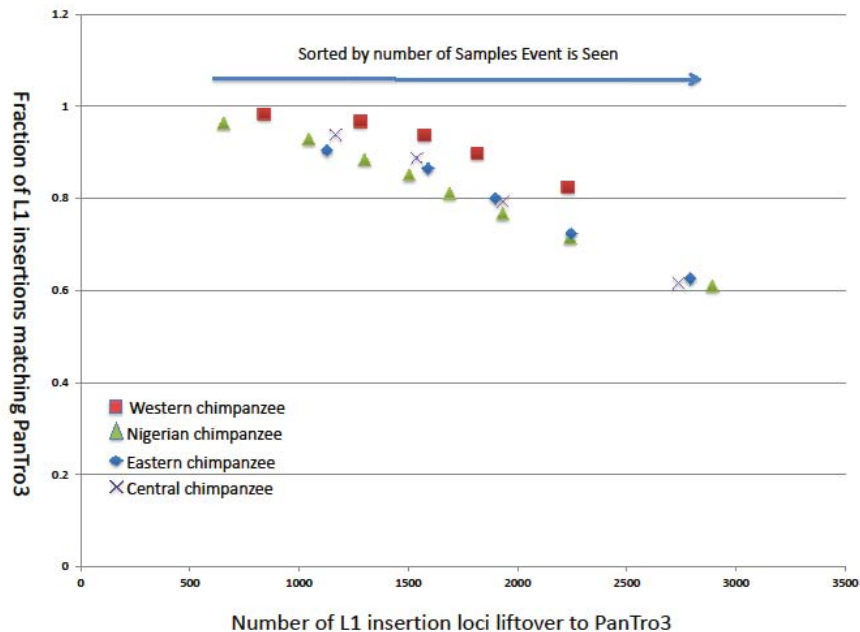


Fig. S5. The ratio of L1 insertion loci validated against panTro3 is plotted in decreasing order of the frequency of the event. Events at higher frequency (further to the left) have a higher validation rate.

PCR validation of MEIs

In order to assess the FDR of our calls, we performed targeted PCR experiments. We randomly selected 13 predicted *Alu* insertion loci and 9 predicted L1 insertion loci and performed PCR on eight (sequenced) samples from multiple species (3 chimpanzees, 1 bonobo, 3 gorillas, and 1 orangutan). The genotyping concordance of our predictions with the PCR results for *Alu* insertions was 86/90 (>95%) and for L1 insertions was 54/55 (>98%). The summary results of these experiments are shown in (Fig. S6–S9).

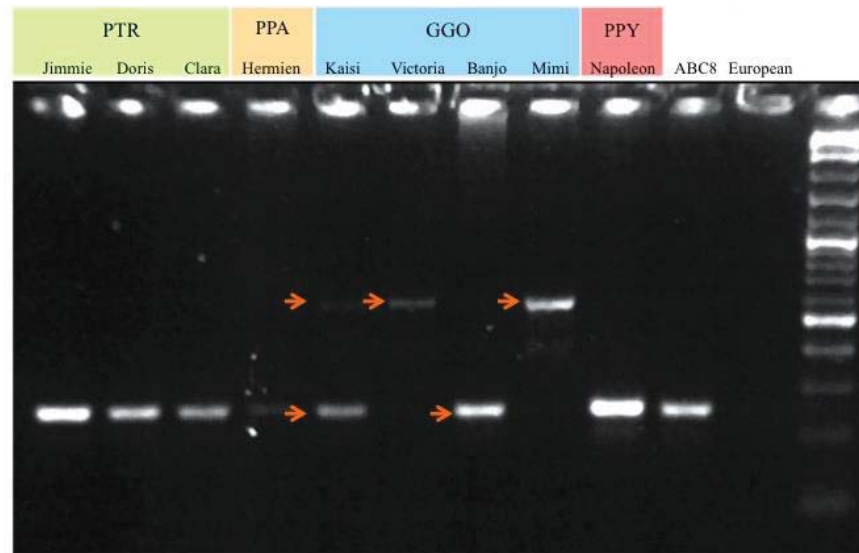


Fig. S6. Targeted PCR validation of chr11:29087269-29087668. The expected PCR product size is 253 bp without the *Alu* insertion and 563 bp with the *Alu* insertion. Gorilla individuals Kaisi, Victoria, and Mimi have the insertion while Banjo does not.

We additionally tested 55 fixed *Alu* insertions (predicted as AIM or ILS) validating 54 of them (98.2%). This puts the FDR of our predictions at less than 1.8% for *Alu* insertions. Of the set of 68 fixed (AIM or ILS) L1 insertions tested, a total of 60 validated, while 5 other loci did not amplify 95.2% (60/63). This puts the FDR for our L1 insertion at the range of 3.2%. In total, the FDR is ~4.3% (6/139).

	PTR			PPA	GGO			PPY	tot
	Jimmie	Doris	Clara	Hermien	Kaisi	Banjo	Mimi	Napoleon	
✓	12	8	8	13	13	7	13	12	86
x	0	2	1	0	0	1	0	0	4
-	1	3	4	0	0	5	0	1	14

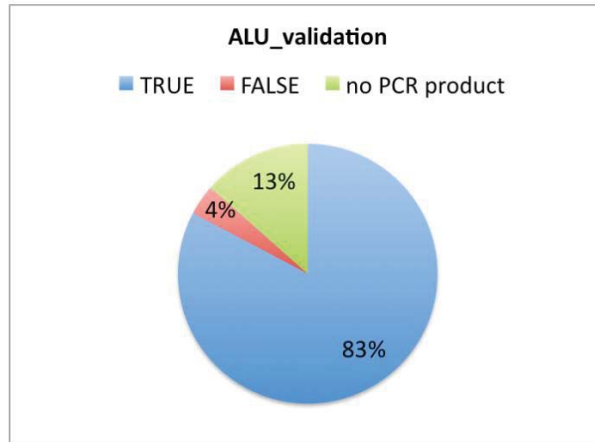


Fig. S7: Breakdown of PCR validation experiments on 13 loci *Alu* insertion loci tested in eight individuals. We observe a 98% concordance.

	PTR			PPA	GGO			PPY	tot
	Jimmie	Doris	Clara	Hermien	Kaisi	Banjo	Mimi	Napoleon	
✓	7	4	4	9	9	3	9	9	54
x	1	0	0	0	0	0	0	0	1
-	1	5	5	0	0	6	0	0	17

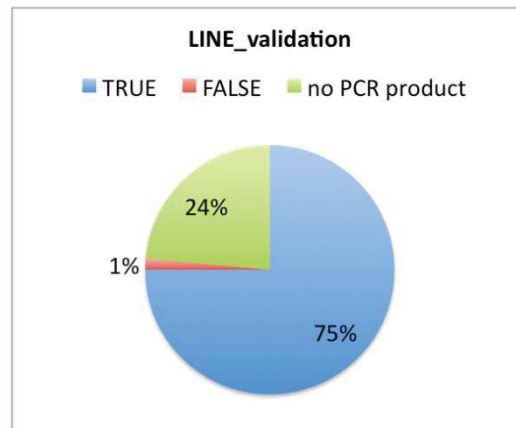
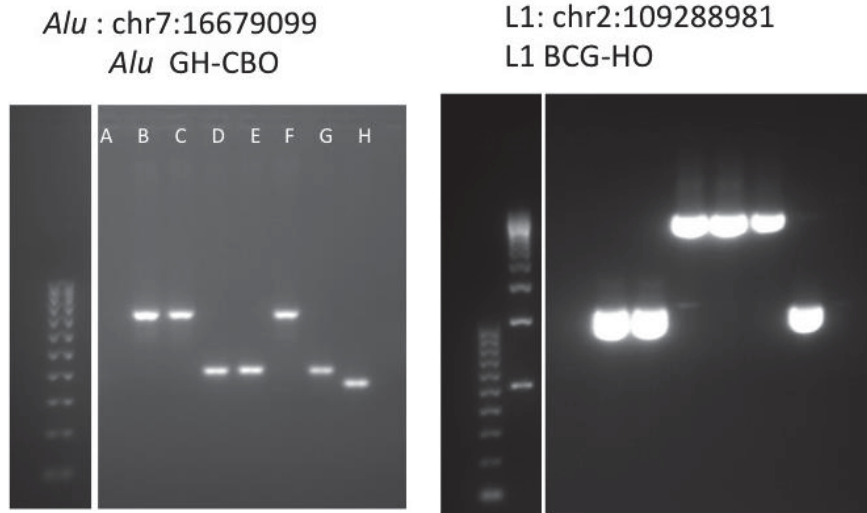


Fig. S8: Breakdown of PCR validation experiments on 9 L1 insertion loci tested in eight individuals. We observe a 96.7% concordance.



	ID	Sample Name	Species Name	Source
A	TLE	negative control	n/a	n/a
B	HeLa (CCL2)	Human	<i>Homo sapiens</i>	ATCC
C	NA19240	Human	<i>Homo sapiens</i>	Coriell
D	NS06006	Chimpanzee ("Clint")	<i>Pan troglodytes</i>	IPBIR
E	PR00661	Bonobo	<i>Pan paniscus</i>	IPBIR
F	AG05251	Gorilla (lowland)	<i>Gorilla gorilla</i>	Coriell
G	GM06213	Sumatran orangutan	<i>Pongo abelii</i>	Coriell
H	CCL70	African Green Monkey	<i>Chlorocebus aethiops</i>	ATCC

Fig. S9: PCR validation of ILS *Alu* and *L1* insertions in a randomly chosen panel of individuals.

Finally, we provide in **Table S5** the summary of all of the PCR tested loci of *Alu* and *L1* insertions.

Table S5. The summary of all PCR tested *Alu* and *L1* insertions.

	<i>Alu</i>	L1				
			# Tested	Validated	Unvalidated	# Tested
Random (For Genotype Concordance)	13	12	1	9	8	1
AIMs	42	42	0	31	23	3
ILS	13	12	1	37	37	0
Total (AIM+ILS)	55	54	1	68	60	3
Total	68	66	2	77	68	4

PCR validations approach for the ILS MEIs

ILS candidate locus insertion coordinates were recorded based on the human reference genome (hg18). For primer design, 600 bp of flanking sequence were added upstream and downstream of each human candidate insertion coordinate, with the nucleotide sequences retrieved using Galaxy. Next, orthologous sequences were retrieved for chimpanzee, gorilla, orangutan, and rhesus macaque using BLAT⁹. A multi-alignment was then performed using BioEdit¹⁰.

Primers were designed flanking the ILS coordinates in regions of high homology among all species. Primer3 was used for primer design. Each primer was checked against the multi-alignment to ensure a high likelihood for amplification in all species. In addition, each primer was BLATed⁹ against the human reference genome to confirm the uniqueness of the primers. If necessary, alternative primers were selected and tested. An *in silico* PCR (genome.ucsc.edu) was performed for each primer combination a) to confirm that only one amplicon was predicted, and b) to determine the size of the predicted filled (MEI present) and empty (insertion absent) PCR product.

Based on the predicted PCR amplicon size, PCR reactions were performed using either Jumpstart *Taq* DNA polymerase (Sigma Aldrich, St. Louis, MO) for PCR products < 2 kb or Takara LA-*Taq* (Clontech Laboratories, Inc., Mountain View, CA) for products predicted to be 2 to 6 kb, in each case following the manufacturer's suggested protocol. PCR products from the smaller set were size fractionated in a horizontal gel chamber on a 2% agarose gel containing 0.1 µg/ml ethidium bromide for 60 minutes at 175V. Larger PCR products were size fractionated on a 1% agarose gel for 90 minutes to 2 hours at 150V. UV-fluorescence was used to visualize the DNA fragments.

Following PCR validations, 3 of the 50 ILS candidate loci remained inconclusive. Sanger sequencing of PCR amplicons was performed to confirm the presence of a shared insertion event. Incomplete lineage sorting was confirmed in all three cases. Of the 50 ILS candidate loci evaluated, 10 *Alu* loci were ILS validated, 3 determined to be false, and all 37 L1 loci validated as ILS events.

Breakpoint analysis

The sequence breakpoints (i.e., the insertion coordinates for each MEI shared by BCG and absent from HO) were determined based on the multiple alignments. For this analysis, each breakpoint was defined as the coordinate closest to the 5' end of the MEI sequence. To allow comparison with the predicted insertion coordinates, the breakpoint coordinates were recorded for the human genome (hg18), even though the insertion is not present in human.

Section 4 – Trees constructed from different genetic markers

Here we show the phylogenetic tree constructed by *Alu* and L1 insertions using UPGMA method. The trees built using UPGMA¹¹ are very similar to the trees constructed using neighbor-joining method shown in the main paper.

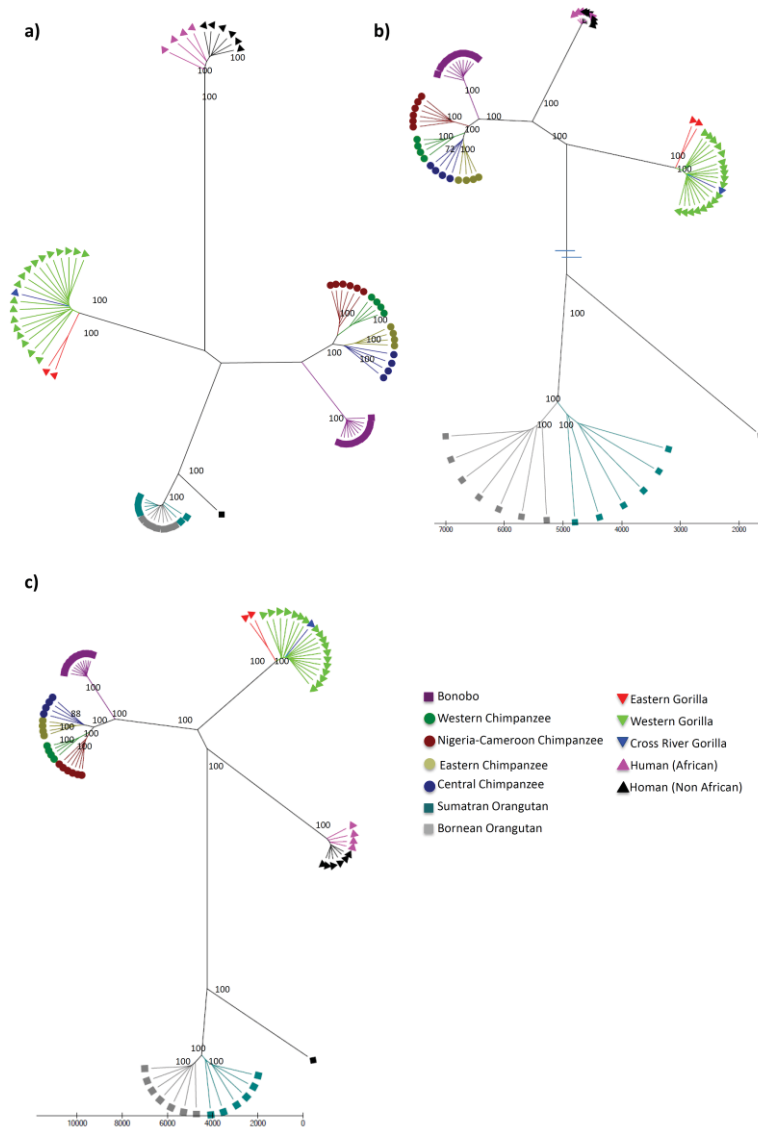


Fig. S10. Phylogenetic tree build using UPGMA algorithm for *Alu*, L1, merged *Alu* - L1 insertions is shown.

Section 5 – Genetic diversity (PCA, SNP heterozygosity and population diversity)

PCA

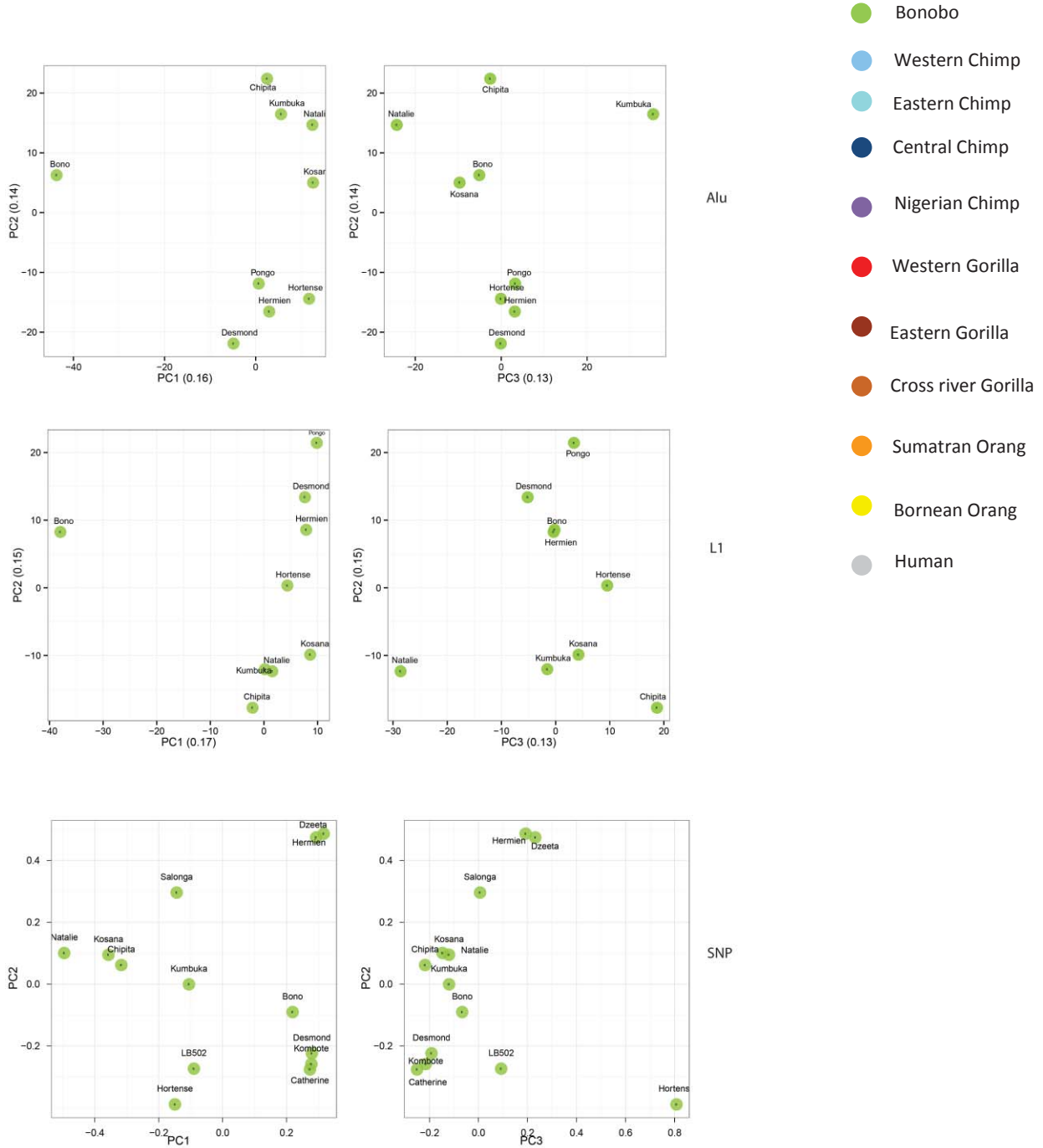


Fig. S11. PCA analysis of bonobo samples using *Alu*, L1 and SNP data.

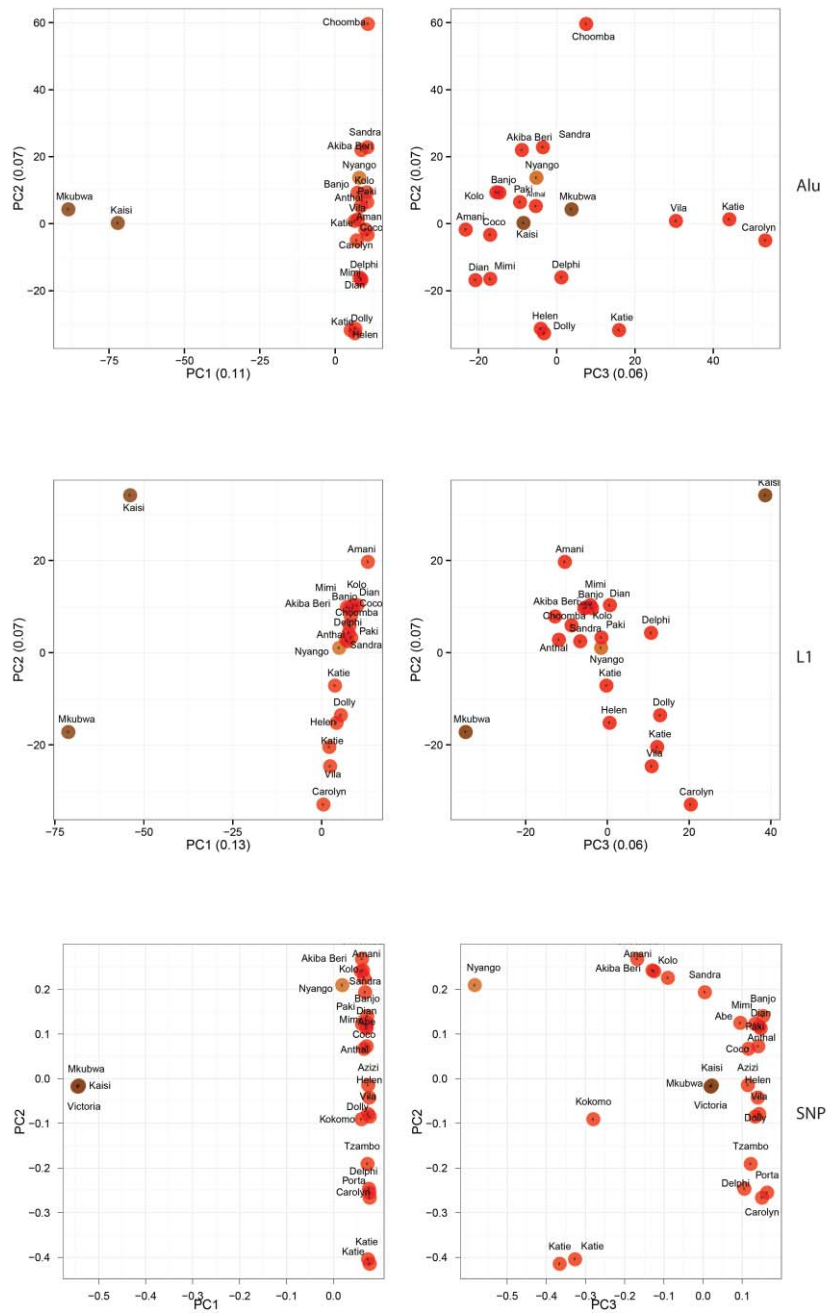


Fig. S12. PCA analysis of gorilla samples using *Alu*, L1 and SNP data.

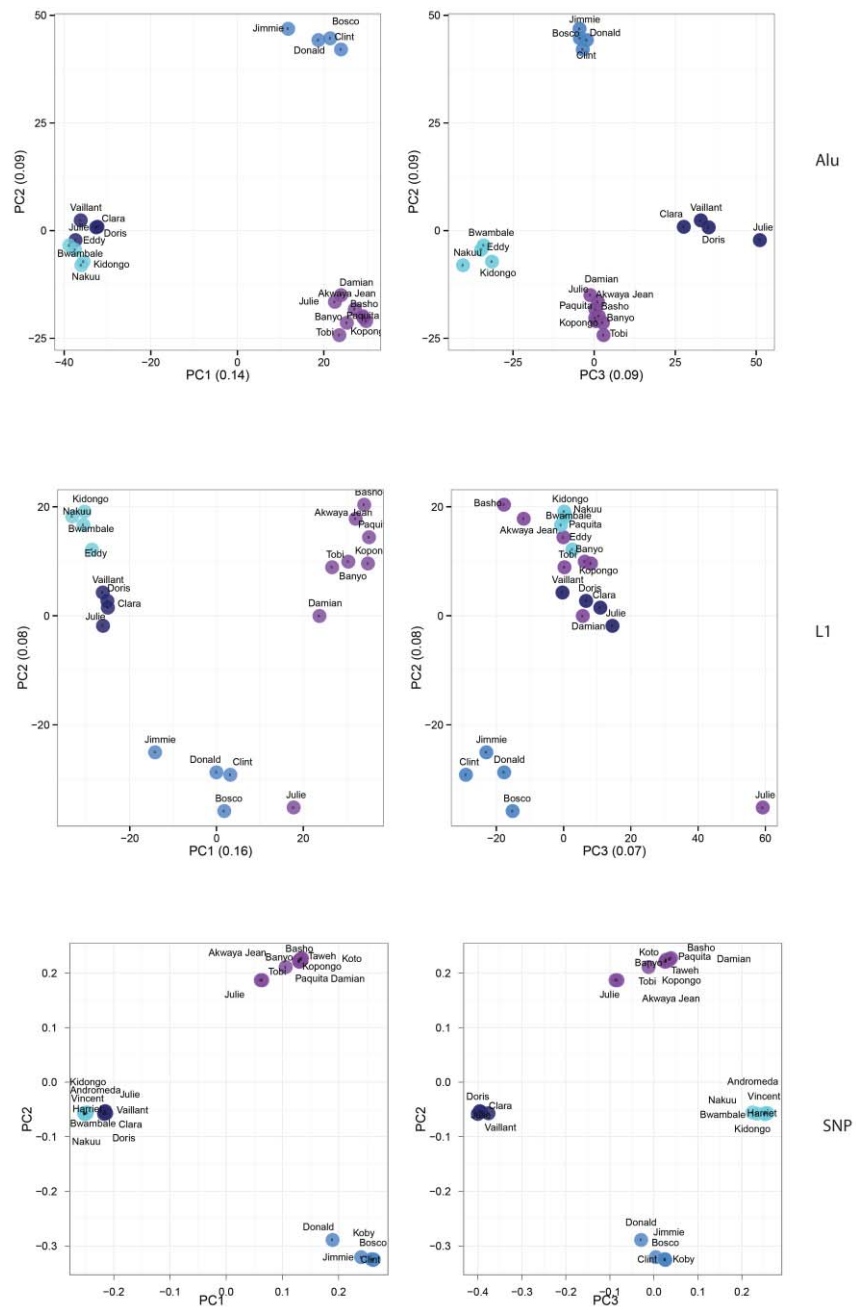


Fig. S13. PCA analysis of chimpanzee samples using *Alu*, L1 and SNP data.

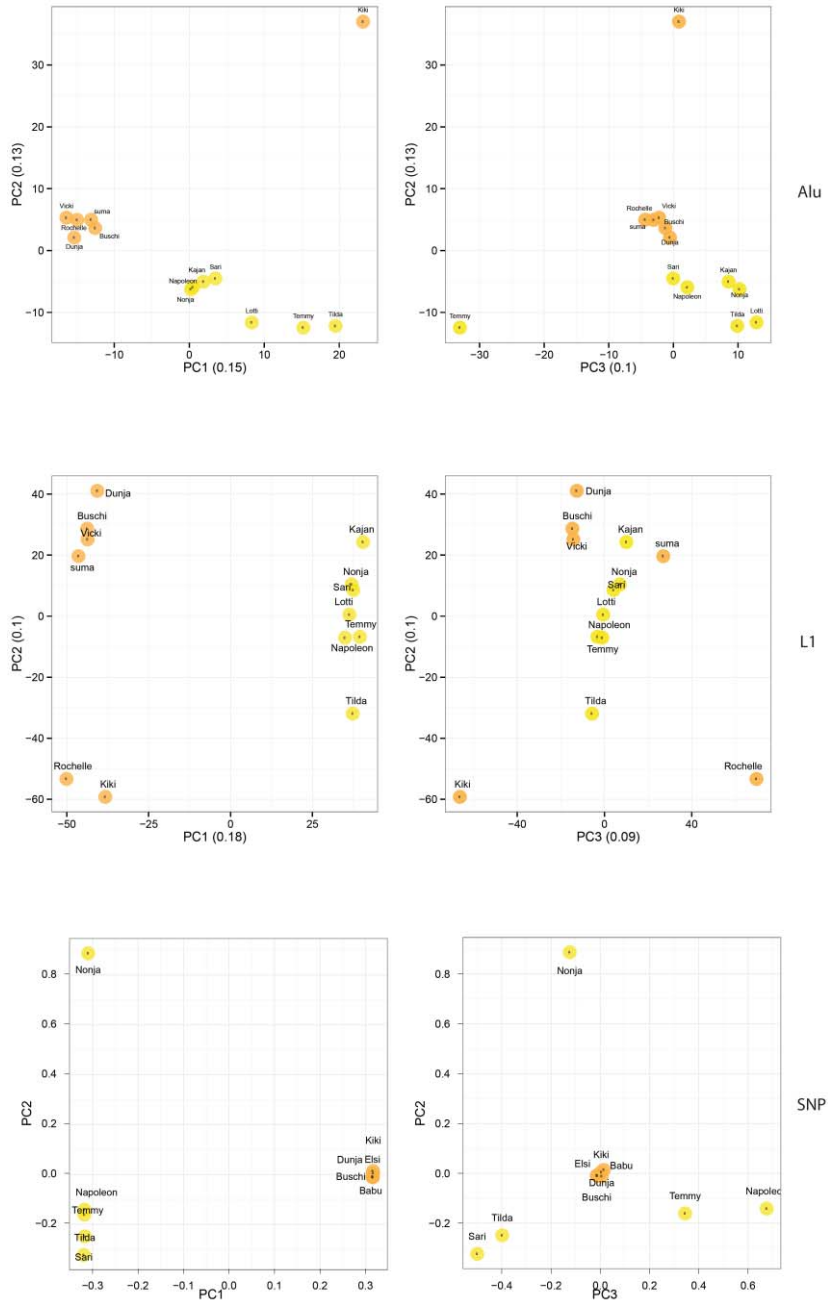


Fig. S14. PCA analysis of orangutan samples using *Alu*, L1 and SNP data.

SNP heterozygosity vs. *Alu* and L1 diversity

We compared the SNP heterozygosities of each subspecies to the *Alu* and L1 diversity. The *Alu* or L1 diversity is calculated as the average pairwise total differences in *Alu* or L1 insertions for each individual for each subspecies. There is no correlation between *Alu* diversity and SNP heterozygosity ($r^2=0.077$), while there is moderate correlation between L1 diversity and SNP heterozygosity ($r^2=0.655$).

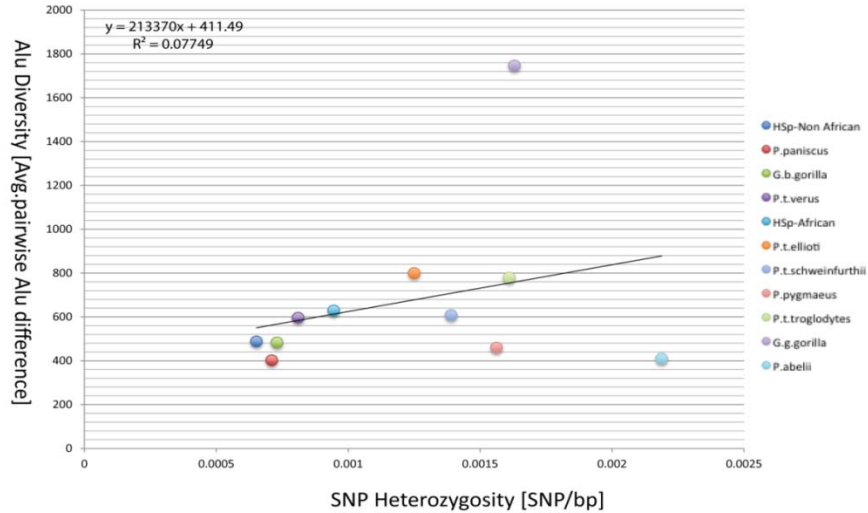


Fig. S15. SNP heterozygosity vs. average *Alu* pairwise differences.

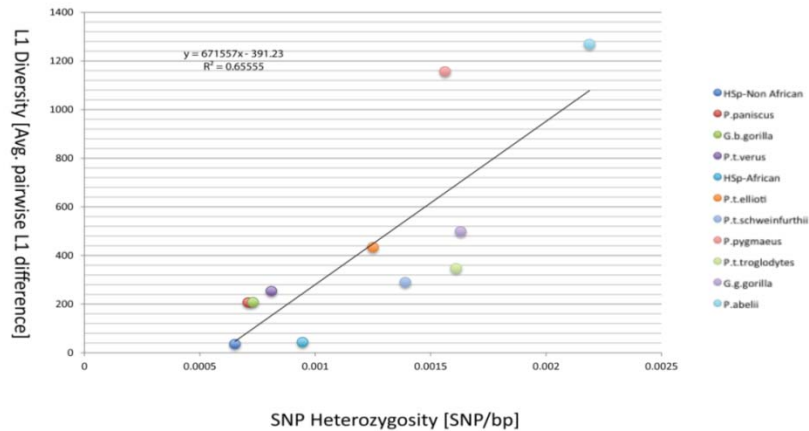


Fig. S16. SNP heterozygosity vs. average L1 pairwise differences.

Population diversity

We looked at the distribution of absolute difference between subspecies/subgroups of each species with each other and with other subspecies of the same species. Orangutan *Alu* insertion is quite similar for the individuals of the two subspecies with each other and between each other. For L1 insertions orangutan shows the most difference.

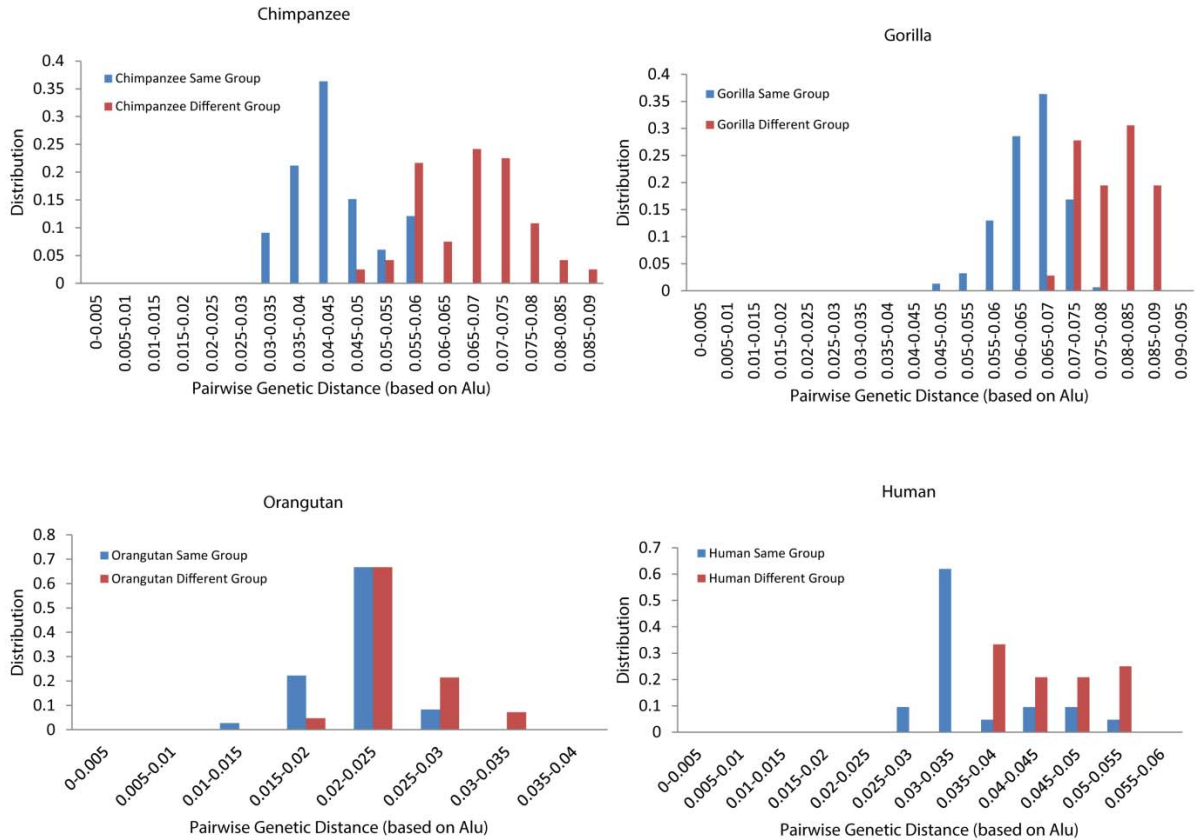


Fig. S17. Average *Alu* difference between individuals in the same group (subspecies) and between different groups (subspecies) for each of the chimpanzee, gorilla, human, and orangutan species.

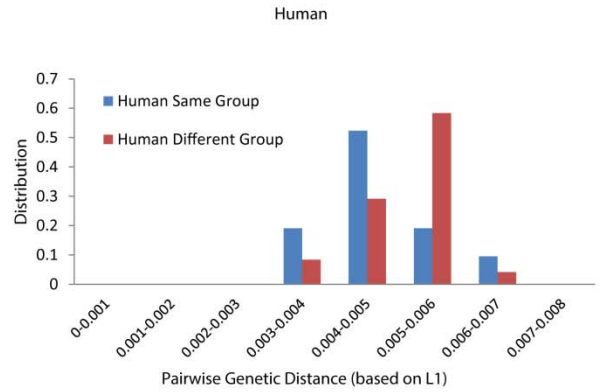
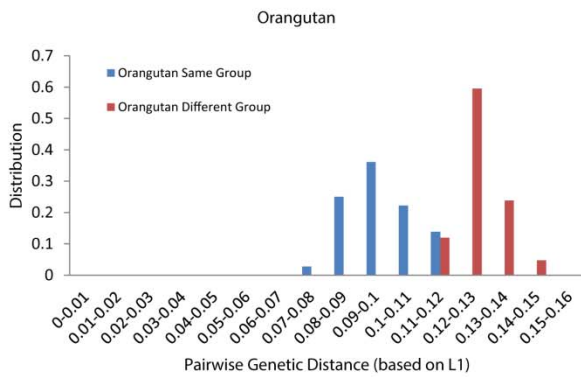
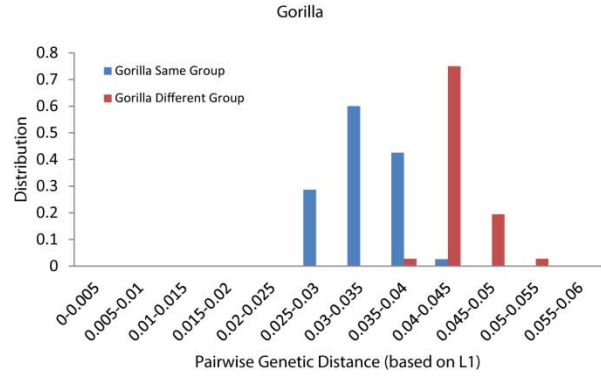
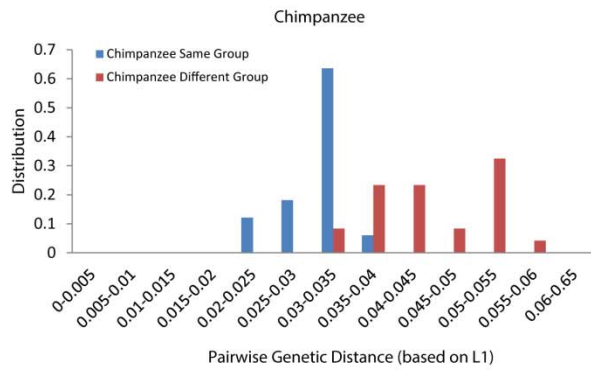


Fig. S18. Average L1 difference between individuals in the same group (subspecies) and between different groups (subspecies) for each of the chimpanzee, gorilla, human, and orangutan species.

Section 6 –Fixed incomplete lineage sorting loci

In this subsection we consider the ILS events that are fixed. In total we predicted two and five total fixed *Alu* ILS insertions in HB-C and HC-B, respectively. We used the same ten high-quality samples per species that we used for rate calculation to assign the discovery ILS MEI loci.

There were more total fixed L1 ILS insertions than fixed *Alu* ILS insertions. We predicted a total of seven and four total fixed L1 ILS insertions in HB-C and HC-B, respectively. The total number of fixed *Alu* and L1 insertions between BC-H is 783 and 708. This puts the fixed ILS rate for HB-C and HC-B around 0.6% using MEIs. For HG-CBO the rate of fixed ILS is around 13.5% and the rate of CBG-HO ILS using MEIs is around 12.2%. Note that the numbers reported in **Fig. S19** and **S20** are the fixed ILS loci (i.e., they are reported to be fixed in all samples of each species), while the number reported in the main paper are all the ILS events (fixed and polymorphic in some species).

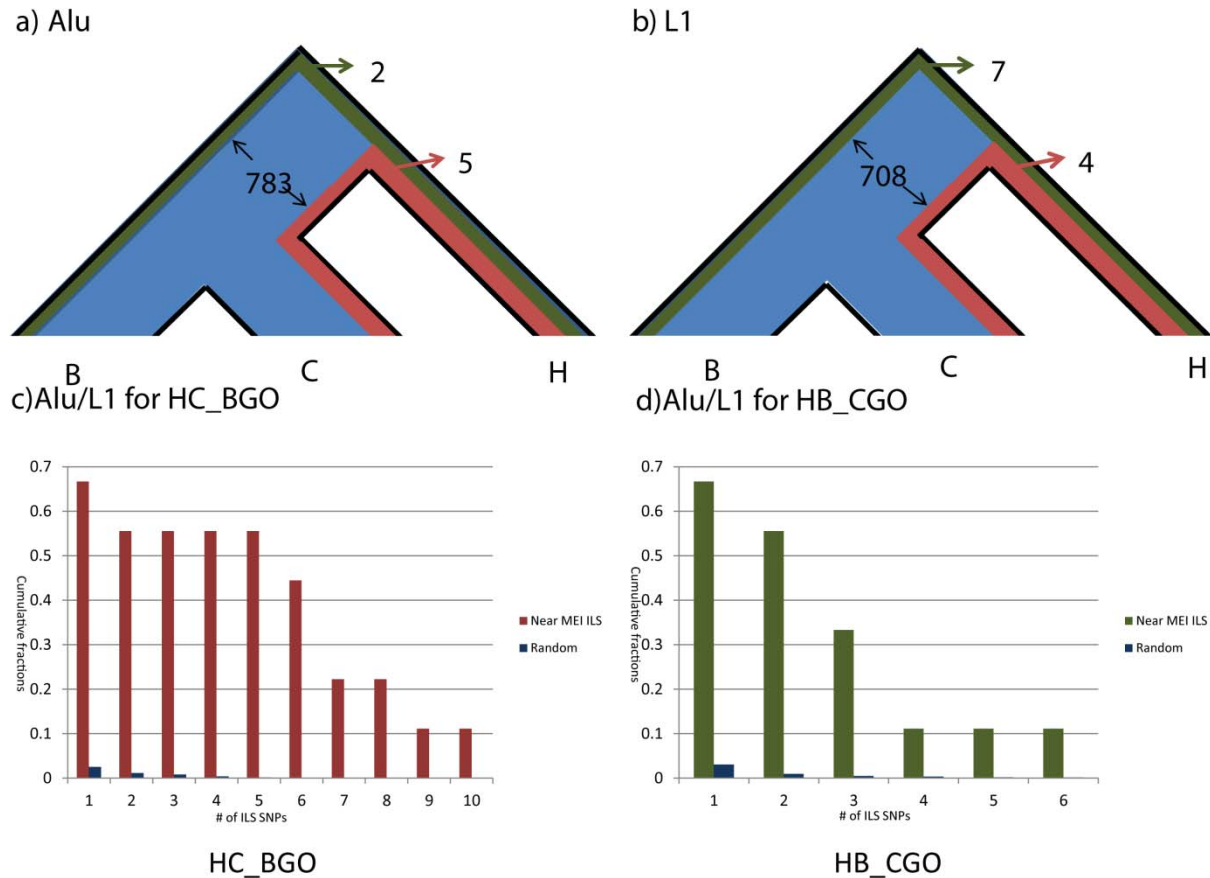


Fig. S19. Total *Alu* ILS insertions common between bonobo (B) and human (H) (green); total *Alu* ILS insertions common between human (H) and chimpanzee (C) (red) with total *Alu* insertions common between chimpanzee (C) and bonobo (B) (blue). b. ILS information for L1 insertions. c. Cumulative distribution of *Alu*/L1 ILS events concordant with SNP

ILS between for HC-B. x-axis shows number of ILS SNPs close to the predicted ILS *Alu/L1* insertion, and y-axis shows the cumulative fraction of events. d. Cumulative distribution of *Alu/L1* ILS events concordance with SNP ILS between HB-C.

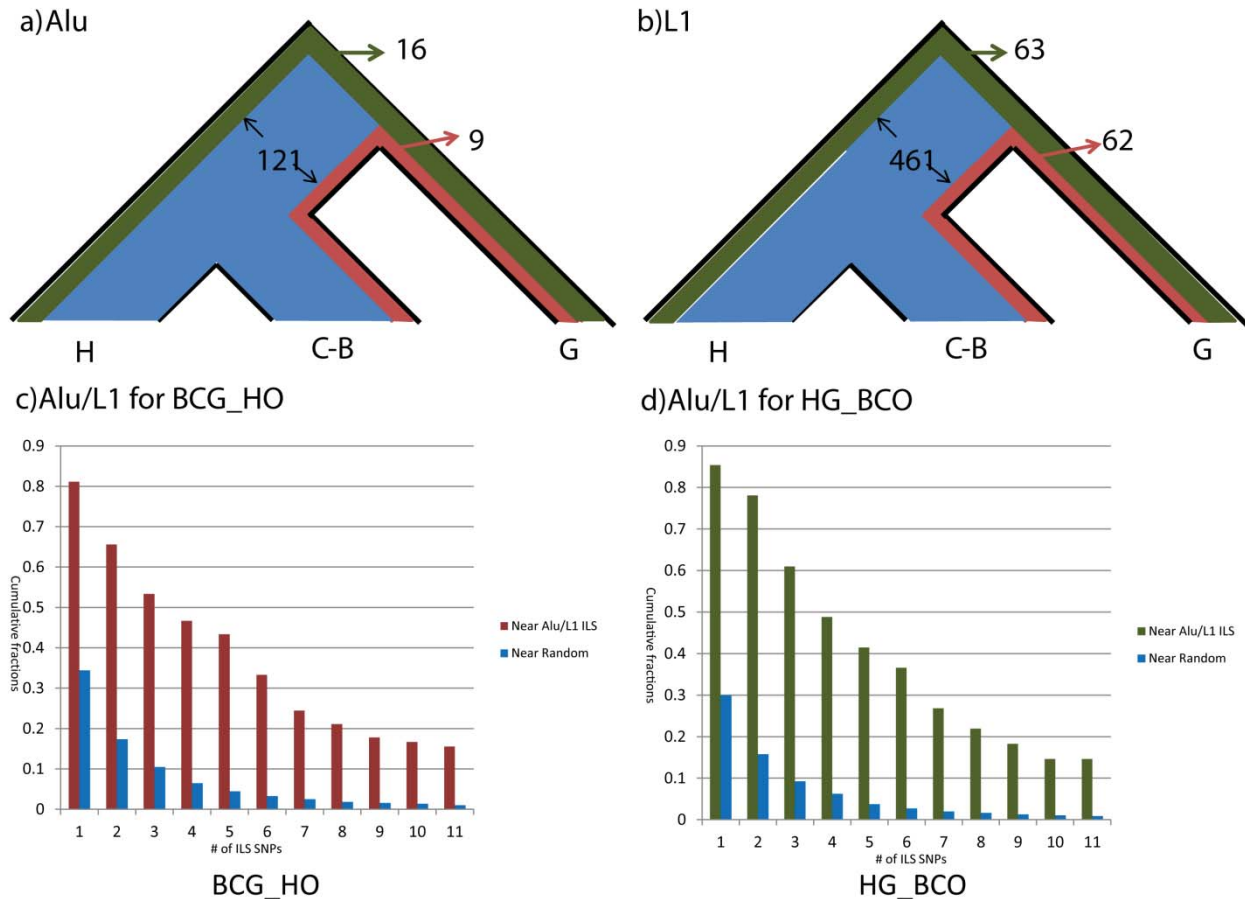


Fig. S20. Total *Alu* ILS insertions common between gorilla (G) and human (H) (green); total *Alu* ILS insertions common between gorilla (G) and chimpanzee-bonobo (C-B) (red) with total *Alu* insertions common between human (H) and chimpanzee-bonobo(C-B) (blue). b. ILS information for *L1* insertions. c. Cumulative distribution of *Alu/L1* ILS events concordant with SNP ILS between for BCG-HO. x-axis shows number of ILS SNPs close to the predicted ILS *Alu/L1* insertion, and y-axis shows the cumulative fraction of events. d. Cumulative distribution of *Alu/L1* ILS events concordant with SNP ILS between GH-BCO.

Section 7 – Inverse correlation of *Alu* and L1 insertions

First, we considered the rate of *Alu* insertions vs. L1 insertions per each branch of evolution. As can be seen in **Fig. S21**, there is a weak inverse correlation between rate of accumulation of *Alu* and L1, where the Pearson correlation is $r=-0.409$ with $p=0.31$.

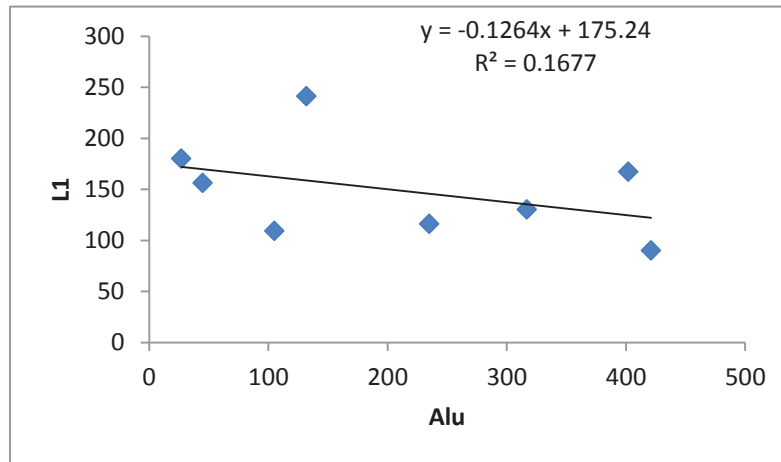


Fig. S21. The inverse correlation between rate of *Alu* and L1 insertions in all branches of GAPE evolution.

However, if we only look at the terminal branches of the evolution (i.e., human, chimpanzee, bonobo, gorilla, and orangutan) the inverse correlation becomes stronger as shown in **Fig. S22** with Pearson correlation of $r=-0.5578$ and $p=0.32$.

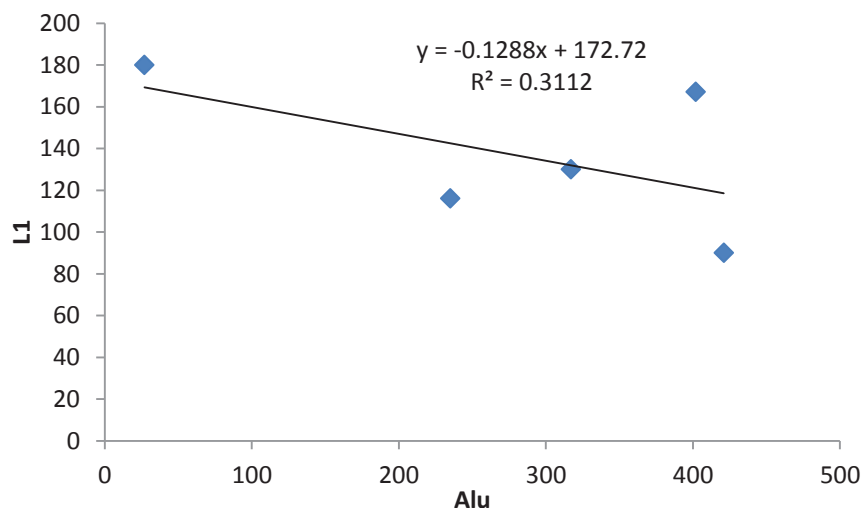


Fig. S22. The inverse correlation between rate of *Alu* and L1 insertions in terminal branches of GAPE evolution.

Section 8 – GC distribution of *Alu* insertions

We also provide the GC distribution for the windows of 50 kbp spanning the loci of *Alu* insertions.

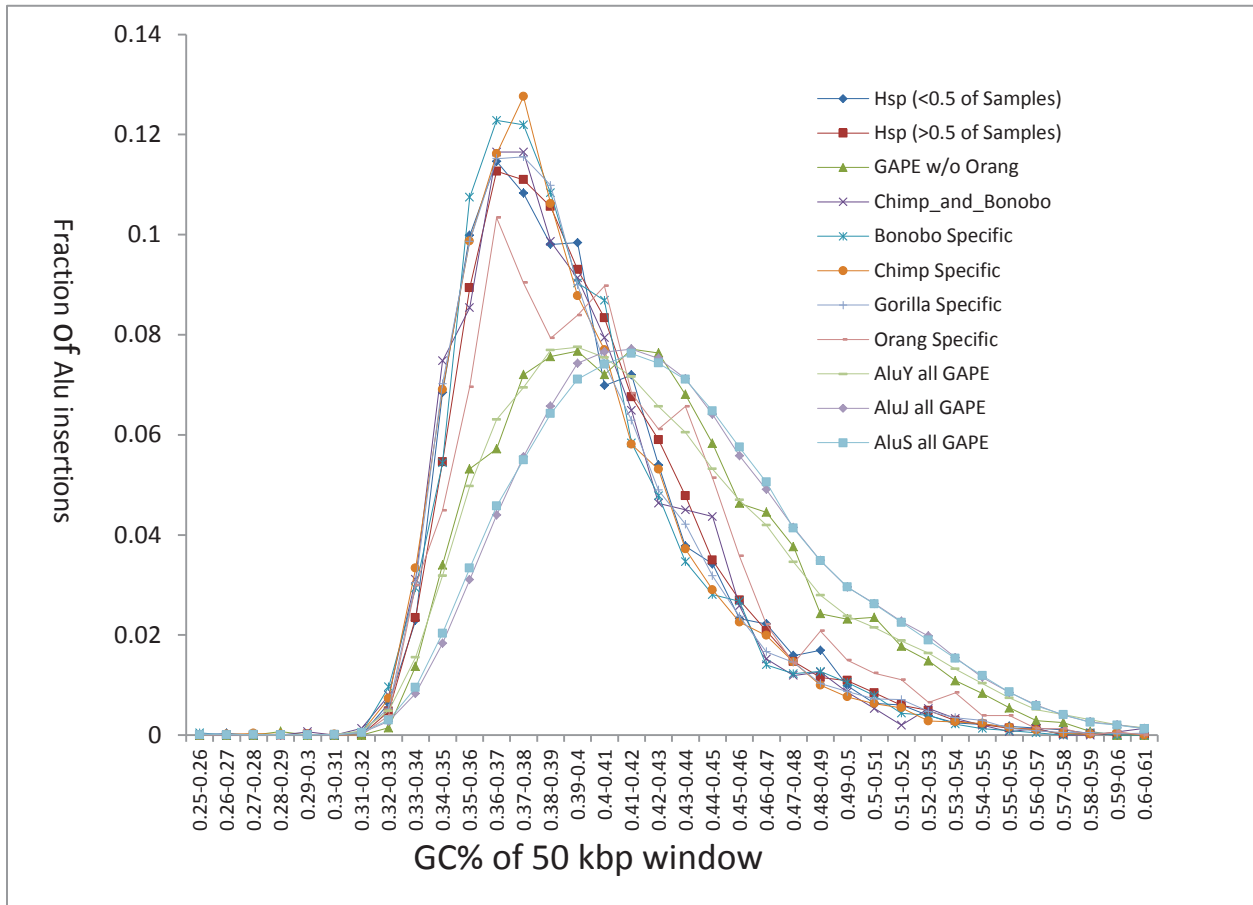


Fig. S23. Distribution of GC ratio of 50 kbp windows of all *Alu* insertions for each lineage. It shows the shift of the distribution from AT-rich regions to GC-rich regions as the insertion becomes older.

References

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
2. Hach F, Hormozdiari F, Alkan C, et al. mrsFAST: A cache-oblivious algorithm for short-read mapping. *Nature methods*. 2010;7(8):576-577.
3. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*. 2009;19(7):1270-1278.
4. Hormozdiari F, Alkan C, Ventura M, et al. Alu repeat discovery and characterization within human genomes. *Genome Res*. 2011;21(6):840-849.
5. Hormozdiari F, Hajirasouliha I, Dao P, et al. Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. *Bioinformatics*. 2010;26(12):i350-i357.
6. Stewart C, Kural D, Strömberg MP, et al. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics*. 2011;7(8):e1002236.
7. Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, Sahinalp SC. Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome Res*. 2011;21(12):2203-2212.
8. Mikkelsen TS, Hillier LDW, Eichler EE, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 2005.
9. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656-664.

10. Hall TA. BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. 1999;41:95-98.

11. Felsenstein J. *Inferring phylogenies*. Vol 2. Sinauer Associates Sunderland; 2004.