

# Supplement: Thousands of human mobile element fragments undergo strong purifying selection near developmental genes

Craig B Lowe<sup>\*</sup>, Gill Bejerano<sup>†\*</sup>, and David Haussler<sup>\* ‡</sup>

<sup>\*</sup>Center for Biomolecular Science and Engineering University of California Santa Cruz, Santa Cruz, CA 95064, USA, and <sup>†</sup>Howard Hughes Medical Institute

Submitted to Proceedings of the National Academy of Sciences of the United States of America

## Supplemental material

Supplemental text, tables, and figures.

### S1 pan-boreoeutherian mobile elements

Mobile elements, as annotated by Repbase [1] and RepeatMasker [2], have a hierarchical naming convention where they are first broken down into classes, then families, and finally subfamilies. In the main text we define “pan-boreoeutherian” subfamilies to be the subfamilies that are present in primates (human, chimp, or rhesus), rodents (rat or mouse), and carnivores (dog). Having a presence in all three of the subtrees means that the mobile element subfamily either predated, or was alive at the time of the boreoeutherian ancestor. These were the only subfamilies used in our study since mobile elements would have to be this old to deposit conserved nonexonic elements (CNEs) that were present in the boreoeutherian ancestor. Figure S1 may aid in visualizing the location of the boreoeutherian ancestor in relation to sequenced extant species.

All these pan-boreoeutherian subfamilies can be grouped at the family and class level to show which types of mobile elements are contributing the most (Table S1). LINEs and SINEs appear to be contributing the majority of the CNEs. This data set can also be shown at the subfamily level and in Table S2 we show the 50 subfamilies that have contributed the most CNEs in our survey. In Table S3 we show the top 50 subfamilies, ranked by how many CNEs they have contributed, in relation to their overall genomic abundance. These subfamilies tend to be older, since this allows for their neutral copies to have drifted far enough from the consensus to be unrecognizable in the extant species.

### S2 formal definitions for enrichment tests

Our first enrichment test assigns exaptations to the closest gene TSS within 1Mb and uses the binomial distribution, for which we must define three parameters:

$x = \text{number of successes}$  This is the number of genes with the given annotation (either a GO term or a pathway name) that were selected by an exaptation.

$n = \text{number of trials}$  This is the number of chances to get a success, so we use the number of exaptations.

$p = \text{probability of success}$  To define the probability of success we divide the number of bases in the genome that closest to a TSS with the given annotation, and divide that number by

the total number of bases in the genome.

Our second enrichment test assigns exaptations to the closest gene TSS within 1Mb and uses the hypergeometric distribution, for which we must define four parameters:

$x = \text{number of successes}$  This is the number of genes with the given annotation that were selected by an exaptation.

$n = \text{number of balls in the hat}$  This is the number of genes that had GO annotation.

$m = \text{number of balls colored for success}$  This is the number of genes that have been assigned the given annotation being tested for enrichment.

$k = \text{the number of draws}$  This is the number of genes that were assigned an exaptation.

Our third enrichment test assigns repeat elements (both those identified as exapted and those not) to gene TSSs. We then “pick” the exapted repeat elements and see if they are enriched for an annotation compared to the repeat element background. For this we use the hypergeometric distribution and need to define four parameters:

$x = \text{number of successes}$  This is the number of repeat instances that were exapted and assigned to a gene with the given annotation.

$n = \text{number of balls in the hat}$  Number of repeat elements.

$m = \text{number of balls colored for success}$  Number of repeat elements that have been assigned to a gene with the given annotation.

$k = \text{the number of draws}$  Number of repeat elements that have been both exapted and assigned to a gene with the given annotation.

### S3 enrichment results for GO and pathways

As described in the main text, the striking enrichment is for terms related to transcription regulation and development. A more specific category that we see quite often is cell-adhesion and its related terms. In Figure S4 through Figure S15 we have shown the most enriched GO terms and colored them to help the reader visualize these strong enrichments. We not only looked at our set as a whole, but also investigated the

---

The authors declare no conflict of interest.

This paper was submitted directly to the PNAS office.

<sup>†</sup>To whom correspondence should be addressed. E-mail: bejerano@stanford.edu. New address: Depts. of Developmental Biology and Computer Science, Stanford University, Stanford, CA 94305, USA

©2006 by The National Academy of Sciences of the USA

enrichment at every level of the repeat taxonomy provided to us: classes, families, and subfamilies.

The pathway annotation is far more sparse than GO, but top enrichments for the entire set are shown in Table S17 and top results for classes, families, and subfamilies are present in Table S18. We used pathway datasets from Biocarta [6], Kegg [7], and Genmapp [8]. The p-values reported are from the test that assumes a uniform distribution over genes and does not allow the same gene to be selected multiple times. We use this test because it might be misleading to say we are enriched for a pathway of 30 genes if 20 exapted elements are all near a single gene in the pathway.

We believe that enrichment for pathways will be a very insightful way to examine sets of cis-regulatory regions in the future, but currently the annotation is too sparse. We suspect that many of these regions, from the same consensus, may drive expression at similar time points. We would not expect the genes they are near to all have a similar function (as specific GO terms usually show), but we would expect many of them to be in the same or related pathways. We followed up on MIRb elements being near all genes known to be in the pathway dealing with the reception of the RELN signal. We looked for conserved transcription factor binding sites (detailed in the main text), and an example of engrailed binding sites is shown in Figure S3.

#### **S4 clouds of exaptations**

During the enrichment tests it became clear that the elements do indeed form large clusters near certain genes (Figure 4, Table S16). We investigated if these clouds of exaptations contained similar sequences or dissimilar sequences. Are these genes exapting many of a certain kind of repeat, or are they looking to grab one of each? We calculated both the entropy and relative entropy of these exaptation clouds. For the relative entropy calculation we used the background distribution of all exaptations identified in this study. In Table S28 we show both the entropies as well as a p-value of getting an entropy that high with the given number of elements in the cloud.

#### **S5 overlap with verified gene regulatory elements**

Our set of conserved nonexonic elements (CNEs) was created with the intention of being putative gene regulatory elements. To investigate if any of these regions have already been experimentally validated as developmental enhancers, we checked to see if there was overlap between any of our elements and those validated in the Enhancer Browser [4].

We found that three of the exapted regions are covered by verified enhancers, but the regions of DNA that were validated were typically 10 times larger than the exaptation, so it is not clear if the exapted region is fully, or even partially responsible for the enhancer activity. The locations of these exaptations and the size of the verified regions is available in Table S19.

#### **S6 overlap with previously unknown transcription start sites**

CAGE sequencing allows the first approximately 20 bases of an mRNA to be sequenced so that the transcription start site

may be identified. When a large set of CAGE tags was published and made publicly available [5] we investigated if our putative cis-regulatory elements could be acting as distal transcription start sites that may be tissue or time specific. This would be a different mechanism than cis-regulation, but in the end it would have a similar effect of driving gene expression at a specific time and/or in a specific tissue.

We utilized both the human and mouse CAGE tags. The mouse tags were first mapped to syntenic locations in the human genome for analysis. All previously known coding regions were filtered out of the CAGE tags so that only the new putative start sites remained. We examined the overlap of the remaining CAGE tags with our set of exapted elements and found 297 exaptations that are overlapped by CAGE tags. This is only an enrichment of 1.3x, which leads us to believe that most of these overlaps are by chance; randomly selecting noncoding regions in the genome would have comparative results. A few of these regions could be functioning as previously unknown transcription start sites, but there is no strong bias towards a certain repeat family or certain tissue type. The number of overlaps can be seen in Table S20.

Tags that could not be mapped uniquely to the genome did not appear in the final set, which biases strongly against mobile elements. We hope that in the future advanced methods will allow experimentalists to deal with interspersed repeats so that the true contribution of these elements may be fully realized.

#### **S7 enrichment for exaptations that have conserved the same section of the consensus**

By comparing the regions of a repeat consensus that were being exapted versus the genomic abundance of that section, we were able to see that certain portions of the repeat are much more likely to be exapted as CNEs (Figure 2, Figure S2). We then investigated if the exapted elements that contributed to each peak may be enriched for a certain GO term or pathway. This would be circumstantial evidence that exaptations from similar sections of the consensus have similar functions. Table S21 and Table S22 show the enrichments for these tests, but no convincing enrichments were found. We look forward to more extensive pathway annotation in the future.

#### **S8 enrichment for groups of exaptations defined by sequence similarity**

Many mobile elements in the genome are chimeras so very similar sequences are often present in different families [9]. For this reason we examined groups of exapted elements where every member of the group has close sequence similarity to all other members in the group.

To do this we did an all-by-all sequence comparison of the exapted elements. This dataset allowed us to create a graph where each exaptation is represented by a node and two nodes (exaptations) are connected by an edge if the sequence alignment between the two is above a given significance threshold. We found the largest fully connected cliques in this graph. A clique is a section of the graph where each node is directly connected to every other node in the section. The nodes in the clique will give us a set of exaptations where each element is highly similar, at the sequence level, to all the other ele-

ments. Table S23 shows the cliques that we discovered and used in further analysis.

We expect that these groups of exaptations may have similar function because of their close sequence identity. To explore this idea we looked to GO and pathway enrichments for these groupings. The results of the GO analysis assuming a uniform distribution over the genome and a uniform distribution over genes can be seen in Table S24 and Table S25. The pathway results for the same null distributions can be seen in Table S26 and Table S27.

## S9 vanishing repeat families

When conducting this survey we faced the problem that a sequence may have been deposited by a mobile element so long ago in the human lineage that we no longer recognize the sequence as coming from a mobile element. Once all members of a mobile element family cease to replicate, it is only a matter of time before the instances decay away beyond recognition and it is no longer evident that the interspersed repeat ever existed. We conducted a simulation to quantify how long ago a repeat family would have needed to cease replicating in order for it to go unnoticed in the extant human genome.

Repeats are often identified because researchers notice that a section of the genome has a number of seemingly paralogous sequences. These sequences are then aligned to each other to generate a consensus sequence, which is used as the most likely sequence of the ancient repeat. This consensus can then be used to iteratively find more elements and refine the consensus. If enough time has gone by then only a few of the sequences will have a significant alignment to each other and the family will not be identified. In a recent paper a repeat was identified in human, only after being found at very high copy number in *Coelacanth* [10]. This SINE, the LF-SINE, had never been reconstructed and annotated in human, even though there is a region in human that has 34 significant alignments to other regions in the human genome. We use the term “pile-up” to describe a region of the genome that has many significant alignments to other sequence in the genome. The largest pile-up of all the LF-SINE instances in human was a pile-up of 34. We use this statistic, that regions seem to need more than 34 paralogous copies in the genome before they will be annotated based on human sequence, to quantify how recently a mobile element must have been alive for us to recognize it in human. Our calculations will give us the expected maximum pile-up in the genome for a repeat family based on the branch length from when it stopped replicating to the extant genome, as well as the number of copies in the genome when replication stopped. If the expected maximum pile-up is under 35 then we label this repeat as having disappeared.

We begin by calculating the chances that two identical sequences will align to each other after they have both undergone a specified distance of independent evolution. This tells us, given that two sequences (repeats) in the genome were identical at a certain time in evolution, what are the chances that they will give significant alignments in the extant genome, providing that they have been under neutral selection. For each branch length from 0 to 1 with a step of 0.2, we did 1,000,000 trials of taking two sequences and decaying each for the specified branch length and recording if they gave a significant alignment. We used the MIR\_Mars

263 base-pair consensus sequence since paralogous alignments must be to the same regions of the sequence to make it apparent that a unit is repeated in the genome. Longer elements, such as LINEs, are more complicated since they may have 20 or more non-overlapping 250 base windows, only one of these needs to have a large pile-up to have the element partially detected, but all would need to have pile-ups to have the element be completely detected. By using a shorter repeat we simplified these issues since the repeat will either be almost entirely found, or not found at all. This simulation gave us the probability that two sequences will align, given that they have both decayed for a specified branch length.

To expand this to a family of  $N$  elements we use a simplifying assumption that all comparisons are independent and we model the density function as a binomial with the probability taken from the simulation and the number of trials being the number of family members minus one,  $N - 1$ . This binomial distribution gives the probability density for a given pile-up in the genome, from a family of size  $N$ , that stopped replicating at a certain point on the human lineage. To know what the probability of this single pile-up being greater than or equal to some value,  $Y$ , we can sum the discrete probability density from  $Y$  to  $N - 1$ . Because we care if any of the  $N$  members of the family have a pile-up greater than  $Y$ , and not only a single instance, we make a second simplifying assumption that alignments between family members are also independent and we can now multiply the probability of a single instance giving us a pile-up of  $Y$  or greater by  $N$  to get the probability of any member of the family giving us a pile-up greater than or equal to  $Y$ .

$$p(\text{maxPileUp} \geq Y) = \min(1, N \cdot \sum_{i=Y}^{N-1} B(i; N-1, p))$$

For branch lengths up to 1 substitution per site and repeat population sizes up to one million, we show in Figure S4 the surface for what the largest expected pile-up in the genome is.

$$\text{largestPileUp} = \underset{k}{\operatorname{argmax}} (N \cdot \sum_{i=k}^{N-1} B(i; N-1, p)) > 0.5$$

If a repeat stopped jumping near the split with dog, there will probably be an instance in the genome that aligns to just about all copies. If the mobile element stopped proliferating around the time of the human opossum split, there will still be clear evidence of the family, but the pile-ups will be much smaller than the original family size. For repeats this old, a researcher will have to reconstruct a consensus and then search again to realize the extent of the repeat’s proliferation. Families that died at the time of the human-chicken speciation are most likely the furthest back that we can currently see. After this, the pile-ups left in the genome will be smaller than 35 and will probably not have been noticed and reconstructed at this time, as evidenced by the LF-SINE not being annotated until a few *Coelacanth* sequences were made public. To annotate mobile elements that died before the human-chicken split, researchers would have to either notice the repeat first in a species with a slower mutation rate, or a significant number of instances would need to be under negative selection to slow their mutation rate.

1. Jurka J (2000) *Trends Genet* 16:418–420.
2. Smit A, Green P (2005) <http://ftp.genome.washington.edu/RM/RepeatMasker.html>.
3. Kamal M, Xie X, Lander ES (2006) *Proc Natl Acad Sci U S A* 103:2740–2745.
4. Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) *Nucleic Acids Res* 35:88–92.
5. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engstrom PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) *Nat Genet* 38:626–635.
6. BioCarta . (2006) <http://www.biocarta.com>.
7. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) *Nucleic Acids Res* 32:277–280.
8. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) *Nat Genet* 31:19–20. Letter t:ttter.
9. Zhi D, Raphael BJ, Price AL, Tang H, Pevzner PA (2006) *Genome Biol* 7:R7.
10. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D (2006) *Nature* 441:87–90.
11. Nishihara H, Terai Y, Okada N (2002) *Mol Biol Evol* 19:1964–1972.
12. Deininger PL, Moran JV, Batzer MA, Kazazian HHJ (2003) *Curr Opin Genet Dev* 13:651–658.

**Table S1. The exaptation of mobile element classes and families**

class	Repeat family	Exapted		Genomic		Genomic/Exapted	
		blocks	bases	blocks	bases	blocks	bases
LINE	CR1	2325	242446	76071	13553443	32.7	55.9
	L2	2052	192636	439841	106691082	214.3	553.8
	L1	1389	145365	662145	309380061	476.7	2128.2
	RTE	500	49763	20225	4114685	40.4	82.6
		6245	629707	1187550	433565705	190.1	688.5
SINE	MIR	3069	280215	617031	96166955	201.0	343.1
	Alu	0	0	5571	873110	-	-
		3069	280215	622537	97038284	202.8	346.2
DNA	MER1.type	315	29503	194991	36545702	619.0	1238.7
	Mariner	80	8574	6507	932209	81.3	108.7
	Tip100	50	4590	25829	6122175	516.5	1333.8
	Tc2	45	5045	7472	1634814	166.0	324.0
	MER2.type	26	2656	25491	8404638	980.4	3164.3
	DNA	21	1645	13796	1858964	656.9	1130.0
	AcHobo	20	1747	18851	3538969	942.5	2025.7
	MER1.type?	4	410	4981	934050	1245.2	2278.1
	hAT	1	61	3097	457999	3097.0	7508.1
	Merlin	0	0	43	17994	-	-
	MuDR	0	0	1534	509074	-	-
	PiggyBac	0	0	466	182189	-	-
		560	54227	298284	60698710	532.6	1119.3
LTR	MaLR	231	25729	219847	73836820	951.7	2869.7
	ERVL	152	16765	112297	39788552	738.7	2373.3
	ERV1	27	2229	52186	21427661	1932.8	9613.1
	ERV	0	0	531	191902	-	-
		407	44483	378055	135112875	928.8	3037.4
Unknown	Unknown	274	34464	1048	212008	3.8	6.1
<b>Total</b>		<b>10402</b>	<b>1035076</b>	<b>2355046</b>	<b>724110376</b>	<b>226.4</b>	<b>699.5</b>

The hierarchical naming scheme of repeat class, family, and subfamily is as defined by RepeatMasker [2] and Repbase [1]. All pan-boreoeutherian mobile elements are grouped here by class and family. The only Alu family is solely comprised of the “Fossil Alu Monomer” subfamily [11], of which dog has a single copy. The unknown category is solely comprised of the MER121 paralog family. For each repeat grouping we first list the number of instances and total number of bases it has contributed to our set of exapted CNEs. We then list, for comparison, the number and total base pair abundance of instances from that family in the genome. Finally, we divide, per instance and per base, the genomic abundance by the exapted abundance, to obtain a “one in” statistic (e.g., one in every 32.7 CR1 instances in the human genome has been identified as exapted by our survey). If orthologous bases are annotated, for example, as an L1 in one species, but as an L2 in another species, these bases will be added to both the L1 and L2 totals, but will appear only once in the total for all LINES. Similarly, for contradictory annotation at the class level. This causes some discrepancy between the totals and the breakdown sums they represent.

**Table S2. The exaptation of mobile element subfamilies**

Repeat Name	Exapted		Genomic		Genomic/Exapted	
	blocks	bases	blocks	bases	blocks	bases
L2	2052	192636	439841	106691082	214.3	553.8
MIRb	1507	141220	334952	52950441	222.2	374.9
L3	1405	152185	55672	10961735	39.6	72
MIR	1035	95799	255893	40300081	247.2	420.6
L3_Mars	699	74807	17741	2544321	25.3	34
L3b	643	68085	10465	1506385	16.2	22.1
L1M5	604	63453	95230	28344661	157.6	446.7
MIR3	579	50391	86987	11397306	150.2	226.1
L4	500	49763	20225	4114685	40.4	82.6
MIR_Mars	493	46250	24437	3348809	49.5	72.4
L1ME4a	474	49721	47303	11096591	99.7	223.1
MIRm	453	38641	60444	6369153	133.4	164.8
MER121	274	34464	1048	212008	3.8	6.1
L1MC	179	19463	28520	8005488	159.3	411.3
THER1_MD	177	15706	27520	3322275	155.4	211.5
L1MC4a	148	15130	35734	12424118	241.4	821.1
L1ME3B	92	10418	31990	11982609	347.7	1150.1
MARNA	78	8431	3693	722941	47.3	85.7
L1M4	74	7608	30362	11855107	410.2	1558.2
L1MD	70	7049	18100	6276365	258.5	890.3
L1MC4	60	7161	33392	11838631	556.5	1653.2
HAL1	60	5652	28567	10436160	476.1	1846.4
MER5A	60	4903	38672	5202703	644.5	1061.1
HAL1b	59	6171	8716	2366506	147.7	383.4
MLT1K	58	5831	20801	4960259	358.6	850.6
L1MEe	57	5213	19253	6882233	337.7	1320.2
L1MEd	53	5547	20897	7592323	394.2	1368.7
MER5B	52	4592	26120	3353832	502.3	730.3
ERVLE	44	5051	12395	6138109	281.7	1215.2
MER117	44	4373	4969	686551	112.9	156.9
Charlie8	43	3844	9520	1646050	221.3	428.2
L1ME3A	34	3393	18730	7230802	550.8	2131
MLT1I	32	3615	12482	2842168	390	786.2
MER102b	30	3044	4868	1086097	162.2	356.7
L1MEc	29	2931	26236	13344450	904.6	4552.8
LTR67	29	2846	7333	1354247	252.8	475.8
L1ME2	28	3262	19419	8691830	693.5	2664.5
MLT1L	28	2771	13327	2917793	475.9	1052.9
L1ME1	28	2729	30950	15867176	1105.3	5814.2
L1M	26	3109	16353	6248253	628.9	2009.7
MLT1J	24	2855	17227	4403127	717.7	1542.2
Kanga1	24	2527	3790	722004	157.9	285.7
MLT1H	21	2747	10942	3173804	521	1155.3
L1M3	21	2097	11962	4631475	569.6	2208.6
L1MC5	20	1628	21522	5957853	1076.1	3659.6
MER113	18	1581	4786	922878	265.8	583.7
L1M2	17	1948	13501	9005300	794.1	4622.8
LTR33	17	1589	9629	2620481	566.4	1649.1
MER103	17	1362	7600	906315	447	665.4
L1MDa	15	1586	13316	7062585	887.7	4453

The 50 subfamilies that contributed the most blocks of exapted bases.

**Table S3. The exaptation of mobile element subfamilies**

Repeat Name	Exapted		Genomic		Genomic/Exapted	
	blocks	bases	blocks	bases	blocks	bases
MER121	274	34464	1048	212008	3.8	6.1
MER57C2	1	72	16	3151	16	43.7
L3b	643	68085	10465	1506385	16.2	22.1
L3_Mars	699	74807	17741	2544321	25.3	34
Kanga1c	5	671	190	76883	38	114.5
L3	1405	152185	55672	10961735	39.6	72
L4	500	49763	20225	4114685	40.4	82.6
MARNA	78	8431	3693	722941	47.3	85.7
MIR_Mars	493	46250	24437	3348809	49.5	72.4
Tigger6b	2	324	100	49159	50	151.7
Charlie6	5	369	271	128113	54.2	347.1
MER102a	8	756	502	132643	62.7	175.4
MER70-int	3	542	190	75974	63.3	140.1
Tigger8	12	1293	911	228041	75.9	176.3
MER45R	3	270	265	140864	88.3	521.7
MER99	3	229	277	98883	92.3	431.8
LTR58	1	50	95	32441	95	648.8
L1ME4a	474	49721	47303	11096591	99.7	223.1
Kanga2_a	14	1484	1472	433115	105.1	291.8
Charlie11	1	58	109	39126	109	674.5
MER117	44	4373	4969	686551	112.9	156.9
MLT1H2-int	1	147	116	115671	116	786.8
FordPrefect	4	464	471	271573	117.7	585.2
LTR68	3	226	369	110614	123	489.4
MIRm	453	38641	60444	6369153	133.4	164.8
LTR69	1	50	140	54630	140	1092.6
HAL1b	59	6171	8716	2366506	147.7	383.4
MIR3	579	50391	86987	11397306	150.2	226.1
LTR52-int	2	318	302	151329	151	475.8
THER1_MD	177	15706	27520	3322275	155.4	211.5
L1M5	604	63453	95230	28344661	157.6	446.7
Kanga1	24	2527	3790	722004	157.9	285.7
L1MC	179	19463	28520	8005488	159.3	411.3
MER102b	30	3044	4868	1086097	162.2	356.7
MER69B	3	270	511	273651	170.3	1013.5
Zaphod2	5	336	871	165551	174.2	492.7
L1M2a1	1	63	179	176275	179	2798
ERV_L	11	1176	2028	843493	184.3	717.2
L2	2052	192636	439841	106691082	214.3	553.8
Charlie8	43	3844	9520	1646050	221.3	428.2
MIRb	1507	141220	334952	52950441	222.2	374.9
MER51-int	1	101	224	136231	224	1348.8
MER91B	7	438	1581	181805	225.8	415
MER97b	1	151	239	76827	239	508.7
L1MC4a	148	15130	35734	12424118	241.4	821.1
LTR65	2	201	484	130221	242	647.8
MIR	1035	95799	255893	40300081	247.2	420.6
Charlie4	8	1180	1999	293395	249.8	248.6
LTR67	29	2846	7333	1354247	252.8	475.8
L1MD	70	7049	18100	6276365	258.5	890.3

The 50 subfamilies that have the most impressive ratios of genomic blocks to exapted blocks.

Top GO enrichment p-values for the set of all exapted regions  
using a uniform null distribution

transcription regulator activity and related terms    
development and related terms    
cell adhesion and related terms  

p-value GO term

1.84E-75	development
1.09E-72	transcription regulator activity
5.45E-58	transcription factor activity
3.52E-55	system development
3.12E-53	nervous system development
9.19E-53	regulation of cellular metabolism
3.73E-52	regulation of transcription, DNA-dependent
9.89E-51	DNA binding
1.17E-50	transcription, DNA-dependent
1.76E-50	regulation of metabolism
2.03E-50	binding
8.56E-49	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
1.41E-48	regulation of transcription
2.92E-46	transcription
8.67E-43	regulation of cellular physiological process
4.43E-42	regulation of physiological process
4.71E-40	regulation of biological process
1.76E-39	nucleic acid binding
1.52E-38	sequence-specific DNA binding
1.55E-37	regulation of cellular process
3.78E-37	organ development
6.10E-34	cell differentiation
7.00E-32	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
9.98E-27	nucleus
1.66E-26	transcription from RNA polymerase II promoter
1.08E-25	cell
1.22E-25	cell part
1.25E-24	protein binding
5.00E-24	transcription factor binding
3.09E-23	cell recognition
1.78E-22	transmembrane receptor protein tyrosine kinase signaling pathway
7.14E-22	neuron recognition
1.74E-21	GPI anchor binding
5.65E-21	enzyme linked receptor protein signaling pathway
1.14E-19	skeletal development
1.17E-19	cellular process
1.45E-19	phosphoinositide binding
2.05E-19	cell development
4.07E-19	morphogenesis
6.46E-18	embryonic development
8.35E-18	hemopoiesis
1.12E-17	hemopoietic or lymphoid organ development
3.34E-17	transcription cofactor activity
1.28E-16	transmembrane receptor protein kinase activity
1.32E-16	transmembrane receptor protein tyrosine kinase activity
1.67E-16	neuron differentiation
2.34E-16	BRE binding
2.34E-16	translation repressor activity, nucleic acid binding
7.26E-16	RNA polymerase II transcription factor activity
2.58E-15	central nervous system development
3.46E-15	translation repressor activity
3.82E-15	neuron development
4.20E-15	osteoblast differentiation
5.40E-15	RNA interference
5.78E-15	negative regulation of cell differentiation
6.17E-15	neurogenesis
9.22E-15	cell glucose homeostasis
1.22E-14	RNA-mediated gene silencing
1.22E-14	RNA-mediated posttranscriptional gene silencing
1.22E-14	posttranscriptional gene silencing
2.06E-14	phospholipid binding
2.23E-14	negative regulation of development
2.51E-14	cell adhesion
3.81E-14	positive regulation of organismal physiological process
5.47E-14	gene silencing

**Table S4.** Most significant p-values for all exapted elements when a uniform null over bases in the genome is used. For an explanation of all GO tests see Supplemental Text S2.



Top GO enrichment p-values for the set of all exapted regions  
using a uniform null distribution where each gene can only be selected once

	transcription regulator activity and related terms	
	development and related terms	
	cell adhesion and related terms	
p-value	GO term	
7.73E-24	development	
1.15E-22	system development	
3.32E-22	nervous system development	
5.22E-19	transcription regulator activity	
2.83E-18	transcription factor activity	
1.68E-16	sequence-specific DNA binding	
4.62E-13	organ development	
5.16E-11	cell adhesion	
6.26E-10	enzyme linked receptor protein signaling pathway	
1.23E-09	central nervous system development	
3.33E-09	ion channel activity	
3.33E-09	channel or pore class transporter activity	
5.51E-09	alpha-type channel activity	
6.48E-09	cation channel activity	
1.65E-08	synapse	
1.72E-08	voltage-gated ion channel activity	
2.44E-08	skeletal development	
2.47E-08	transmembrane receptor protein tyrosine kinase signaling pathway	
6.46E-08	cell-cell adhesion	
1.22E-07	morphogenesis	
1.28E-07	metal ion transport	
1.63E-07	calcium ion binding	
2.09E-07	cell communication	
2.55E-07	plasma membrane part	
2.88E-07	potassium channel activity	
5.02E-07	binding	
5.13E-07	protein binding	
5.50E-07	transmembrane receptor protein kinase activity	
6.99E-07	plasma membrane	
7.02E-07	transcription from RNA polymerase II promoter	
7.28E-07	regulation of biological process	
1.10E-06	voltage-gated potassium channel activity	
1.16E-06	calcium ion transport	
1.45E-06	regulation of transcription, DNA-dependent	
1.46E-06	potassium ion transport	
2.31E-06	regulation of transcription	
2.68E-06	brain development	
3.69E-06	voltage-gated potassium channel complex	
4.75E-06	intrinsic to plasma membrane	
5.39E-06	cation transport	
5.74E-06	transcription, DNA-dependent	
6.06E-06	signal transduction	
6.39E-06	homophilic cell adhesion	
7.21E-06	integral to plasma membrane	
7.35E-06	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
9.65E-06	neurogenesis	
1.01E-05	regulation of cellular process	
1.04E-05	ion transporter activity	
1.05E-05	neuron differentiation	
1.22E-05	extracellular matrix part	
1.36E-05	extracellular matrix	
1.38E-05	ion transport	
1.47E-05	glutamate receptor activity	
1.48E-05	transcription	
2.02E-05	cation transporter activity	
2.28E-05	embryonic development	
2.64E-05	DNA binding	
2.75E-05	regulation of physiological process	
2.92E-05	glutamate-gated ion channel activity	
2.92E-05	ionotropic glutamate receptor activity	
3.05E-05	regulation of cellular metabolism	
3.11E-05	cell differentiation	
3.54E-05	transmembrane receptor protein tyrosine kinase activity	
3.77E-05	membrane	
4.26E-05	cellular morphogenesis	

**Table S5.** Most significant p-values for all exapted elements when a uniform null over genes in the genome is used and each gene can only be selected once. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for the set of all exapted regions  
using the location of all repeat insertions as the null distribution

	transcription regulator activity and related terms	
	development and related terms	
	cell adhesion and related terms	
p-value	GO term	
2.24E-64	transcription regulator activity	
2.33E-60	development	
3.47E-51	transcription factor activity	
8.37E-48	regulation of cellular metabolism	
9.86E-48	DNA binding	
1.48E-47	regulation of transcription, DNA-dependent	
5.19E-46	transcription, DNA-dependent	
1.31E-45	regulation of metabolism	
3.81E-45	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
3.87E-45	system development	
8.27E-45	regulation of transcription	
2.11E-43	nervous system development	
1.58E-42	transcription	
2.01E-37	nucleic acid binding	
3.69E-37	sequence-specific DNA binding	
1.34E-35	regulation of cellular physiological process	
3.39E-34	binding	
4.16E-34	regulation of physiological process	
1.42E-31	regulation of biological process	
1.26E-30	regulation of cellular process	
2.47E-30	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
4.70E-29	cell differentiation	
8.64E-29	organ development	
4.28E-25	nucleus	
3.78E-22	transcription from RNA polymerase II promoter	
1.13E-21	GPI anchor binding	
5.08E-21	cell recognition	
2.62E-20	neuron recognition	
1.15E-19	transcription factor binding	
5.10E-19	phosphoinositide binding	
8.27E-18	transmembrane receptor protein tyrosine kinase signaling pathway	
1.08E-16	cell development	
3.51E-16	enzyme linked receptor protein signaling pathway	
3.80E-16	BRE binding	
3.80E-16	translation repressor activity, nucleic acid binding	
8.45E-16	hemopoiesis	
1.14E-15	hemopoietic or lymphoid organ development	
1.37E-15	cell adhesion	
2.33E-15	translation repressor activity	
5.94E-15	cell	
6.34E-15	cell part	
7.32E-15	skeletal development	
8.15E-15	transmembrane receptor protein kinase activity	
8.19E-15	protein binding	
9.58E-15	transmembrane receptor protein tyrosine kinase activity	
1.06E-14	transcription cofactor activity	
2.84E-14	RNA interference	
4.05E-14	morphogenesis	
4.06E-14	embryonic development	
4.31E-14	phospholipid binding	
5.82E-14	RNA-mediated gene silencing	
5.82E-14	RNA-mediated posttranscriptional gene silencing	
5.82E-14	posttranscriptional gene silencing	
6.95E-14	RNA polymerase II transcription factor activity	
2.23E-13	neuron differentiation	
3.31E-13	cell-cell adhesion	
6.15E-13	neuron development	
6.20E-13	cell glucose homeostasis	
5.14E-12	neurogenesis	
6.70E-12	osteoblast differentiation	
8.78E-12	response to starvation	
1.00E-11	gene silencing	
1.32E-11	brown fat cell differentiation	
1.32E-11	positive regulation of histone acetylation	
1.38E-11	germ cell development	

**Table S6.** Most significant p-values for all exapted elements when the distribution of all mobile elements in the genome is used as the null distribution. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for classes of exapted regions  
using a uniform null distribution

transcription regulator activity and related terms    
development and related terms    
cell adhesion and related terms  

class	p-value	GO term
LINE	7.77E-43	transcription regulator activity
LINE	3.90E-36	development
LINE	1.91E-34	transcription factor activity
LINE	2.48E-34	system development
LINE	1.12E-33	nervous system development
LINE	1.99E-30	DNA binding
LINE	3.58E-30	regulation of transcription, DNA-dependent
LINE	4.64E-30	regulation of cellular metabolism
SINE	2.50E-29	development
LINE	2.63E-29	transcription, DNA-dependent
LINE	5.58E-29	regulation of metabolism
LINE	1.15E-28	regulation of transcription
LINE	1.67E-28	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
LINE	3.27E-27	transcription
LINE	1.15E-25	binding
SINE	3.04E-24	transcription regulator activity
LINE	1.00E-22	regulation of cellular physiological process
LINE	3.89E-22	sequence-specific DNA binding
LINE	4.69E-22	cell
LINE	4.85E-22	cell part
LINE	4.90E-22	regulation of physiological process
LINE	5.16E-22	nucleic acid binding
SINE	1.22E-20	regulation of cellular metabolism
SINE	2.19E-20	transcription factor activity
SINE	3.49E-20	binding
LINE	3.69E-20	organ development
SINE	1.82E-19	regulation of metabolism
LINE	1.96E-19	regulation of biological process
SINE	4.41E-19	regulation of transcription, DNA-dependent
LINE	6.80E-19	regulation of cellular process
LINE	9.71E-19	transmembrane receptor protein tyrosine kinase signaling pathway
SINE	1.06E-18	transcription, DNA-dependent
SINE	1.49E-18	DNA binding
SINE	1.52E-18	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
SINE	6.62E-18	regulation of transcription
LINE	9.16E-18	cell differentiation
LINE	1.45E-17	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
SINE	2.73E-17	transcription
SINE	2.75E-17	nucleic acid binding
SINE	4.61E-17	regulation of physiological process
LINE	4.68E-17	enzyme linked receptor protein signaling pathway
LINE	5.81E-17	nucleus
SINE	6.14E-17	regulation of cellular physiological process
LINE	7.08E-17	cell recognition
SINE	7.60E-17	regulation of biological process
LINE	2.88E-16	GPI anchor binding
SINE	1.12E-15	system development
SINE	1.76E-15	regulation of cellular process
LINE	2.22E-15	transmembrane receptor protein tyrosine kinase activity
SINE	4.52E-15	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
SINE	1.70E-14	nervous system development
LINE	2.26E-14	neuron recognition
LINE	2.47E-14	transmembrane receptor protein kinase activity
SINE	3.04E-14	sequence-specific DNA binding
LINE	4.33E-14	transcription from RNA polymerase II promoter
LINE	1.61E-13	phosphoinositide binding
Unknown	1.77E-13	development
SINE	1.94E-13	hemopoiesis
SINE	2.24E-13	hemopoietic or lymphoid organ development
LINE	6.06E-13	transcription factor binding
SINE	6.57E-13	cell differentiation
LINE	8.72E-13	protein binding
SINE	1.53E-12	organ development
SINE	3.38E-12	embryonic development
SINE	8.83E-12	transcription from RNA polymerase II promoter

**Table S7.** Most significant p-values for classes of exapted elements when a uniform null over bases in the genome is used. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for classes of exapted regions  
using a uniform null distribution where each gene can only be selected once

		transcription regulator activity and related terms	
		development and related terms	
		cell adhesion and related terms	
class	p-value	GO term	
SINE	1.76E-20	development	
LINE	1.79E-20	system development	
LINE	1.93E-20	transcription regulator activity	
LINE	4.11E-20	development	
LINE	6.30E-20	nervous system development	
LINE	3.95E-18	transcription factor activity	
SINE	8.30E-18	system development	
SINE	3.87E-17	nervous system development	
SINE	3.58E-15	transcription factor activity	
SINE	1.15E-14	transcription regulator activity	
LINE	1.52E-14	sequence-specific DNA binding	
SINE	6.50E-14	sequence-specific DNA binding	
LINE	1.05E-13	organ development	
Unknown	1.73E-12	development	
LINE	2.06E-11	cell adhesion	
LINE	3.41E-10	enzyme linked receptor protein signaling pathway	
Unknown	8.74E-10	organ development	
LINE	1.76E-09	central nervous system development	
Unknown	3.18E-09	transcription factor activity	
DNA	3.30E-09	development	
LINE	5.12E-09	protein binding	
DNA	6.58E-09	nervous system development	
SINE	7.15E-09	cell-cell adhesion	
DNA	7.97E-09	system development	
LTR	1.13E-08	development	
SINE	1.58E-08	central nervous system development	
LINE	2.21E-08	transmembrane receptor protein kinase activity	
LINE	2.61E-08	regulation of transcription, DNA-dependent	
Unknown	3.68E-08	transcription regulator activity	
LINE	3.74E-08	regulation of transcription	
LINE	4.46E-08	transmembrane receptor protein tyrosine kinase signaling pathway	
SINE	4.61E-08	cell adhesion	
LINE	6.16E-08	synapse	
LTR	7.41E-08	organ development	
SINE	9.49E-08	organ development	
LINE	1.04E-07	cell-cell adhesion	
LINE	1.25E-07	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
SINE	1.45E-07	ion channel activity	
LINE	1.71E-07	transcription, DNA-dependent	
SINE	1.80E-07	binding	
SINE	2.79E-07	homophilic cell adhesion	
SINE	3.49E-07	morphogenesis	
SINE	3.70E-07	alpha-type channel activity	
SINE	4.37E-07	calcium ion binding	
SINE	4.40E-07	channel or pore class transporter activity	
SINE	4.78E-07	cell communication	
LINE	4.96E-07	regulation of biological process	
Unknown	4.98E-07	cell differentiation	
LINE	5.59E-07	transcription	
Unknown	6.68E-07	sequence-specific DNA binding	
LINE	7.76E-07	regulation of cellular metabolism	
SINE	9.64E-07	brain development	
LINE	1.19E-06	skeletal development	
SINE	1.30E-06	transcription from RNA polymerase II promoter	
SINE	1.43E-06	regulation of transcription, DNA-dependent	
SINE	1.49E-06	regulation of biological process	
SINE	1.62E-06	cation channel activity	
LINE	1.63E-06	ionotropic glutamate receptor activity	
LINE	1.63E-06	glutamate-gated ion channel activity	
LINE	1.72E-06	transmembrane receptor protein tyrosine kinase activity	
LINE	1.76E-06	cell differentiation	
LTR	1.79E-06	homophilic cell adhesion	
LINE	1.91E-06	membrane	
LINE	1.96E-06	morphogenesis	
SINE	1.98E-06	DNA binding	

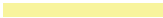
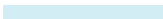

**Table S8.** Most significant p-values for classes of exapted elements when a uniform null over genes is used and each gene can only be selected once. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for classes of exapted regions  
using the location of all repeat insertions as the null distribution

		transcription regulator activity and related terms	
		development and related terms	
		cell adhesion and related terms	
class	p-value	GO term	
LINE	7.82E-41	transcription regulator activity	
LINE	2.17E-33	transcription factor activity	
LINE	7.28E-33	development	
LINE	5.44E-32	system development	
LINE	1.71E-31	nervous system development	
LINE	9.21E-28	regulation of transcription, DNA-dependent	
LINE	1.74E-27	DNA binding	
LINE	4.21E-27	regulation of cellular metabolism	
LINE	8.82E-27	transcription, DNA-dependent	
LINE	1.93E-26	regulation of transcription	
LINE	2.85E-26	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
LINE	4.72E-26	regulation of metabolism	
LINE	6.97E-25	transcription	
LINE	4.52E-23	sequence-specific DNA binding	
LINE	3.25E-19	regulation of cellular physiological process	
LINE	9.58E-19	nucleic acid binding	
LINE	2.54E-18	organ development	
LINE	4.17E-18	regulation of physiological process	
LINE	6.20E-18	GPI anchor binding	
LINE	2.51E-17	cell differentiation	
LINE	5.22E-17	binding	
LINE	9.60E-17	cell recognition	
LINE	2.41E-16	transmembrane receptor protein tyrosine kinase signaling pathway	
SINE	4.84E-16	development	
LINE	5.02E-16	regulation of cellular process	
SINE	7.05E-16	nucleic acid binding	
LINE	7.48E-16	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
LINE	9.15E-16	regulation of biological process	
SINE	3.07E-15	DNA binding	
SINE	5.02E-15	regulation of cellular metabolism	
LINE	6.79E-15	neuron recognition	
LINE	1.12E-14	enzyme linked receptor protein signaling pathway	
LINE	1.89E-14	cell part	
LINE	1.92E-14	cell	
SINE	2.22E-14	transcription regulator activity	
SINE	2.93E-14	regulation of metabolism	
LINE	4.57E-14	transmembrane receptor protein tyrosine kinase activity	
LINE	4.71E-14	phosphoinositide binding	
SINE	5.58E-14	regulation of transcription, DNA-dependent	
LINE	5.97E-14	nucleus	
SINE	9.46E-14	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
SINE	1.15E-13	transcription, DNA-dependent	
LINE	2.79E-13	transmembrane receptor protein kinase activity	
SINE	4.44E-13	regulation of transcription	
LINE	8.75E-13	transcription from RNA polymerase II promoter	
SINE	1.31E-12	transcription	
SINE	2.71E-12	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
SINE	4.25E-12	transcription factor activity	
SINE	2.55E-11	binding	
LINE	4.76E-11	cell adhesion	
LINE	5.30E-11	skeletal development	
LINE	5.78E-11	transcription factor binding	
SINE	6.07E-11	regulation of cellular physiological process	
LINE	1.06E-10	RNA polymerase II transcription factor activity	
SINE	1.49E-10	regulation of physiological process	
LINE	1.58E-10	phospholipid binding	
LINE	2.53E-10	cell development	
SINE	2.66E-10	sequence-specific DNA binding	
LINE	4.78E-10	gene silencing	
SINE	6.26E-10	regulation of biological process	
SINE	7.80E-10	hemopoiesis	
SINE	9.37E-10	hemopoietic or lymphoid organ development	
LINE	1.69E-09	translation repressor activity	
LINE	1.89E-09	translation repressor activity, nucleic acid binding	
LINE	1.89E-09	BRE binding	

**Table S9.** Most significant p-values for classes of exapted elements when the distribution of all mobile elements in the genome is used as the null distribution. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for families of exapted regions  
using a uniform null distribution

transcription regulator activity and related terms   
development and related terms   
cell adhesion and related terms 

family	p-value	GO term
MIR	2.50E-29	development
MIR	3.04E-24	transcription regulator activity
MIR	1.22E-20	regulation of cellular metabolism
MIR	2.19E-20	transcription factor activity
MIR	3.49E-20	binding
MIR	1.82E-19	regulation of metabolism
MIR	4.41E-19	regulation of transcription, DNA-dependent
MIR	1.06E-18	transcription, DNA-dependent
MIR	1.49E-18	DNA binding
MIR	1.52E-18	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
MIR	6.62E-18	regulation of transcription
L2	9.46E-18	development
MIR	2.73E-17	transcription
MIR	2.75E-17	nucleic acid binding
MIR	4.61E-17	regulation of physiological process
CR1	5.36E-17	transcription regulator activity
MIR	6.14E-17	regulation of cellular physiological process
MIR	7.60E-17	regulation of biological process
CR1	1.80E-16	development
MIR	1.12E-15	system development
CR1	1.55E-15	transcription factor activity
MIR	1.76E-15	regulation of cellular process
MIR	4.52E-15	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L2	5.85E-15	system development
MIR	1.70E-14	nervous system development
CR1	2.26E-14	sequence-specific DNA binding
L2	2.99E-14	nervous system development
MIR	3.04E-14	sequence-specific DNA binding
CR1	1.19E-13	system development
Unknown	1.77E-13	development
MIR	1.94E-13	hemopoiesis
MIR	2.24E-13	hemopoietic or lymphoid organ development
CR1	2.65E-13	nervous system development
MIR	6.57E-13	cell differentiation
L1	1.00E-12	transcription regulator activity
MIR	1.53E-12	organ development
CR1	1.57E-12	DNA binding
MIR	3.38E-12	embryonic development
CR1	6.53E-12	binding
MIR	8.83E-12	transcription from RNA polymerase II promoter
L2	1.17E-11	transcription regulator activity
L1	3.61E-11	regulation of transcription, DNA-dependent
L1	4.02E-11	transcription, DNA-dependent
CR1	4.92E-11	regulation of transcription, DNA-dependent
CR1	5.74E-11	transmembrane receptor protein tyrosine kinase activity
L2	6.27E-11	transcription factor activity
L1	8.02E-11	regulation of cellular metabolism
L1	9.76E-11	regulation of metabolism
CR1	1.12E-10	transcription, DNA-dependent
CR1	1.39E-10	regulation of cellular metabolism
L2	1.57E-10	DNA binding
CR1	1.60E-10	transmembrane receptor protein kinase activity
CR1	1.65E-10	regulation of transcription
L1	1.71E-10	transcription
L1	1.84E-10	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L1	1.88E-10	regulation of transcription
CR1	2.19E-10	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
MIR	2.77E-10	protein binding
L1	3.70E-10	transmembrane receptor protein tyrosine kinase signaling pathway
CR1	3.97E-10	regulation of metabolism
CR1	4.19E-10	nucleic acid binding
MIR	5.14E-10	nucleus
L1	7.89E-10	binding
MIR	8.02E-10	osteoblast differentiation
Unknown	8.32E-10	organ development

**Table S10.** Most significant p-values for families of exapted elements when a uniform null over bases in the genome is used. For an explanation of all GO tests see the Supplemental Text S2.

Top GO enrichment p-values for families of exapted regions  
using a uniform null distribution where each gene can only be selected once

		transcription regulator activity and related terms	
		development and related terms	
		cell adhesion and related terms	
family	p-value	GO term	
MIR	1.76E-20	development	
L2	1.88E-20	system development	
L2	1.22E-19	nervous system development	
CR1	7.40E-18	development	
MIR	8.30E-18	system development	
MIR	3.87E-17	nervous system development	
L2	8.95E-16	development	
CR1	1.40E-15	system development	
CR1	2.59E-15	nervous system development	
MIR	3.58E-15	transcription factor activity	
MIR	1.15E-14	transcription regulator activity	
MIR	6.50E-14	sequence-specific DNA binding	
L1	1.41E-13	transcription regulator activity	
CR1	5.74E-13	transcription regulator activity	
Unknown	1.73E-12	development	
L1	4.09E-12	nervous system development	
CR1	5.41E-12	sequence-specific DNA binding	
L1	5.63E-12	system development	
L2	1.66E-11	transcription regulator activity	
L1	1.86E-11	transcription factor activity	
L2	1.96E-11	transcription factor activity	
CR1	2.48E-11	transcription factor activity	
CR1	3.35E-10	organ development	
CR1	3.61E-10	cell adhesion	
Unknown	8.74E-10	organ development	
Unknown	3.18E-09	transcription factor activity	
CR1	4.35E-09	cell differentiation	
MIR	7.15E-09	cell-cell adhesion	
L1	1.29E-08	organ development	
MIR	1.58E-08	central nervous system development	
L2	1.85E-08	sequence-specific DNA binding	
L2	2.17E-08	central nervous system development	
L1	2.75E-08	cell adhesion	
Unknown	3.68E-08	transcription regulator activity	
L1	4.28E-08	cell-cell adhesion	
MIR	4.61E-08	cell adhesion	
L1	6.05E-08	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
L1	6.19E-08	regulation of cellular metabolism	
CR1	6.86E-08	transmembrane receptor protein kinase activity	
L1	6.87E-08	development	
L1	6.97E-08	regulation of transcription, DNA-dependent	
L1	7.54E-08	regulation of transcription	
L2	8.85E-08	neuron differentiation	
MIR	9.49E-08	organ development	
CR1	9.64E-08	cell-cell adhesion	
L1	9.83E-08	transcription, DNA-dependent	
L1	9.88E-08	transcription	
RTE	1.17E-07	transcription regulator activity	
L1	1.35E-07	sequence-specific DNA binding	
L2	1.44E-07	organ development	
MIR	1.45E-07	ion channel activity	
MIR	1.80E-07	binding	
CR1	1.96E-07	transmembrane receptor protein tyrosine kinase activity	
MIR	2.79E-07	homophilic cell adhesion	
L1	2.81E-07	regulation of metabolism	
L1	3.31E-07	enzyme linked receptor protein signaling pathway	
MIR	3.49E-07	morphogenesis	
MIR	3.70E-07	alpha-type channel activity	
L2	3.81E-07	neurogenesis	
CR1	3.88E-07	homophilic cell adhesion	
MIR	4.37E-07	calcium ion binding	
MIR	4.40E-07	channel or pore class transporter activity	
MIR	4.78E-07	cell communication	
CR1	4.88E-07	cell	
Unknown	4.98E-07	cell differentiation	

**Table S11.** Most significant p-values for families of exapted elements when a uniform null over genes is used and each gene can only be selected once. For an explanation of all GO tests see Supplemental Text S2.



Top GO enrichment p-values for families of exapted regions  
using the location of all repeat insertions as the null distribution

family	p-value	GO term
MIR	5.19E-16	nucleic acid binding
MIR	7.85E-16	development
MIR	2.56E-15	DNA binding
MIR	4.35E-15	regulation of cellular metabolism
MIR	2.37E-14	transcription regulator activity
MIR	2.50E-14	regulation of metabolism
MIR	5.07E-14	regulation of transcription, DNA-dependent
MIR	8.25E-14	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
MIR	1.02E-13	transcription, DNA-dependent
L1	3.42E-13	transcription regulator activity
MIR	3.95E-13	regulation of transcription
CR1	7.96E-13	transcription regulator activity
MIR	1.14E-12	transcription
MIR	2.01E-12	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L1	3.64E-12	nervous system development
MIR	5.15E-12	transcription factor activity
CR1	5.45E-12	regulation of transcription, DNA-dependent
CR1	5.99E-12	DNA binding
L1	7.58E-12	system development
CR1	9.88E-12	sequence-specific DNA binding
CR1	1.30E-11	transcription, DNA-dependent
L2	1.72E-11	development
MIR	2.09E-11	binding
CR1	2.09E-11	regulation of transcription
CR1	2.27E-11	transcription factor activity
CR1	3.24E-11	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
CR1	4.31E-11	regulation of cellular metabolism
CR1	5.24E-11	nucleic acid binding
MIR	5.49E-11	regulation of cellular physiological process
CR1	8.72E-11	transcription
CR1	1.09E-10	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L2	1.38E-10	system development
MIR	1.40E-10	regulation of physiological process
CR1	1.88E-10	regulation of metabolism
L1	2.09E-10	regulation of transcription, DNA-dependent
MIR	3.19E-10	sequence-specific DNA binding
L1	3.36E-10	transcription, DNA-dependent
L2	3.61E-10	nervous system development
MIR	6.12E-10	regulation of biological process
L1	7.52E-10	regulation of cellular metabolism
MIR	8.93E-10	hemopoiesis
L2	9.17E-10	DNA binding
MIR	1.07E-09	hemopoietic or lymphoid organ development
L1	1.26E-09	regulation of metabolism
L1	1.37E-09	transcription factor activity
L2	1.67E-09	transcription regulator activity
L1	1.73E-09	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L1	1.76E-09	regulation of transcription
MIR	2.05E-09	regulation of cellular process
L1	2.29E-09	transcription
L1	2.92E-09	transmembrane receptor protein tyrosine kinase signaling pathway
L1	2.96E-09	enzyme linked receptor protein signaling pathway
MIR	5.64E-09	system development
L2	7.09E-09	transcription factor activity
MIR	7.69E-09	embryonic development
MIR	7.94E-09	nucleus
MIR	1.75E-08	cell-cell adhesion
L1	2.52E-08	cell part
CR1	2.69E-08	development
L1	2.76E-08	cell
L2	2.92E-08	regulation of transcription, DNA-dependent
MIR	3.96E-08	nervous system development
L2	5.86E-08	regulation of transcription
MIR	5.91E-08	cell differentiation
L2	7.54E-08	transcription, DNA-dependent

**Table S12.** Most significant p-values for families of exapted elements when the distribution of all mobile elements in the genome is used as the null distribution. For an explanation of all GO tests see Supplemental Text S2.



Top GO enrichment p-values for subfamilies of exapted regions  
using a uniform null distribution

subfamily	p-value	GO term	
MIRb	8.53E-18	development	
L2	9.46E-18	development	
MIRb	3.27E-15	transcription regulator activity	
L2	5.85E-15	system development	
L2	2.99E-14	nervous system development	
MER121	1.77E-13	development	
MIRb	7.32E-13	transcription factor activity	
L3 Mars	9.38E-12	transcription regulator activity	
MIRb	9.64E-12	system development	
L2	1.17E-11	transcription regulator activity	
L3 Mars	1.37E-11	sequence-specific DNA binding	
MIRb	3.05E-11	regulation of transcription, DNA-dependent	
MIRb	6.00E-11	transcription, DNA-dependent	
L2	6.27E-11	transcription factor activity	
MIRb	6.57E-11	nervous system development	
MIR	7.20E-11	transcription regulator activity	
MIRb	7.21E-11	regulation of cellular metabolism	
L3	1.45E-10	development	
L2	1.57E-10	DNA binding	
L3 Mars	1.71E-10	transcription factor activity	
MIRb	1.72E-10	regulation of metabolism	
MIR	2.65E-10	development	
L3	2.91E-10	nervous system development	
L3	2.97E-10	system development	
MIR	3.65E-10	binding	
MER121	8.32E-10	organ development	
MIRb	8.37E-10	regulation of transcription	
MIRb	9.01E-10	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
MIRb	2.04E-09	transcription	
MIRb	2.22E-09	transcription from RNA polymerase II promoter	
MIRb	3.10E-09	DNA binding	
MIR	3.41E-09	transcription factor activity	
L2	4.90E-09	neuron recognition	
MER121	5.32E-09	growth factor activity	
L2	6.27E-09	osteoblast differentiation	
MIRb	7.58E-09	regulation of physiological process	
L3	8.18E-09	transmembrane receptor protein kinase activity	
L4	8.89E-09	transcription regulator activity	
L2	9.33E-09	regulation of transcription, DNA-dependent	
MIRb	9.65E-09	regulation of biological process	
MIR	1.15E-08	nucleic acid binding	
MIRb	1.60E-08	regulation of cellular physiological process	
MIR	1.66E-08	sequence-specific DNA binding	
L2	1.74E-08	regulation of transcription	
MIRb	1.80E-08	embryonic epithelial tube formation	
MIRb	1.80E-08	neural plate morphogenesis	
MIRb	1.80E-08	neural tube formation	
MIRb	1.80E-08	neural tube closure	
MIRb	1.80E-08	primary neural tube formation	
MIRb	1.80E-08	morphogenesis of embryonic epithelium	
L3	1.97E-08	transmembrane receptor protein tyrosine kinase activity	
L2	2.06E-08	regulation of cellular metabolism	
L2	2.26E-08	transcription, DNA-dependent	
L2	2.61E-08	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	
MIRb	2.90E-08	regulation of cellular process	
MER121	3.12E-08	morphogenesis	
L2	3.34E-08	cell glucose homeostasis	
MIRb	3.41E-08	sequence-specific DNA binding	
L2	4.03E-08	cell differentiation	
MIRb	4.37E-08	GPI anchor binding	
L2	4.57E-08	organ development	
MIRb	4.67E-08	binding	
L2	4.71E-08	regulation of metabolism	
MIRb	5.01E-08	neuron recognition	
MIR	5.59E-08	regulation of cellular metabolism	

**Table S13.** Most significant p-values for subfamilies of exapted elements when a uniform null over bases in the genome is used. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for subfamilies of exapted regions  
using a uniform null distribution where each gene can only be selected once

		transcription regulator activity and related terms	
		development and related terms	
		cell adhesion and related terms	
subfamily	p-value	GO term	
L2	1.88E-20	system development	
L2	1.22E-19	nervous system development	
MIRb	2.83E-19	development	
MIRb	3.22E-17	system development	
MIRb	2.33E-16	nervous system development	
L3	6.15E-16	development	
L2	8.95E-16	development	
MIR	9.71E-15	development	
MIR	8.07E-14	transcription regulator activity	
L3	1.91E-13	system development	
MIR	2.84E-13	transcription factor activity	
L3	4.15E-13	nervous system development	
L3 Mars	1.35E-12	transcription regulator activity	
MER121	1.73E-12	development	
MIR	3.58E-12	sequence-specific DNA binding	
MIRb	1.04E-11	sequence-specific DNA binding	
L2	1.66E-11	transcription regulator activity	
L2	1.96E-11	transcription factor activity	
L3 Mars	4.49E-11	sequence-specific DNA binding	
L3	5.85E-11	cell adhesion	
L3 Mars	1.19E-10	development	
MIRb	1.75E-10	transcription regulator activity	
L3 Mars	1.89E-10	transcription factor activity	
MIRb	1.95E-10	transcription factor activity	
MER121	8.74E-10	organ development	
MIR	1.20E-09	organ development	
L3	1.52E-09	transcription regulator activity	
MIR	1.65E-09	binding	
MIR Mars	1.85E-09	system development	
L3 Mars	2.51E-09	system development	
MER121	3.18E-09	transcription factor activity	
MIR Mars	5.34E-09	nervous system development	
L3 Mars	6.36E-09	nervous system development	
MIRm	6.57E-09	system development	
L3	9.55E-09	cell differentiation	
MIR3	9.63E-09	development	
L3	1.12E-08	transmembrane receptor protein kinase activity	
L3	1.18E-08	organ development	
L3b	1.25E-08	sequence-specific DNA binding	
L1ME4a	1.36E-08	nervous system development	
L1ME4a	1.62E-08	system development	
L2	1.85E-08	sequence-specific DNA binding	
MIRm	1.90E-08	nervous system development	
L2	2.17E-08	central nervous system development	
L1M5	2.81E-08	organ development	
MIRb	3.60E-08	cell-cell adhesion	
L3 Mars	3.67E-08	regulation of transcription, DNA-dependent	
MER121	3.68E-08	transcription regulator activity	
L3	3.75E-08	transcription factor activity	
MIRb	4.92E-08	homophilic cell adhesion	
L3b	6.77E-08	development	
MIR Mars	6.90E-08	organ development	
L3 Mars	8.14E-08	transcription, DNA-dependent	
L3	8.42E-08	transmembrane receptor protein tyrosine kinase activity	
L2	8.85E-08	neuron differentiation	
MIRb	1.15E-07	central nervous system development	
L4	1.17E-07	transcription regulator activity	
L1M5	1.19E-07	nervous system development	
L3 Mars	1.27E-07	regulation of transcription	
L3	1.40E-07	cell-cell adhesion	
L1M5	1.41E-07	system development	
MIR Mars	1.42E-07	development	
L2	1.44E-07	organ development	
MIRb	1.53E-07	regulation of development	
MIR	1.59E-07	system development	

**Table S14.** Most significant p-values for subfamilies of exapted elements when a uniform null over genes is used and each gene can only be selected once. For an explanation of all GO tests see Supplemental Text S2.

Top GO enrichment p-values for subfamilies of exapted regions  
using the location of all repeat insertions as the null distribution

subfamily	p-value	GO term
		transcription regulator activity and related terms
		development and related terms
		cell adhesion and related terms
L2	1.72E-11	development
L2	1.38E-10	system development
L2	3.61E-10	nervous system development
MIRb	4.85E-10	development
L2	9.17E-10	DNA binding
L3 Mars	1.61E-09	sequence-specific DNA binding
L2	1.67E-09	transcription regulator activity
MIRb	2.08E-09	transcription regulator activity
L2	7.09E-09	transcription factor activity
MIR	7.42E-09	nucleic acid binding
L3 Mars	8.35E-09	transcription regulator activity
MIRb	2.34E-08	regulation of transcription, DNA-dependent
L2	2.92E-08	regulation of transcription, DNA-dependent
MIRb	3.74E-08	transcription, DNA-dependent
MIRb	4.38E-08	transcription factor activity
MIR	5.30E-08	transcription regulator activity
L2	5.86E-08	regulation of transcription
MIRb	6.73E-08	regulation of cellular metabolism
L2	7.54E-08	transcription, DNA-dependent
L2	8.43E-08	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
MIRb	8.96E-08	DNA binding
L3 Mars	8.98E-08	regulation of transcription, DNA-dependent
MIRb	9.84E-08	regulation of metabolism
L3 Mars	1.08E-07	transcription factor activity
MIRb	1.23E-07	GPI anchor binding
L3 Mars	1.26E-07	transcription, DNA-dependent
L4	1.41E-07	transcription regulator activity
L2	1.73E-07	neuron recognition
MIRb	1.80E-07	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L2	1.88E-07	regulation of cellular metabolism
MIR	1.95E-07	DNA binding
L3 Mars	2.11E-07	regulation of transcription
MIRb	2.13E-07	regulation of transcription
L2	2.23E-07	transcription
MIRb	2.42E-07	embryonic epithelial tube formation
MIRb	2.42E-07	neural plate morphogenesis
MIRb	2.42E-07	neural tube formation
MIRb	2.42E-07	neural tube closure
MIRb	2.42E-07	primary neural tube formation
MIRb	2.42E-07	morphogenesis of embryonic epithelium
L2	2.65E-07	regulation of metabolism
L3 Mars	2.81E-07	regulation of cellular metabolism
MIR	2.91E-07	nucleus
L3 Mars	3.57E-07	transcription
MIRb	3.98E-07	transcription
MIR	3.99E-07	regulation of cellular metabolism
MIRb	4.55E-07	system development
MIR	4.79E-07	sequence-specific DNA binding
L3 Mars	4.91E-07	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
MIR	7.00E-07	binding
HAL1b	8.31E-07	regulation of transcription, DNA-dependent
MIR	8.50E-07	regulation of metabolism
MIRb	9.86E-07	neuron recognition
HAL1b	1.02E-06	transcription, DNA-dependent
MIRb	1.07E-06	phosphoinositide binding
MIR	1.14E-06	transcription factor activity
L2	1.54E-06	cell recognition
MIR	1.56E-06	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
L3 Mars	1.56E-06	regulation of metabolism
L3	1.80E-06	GPI anchor binding
MIRb	1.87E-06	nervous system development
HAL1b	1.94E-06	transcription regulator activity
L2	1.97E-06	cell glucose homeostasis
MIRb	2.00E-06	cell-cell adhesion
HAL1b	2.21E-06	regulation of transcription

**Table S15.** Most significant p-values for subfamilies of exapted elements when the distribution of all mobile elements in the genome is used as the null distribution. For an explanation of all GO tests see Supplemental Text S2.

**Table S16. The largest clusters of exapted elements and a nearby gene they may regulate**

Chrom	#Elements	Region (Mb)	Gene	Gene function
11	44	1.57	ODZ4	signaling and transcription regulation during development
11	40	1.32	HNT	cell adhesion and neurite outgrowth
18	37	1.24	BRUNOL4	developmental splicing regulation
3	35	1.07	EPHB1	guidance receptor for neuronal connectivity
2	32	1.04	Unknown	Unknown
9	31	1.00	FCMD	migration and assembly of neurons
9	31	1.00	CDK5RAP2	regulation of neuronal differentiation
13	31	1.00	DIAPH3	cellular motility, adhesion, and cytokinesis
10	30	1.00	CRTAC1	cell adhesion

This table lists chromosomal location and abundance of the densest 1Mb locales in the human genome for surveyed CNE exaptations. These regions most often overlap large gene deserts (Figure 4), presumably there to harbor the regulatory regions governing the expression of a nearby, most often developmental gene. The putative target genes are often involved in neuronal development, which is clearly seen in this short list.

**Table S17. Enrichment of all exaptations for annotated pathways**

P-value	Pathway
4.57611e-05	Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy pathway
0.00010284	Wnt Signaling Pathway
0.000109369	ALK in cardiac myocytes pathway
0.000571029	G Protein Signaling Pathways
0.000952423	Reelin reception
0.0016645	Corticosteroids and cardioprotection pathway
0.0016645	Effects of calcineurin in Keratinocyte Differentiation pathway
0.00204904	Chondroitin / Heparan sulfate biosynthesis
0.00256153	Reelin Signaling Pathway pathway
0.0034129	Calcium Channels
0.00349215	Signaling Pathway from G-Protein Families pathway
0.00368649	Nuclear Receptors
0.00398588	Cell Cycle G1 S Check Point pathway
0.0069549	Neuropeptides VIP and PACAP inhibit the apoptosis of activated T cells pathway
0.0074094	Pertussis toxin-insensitive CCR5 Signaling in Macrophage pathway
0.00752601	Control of skeletal myogenesis by HDAC and calcium calmodulin-dependent kinase CaMK pathway
0.010302	Cell to Cell Adhesion Signaling pathway
0.0121154	CXCR4 Signaling Pathway pathway
0.0125916	TGF Beta Signaling Pathway
0.0131346	Fc Epsilon Receptor I Signaling in Mast Cells pathway
0.0131346	Signaling of Hepatocyte Growth Factor Receptor pathway
0.0163225	Ca Calmodulin-dependent Protein Kinase Activation pathway
0.0188365	Function of SLRP in Bone An Integrated View pathway
0.0210557	nos1Pathway pathway
0.0212401	Regulation of Spermatogenesis by CREM pathway
0.0217451	WNT Signaling Pathway pathway
0.0217918	Erk1 Erk2 Mapk Signaling pathway pathway
0.0282009	Galactose metabolism
0.0306238	NFAT and Hypertrophy of the heart Transcription in the broken heart pathway
0.0308757	Alpha-synuclein and Parkin-mediated proteolysis in Parkinson's disease pathway
0.0346002	O-Glycans biosynthesis
0.0347328	The IGF-1 Receptor and Longevity pathway
0.0358894	Inactivation of Gsk3 by AKT causes accumulation of b-catenin in Alveolar Macrophages pathway
0.0364711	Role of Tob in T-cell activation pathway
0.036641	Regulation of BAD phosphorylation pathway
0.036641	Regulation of PGC-1a pathway
0.036641	Monoamine GPCRs
0.0377373	p38 MAPK Signaling Pathway pathway
0.0394605	MAPKinase Signaling Pathway pathway
0.0402183	Bioactive Peptide Induced Signaling Pathway pathway
0.0445893	Gamma-aminobutyric Acid Receptor Life Cycle pathway
0.0448973	Actions of Nitric Oxide in the Heart pathway
0.045102	T Cell Receptor Signaling Pathway pathway
0.0478432	Inhibition of Cellular Proliferation by Gleevec pathway

This test used a uniform null over genes and each gene could be selected at most once. The p-values will show if our whole set of exaptations is enriched for a certain pathway that has been annotated.

**Table S18. Enrichment of classes, families, and subfamilies of exapted elements for annotated pathways**

level	name	p-value	Pathway
subfamily	MIRm	3.58086e-06	Signaling Pathway from G-Protein Families pathway
family	L1	4.44207e-06	Reelin reception
subfamily	MIRb	6.7485e-06	Reelin reception
class	LINE	2.73547e-05	Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy pathway
subfamily	L1MC4_3endX	4.37259e-05	Signaling of Hepatocyte Growth Factor Receptor pathway
class	SINE	4.94168e-05	G Protein Signaling Pathways
family	MIR	4.94168e-05	G Protein Signaling Pathways
class	SINE	5.76333e-05	Reelin reception
family	MIR	5.76333e-05	Reelin reception
class	SINE	6.30211e-05	Wnt Signaling Pathway
family	MIR	6.30211e-05	Wnt Signaling Pathway
subfamily	MLT1K	0.00010579	ALK in cardiac myocytes pathway
subfamily	MIRm	0.000108228	Neuropeptides VIP and PACAP inhibit the apoptosis of activated T cells pathway
family	L1	0.000139029	Reelin Signaling Pathway pathway
subfamily	MLT1K	0.000158219	p38 MAPK Signaling Pathway pathway
subfamily	MIRm	0.000190758	Actions of Nitric Oxide in the Heart pathway
subfamily	MIRb	0.000208574	Reelin Signaling Pathway pathway
subfamily	MIRm	0.000282139	Endocytotic role of NDK Phosphins and Dynamin pathway
class	LINE	0.000314906	Reelin reception
class	LTR	0.000328462	ALK in cardiac myocytes pathway
subfamily	L1MC	0.000339889	Reelin reception
class	LINE	0.000342128	ALK in cardiac myocytes pathway
family	L2	0.000371015	ALK in cardiac myocytes pathway
subfamily	L2	0.000371015	ALK in cardiac myocytes pathway
family	MaLR	0.000401576	ALK in cardiac myocytes pathway
class	SINE	0.000461285	Signaling Pathway from G-Protein Families pathway
family	MIR	0.000461285	Signaling Pathway from G-Protein Families pathway
family	L1	0.0004647	Signal Dependent Regulation of Myogenesis by Corepressor MITR pathway
family	L2	0.00046567	Chondroitin / Heparan sulfate biosynthesis
subfamily	L2	0.00046567	Chondroitin / Heparan sulfate biosynthesis
class	LINE	0.000470695	Chondroitin / Heparan sulfate biosynthesis
subfamily	L1ME3	0.000486647	Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa alpha pathway
class	SINE	0.00050867	Effects of calcineurin in Keratinocyte Differentiation pathway
family	MIR	0.00050867	Effects of calcineurin in Keratinocyte Differentiation pathway
subfamily	MIR	0.000522392	Effects of calcineurin in Keratinocyte Differentiation pathway
subfamily	MIRm	0.000563079	Fc Epsilon Receptor I Signaling in Mast Cells pathway
subfamily	L1ME4a	0.000582401	GPCRs Class B Secretin-like
subfamily	MIRm	0.000611983	G Protein Signaling Pathways
subfamily	HAL1	0.00064503	p53 Signaling Pathway pathway
subfamily	MIRm	0.000666856	nos1Pathway pathway
class	LINE	0.000696149	Reelin Signaling Pathway pathway
subfamily	HAL1	0.000735985	Cell Cycle Control in G1/S Phase
subfamily	MLT1E3	0.000793611	GATA3 participate in activating the Th2 cytokine genes expression pathway
subfamily	MLT2C1	0.000793611	GATA3 participate in activating the Th2 cytokine genes expression pathway
subfamily	L1M1	0.000798139	GPCRs Class A Rhodopsin-like
family	CR1	0.000807272	Nuclear Receptors
family	MaLR	0.00085304	TGF beta signaling pathway pathway
family	MaLR	0.000889185	TGF Beta Signaling Pathway
subfamily	L1M5	0.000906048	Signal Dependent Regulation of Myogenesis by Corepressor MITR pathway
subfamily	MER70B	0.000942414	Blood Clotting Cascade

This test used a uniform null over genes and each gene could be selected at most once. The p-values will show if any class, family, or subfamily of our set of exaptations is enriched for a certain pathway that has been annotated.

**Table S19. Exaptations from our set that have overlaps with verified developmental enhancers**

chrom	start	end	name	length	lengthOfConfirmed
chr1	50807508	50807558	exap212	50bp	1743bp
chr11	32154993	32155091	exap1484	98bp	1244bp
chr16	70962838	70962939	exap3631	101bp	3058bp

We compared our set of putative cis-regulatory elements exapted from mobile elements to verified developmental enhancers. The above three exaptations out of our set of 10402 have overlaps with verified enhancers available through the Enhancer Browser [4]. In the last column we show the length of the overlapping region that has been verified as a developmental enhancer. Due to the fact that these validated regions are so much larger, it is not clear if our exaptation is the entire, part of, or outside the functional unit.

**Table S20. Overlaps with human and mouse CAGE tags**

#overlaps	enrichment	tissue	species
1	3.16x	cerebral cortex	mouse
1	1.91x	diencephalon	mouse
1	1.91x	epididymis	human
1	1.08x	renal artery	human
2	6.10x	mammary gland	human
2	1.99x	prostate gland	human
2	2.16x	prostate gland	mouse
2	1.45x	testis	mouse
2	2.26x	ureter	human
2	0.40x	urinary bladder	human
3	6.37x	heart	mouse
4	1.78x	undefined tissue	mouse
4	0.79x	adipose	mouse
4	2.54x	brain	mouse
5	0.86x	pancreas	human
5	1.19x	rectum	human
5	1.71x	somatosensory cortex	mouse
6	0.99x	kidney	human
6	0.83x	small intestine	human
6	1.93x	visual cortex	mouse
7	2.05x	cerebellum	mouse
13	0.63x	cecum	human
13	1.57x	embryo	mouse
20	0.85x	adipose	human
22	2.16x	macrophage	mouse
24	0.99x	undefined tissue	human
25	0.76x	large intestine	human
25	1.58x	liver	mouse
33	2.25x	lung	mouse
34	1.25x	liver	human
53	1.12x	cerebrum	human
120	2.12x	all tissues pooled	mouse
183	1.06x	all tissues pooled	human
297	1.34x	all tissues pooled	both

CAGE sequencing can identify uncommon transcription start sites that may be tissue or time specific. We investigated the intersection of our set with CAGE data that recently became available, but no significant enrichments with a decent sample size were found. Many CAGE tags that did not map uniquely to the genome were discarded, so possibly the full contribution of interspersed repeats acting as alternative start sites is not being realized. While some of the exaptations in our set may act as alternative start sites for neighboring genes, we believe the vast majority function as cis-regulatory elements.

**Table S21. GO Enrichment for elements contributing to peaks of conservation**

subfamily	peak	p-value	GO term
L2	peak4	1.70521e-06	development
L3	peak3	1.87747e-06	nervous system development
L3	peak3	2.01658e-06	system development
L2	peak4	6.96576e-06	nervous system development
L2	peak4	7.36326e-06	system development
L2	peak2	1.74886e-05	positive regulation of receptor mediated endocytosis
L3	peak3	2.28994e-05	homophilic cell adhesion
L2	peak4	2.50304e-05	transcription regulator activity
L2	peak4	2.74366e-05	sequence-specific DNA binding
L2	peak4	2.77285e-05	homophilic cell adhesion
L2	peak2	3.08003e-05	DNA replication factor A complex
L2	peak1	3.57795e-05	regulation of MAPK activity
L2	peak1	3.90588e-05	steroid hormone receptor activity
L2	peak3	4.36072e-05	axon guidance receptor activity
L2	peak1	4.67597e-05	skeletal muscle development
L2	peak1	4.67597e-05	muscle fiber development
L2	peak1	4.69023e-05	ligand-dependent nuclear receptor activity
L3	peak3	5.48026e-05	ligand-regulated transcription factor activity
L3	peak3	5.98758e-05	transmembrane receptor protein tyrosine kinase activity
L2	peak4	6.95206e-05	transcription factor activity
L2	peak4	7.18001e-05	ureteric bud development
L3	peak3	7.66953e-05	potassium ion binding
L3	peak2	8.81003e-05	neuron recognition
L2	peak2	9.20735e-05	replication fork (sensu Eukaryota)
L2	peak2	9.20735e-05	replisome (sensu Eukaryota)
L2	peak1	0.000103508	vesicular fraction
L2	peak3	0.000106477	receptor activity
L2	peak4	0.000114908	specific RNA polymerase II transcription factor activity
L2	peak1	0.000116145	ephrin receptor activity
L3	peak3	0.000117494	regulation of transcription, DNA-dependent

Most significant GO enrichments for elements contributing to the peaks of preferential exaptation in the L2 and L3 consensus sequences. These peaks of preferential exaptation are defined in Figure 2 of the main text. The L2 element (Figure 2A) has four peaks and the L3 element (Figure 2B) has three peaks. Peaks are numbered from left to right.



**Table S22. Pathway enrichment for elements contributing to peaks of conservation**

subfamily	peak	p-value	Pathway
L3	peak3	7.37409e-05	Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy pathway
L2	peak4	0.000102791	Nuclear Receptors
L2	peak1	0.000180276	Androgen and estrogen metabolism
L2	peak1	0.00029759	1 4-Dichlorobenzene degradation
L2	peak1	0.000436606	G13 Signaling Pathway
L3	peak2	0.000440204	Corticosteroids and cardioprotection pathway
L2	peak1	0.000495958	Tetrachloroethene degradation
L2	peak1	0.000694307	Alkaloid biosynthesis I
L3	peak2	0.000793611	GATA3 participate in activating the Th2 cytokine genes expression pathway
L2	peak1	0.000892636	Methane metabolism
L2	peak3	0.000951098	Multi-step Regulation of Transcription by Pitx2 pathway
L2	peak1	0.000991793	Nucleotide GPCRs
L2	peak1	0.000991793	Flavonoids stilbene and lignin biosynthesis
L2	peak2	0.00187986	Chondroitin / Heparan sulfate biosynthesis
L2	peak1	0.00218265	G Protein Signaling Pathways
L3	peak3	0.00218265	IGF-1 Signaling Pathway pathway
L3	peak2	0.00231703	Glycogen Metabolism
L3	peak2	0.0025283	ALK in cardiac myocytes pathway
L2	peak1	0.00267671	Bile acid biosynthesis
L2	peak3	0.00310825	Inactivation of Gsk3 by AKT causes accumulation of b-catenin in Alveolar Macrophages pathway
L2	peak1	0.00396216	Tumor Suppressor Arf Inhibits Ribosomal Biogenesis pathway
L3	peak1	0.00396216	Ion Channel and Phorbol Esters Signaling Pathway pathway
L3	peak1	0.00396216	Phospholipase C d1 in phospholipid associated cell signaling pathway
L2	peak1	0.00415965	Glucocorticoid and Mineralcorticoid Metabolism
L2	peak3	0.0044561	Eph Kinases and ephrins support platelet aggregation pathway
L3	peak1	0.00475282	g-Secretase mediated ErbB4 Signaling Pathway pathway
L2	peak1	0.0053456	Sprouty regulation of tyrosine kinase signals pathway
L2	peak2	0.00534758	Glycogen Metabolism
L2	peak4	0.00589178	GPCRs Class B Secretin-like
L2	peak4	0.00594058	HIV-1 defeats host-mediated resistance by CEM15 pathway

Most significant GO enrichments for elements contributing to the peaks of preferential exaptation in the L2 and L3 consensus sequences. These peaks of preferential exaptation are defined in Figure 2 of the main text. The L2 element (Figure 2A) has four peaks and the L3 element (Figure 2B) has three peaks. Peaks are numbered from left to right.

**Table S23. Groups of exaptations with high sequence similarity**

clique	chrom	start	end	name
clique1	chr10	8684835	8685220	exap936
	chr18	53007470	53007671	exap4255
	chr2	80116457	80116672	exap4750
	chr2	114811708	114811968	exap4807
	chr3	136994724	136994988	exap6335
	chr5	165696963	165697116	exap7708
	chr8	106583914	106584260	exap9131
	chr9	81545776	81546019	exap9526
clique2	chr1	6766244	6766378	exap5
	chr1	36778713	36778982	exap109
	chr14	99928779	99929072	exap3047
	chr15	55737165	55737374	exap3163
	chr17	67835017	67835349	exap3966
	chr5	92669061	92669406	exap7440
	chr5	177644035	177644205	exap7787
clique3	chr10	8684835	8685220	exap936
	chr8	106583914	106584260	exap9131
	chr9	81545776	81546019	exap9526
clique4	chr1	107144025	107144173	exap510
	chr10	115273519	115273717	exap1293
	chr11	19664983	19665165	exap1437
	chr11	79530571	79530719	exap1649
	chr11	114504075	114504252	exap1755
	chr11	115568338	115568460	exap1774
	chr18	33405926	33406103	exap4097
	chr20	39066823	39067075	exap5501
	chr3	138892127	138892327	exap6344
	chr6	116374148	116374323	exap8134
clique5	chr10	115273519	115273717	exap1293
	chr14	69465869	69466011	exap2910
	chr22	33715412	33715545	exap5694
	chr20	39066823	39067075	exap5501
	chr18	33405926	33406103	exap4097
	chr8	119321831	119321989	exap9195
	chr11	79530571	79530719	exap1649
	chr11	19664983	19665165	exap1437
	chr12	76755392	76755554	exap2228
	chr7	147324503	147324609	exap8804
	chr9	8162033	8162218	exap9339
	chr10	86432276	86432453	exap1149
	chr20	49610156	49610280	exap5565
	chr18	35310049	35310264	exap4135
	chr20	14549693	14549878	exap5400
	chr11	114504075	114504252	exap1755
	chr18	34138838	34139057	exap4121
	chr6	116374148	116374323	exap8134
	chr11	115568338	115568460	exap1774
	chr11	96296033	96296233	exap1709
	chr2	233419675	233419834	exap5339
	chr1	80586846	80587042	exap369
	chr3	138892127	138892327	exap6344
	chr11	128966346	128966568	exap1890
	chr2	211371345	211371436	exap5218
	chr1	87886389	87886581	exap428
	chr16	24179203	24179333	exap3458

These five cliques consist of groups of sequences that may or may not have come from the same consensus, but are all similar to each other at the sequence level. We expect that these exaptations, since they all have sequence similarity to each other, may all have similar functions. Some cliques are contained in others. The larger sets were made when we relaxed the threshold for sequence similarity. We investigated the enrichment of these cliques for being near genes with similar GO annotation or pathway annotation (Table S24, Table S25, Table S26, and Table S27).

**Table S24. Top GO enrichments for exaptations with similar sequence composition, assuming a uniform null over bases**

clique	p-value	GO term
clique5	0.000121181	RNA interference
clique5	0.000128404	RNA-mediated gene silencing
clique5	0.000128404	RNA-mediated posttranscriptional gene silencing
clique5	0.000128404	posttranscriptional gene silencing
clique5	0.000132331	translation repressor activity
clique5	0.000234344	germ cell development
clique2	0.000277744	specific RNA polymerase II transcription factor activity
clique5	0.000384721	gene silencing
clique2	0.000735728	tryptophan-tRNA ligase activity
clique2	0.000735728	tryptophanyl-tRNA aminoacylation
clique5	0.00141147	embryonic development
clique4	0.00143248	DNA topoisomerase type I activity
clique4	0.00202342	DNA topoisomerase (ATP-hydrolyzing) activity
clique3	0.00213054	regulation of transcription, DNA-dependent
clique3	0.00223043	transcription, DNA-dependent
clique4	0.00243573	DNA topoisomerase activity
clique3	0.00233803	regulation of transcription
clique3	0.00244173	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
clique4	0.00245744	DNA topological change
clique3	0.00272773	transcription

Table S23 shows the exaptations that make up each clique. We see no strong enrichments using the GO annotation of nearby genes and the enrichment test that assumes a uniform null over bases.

**Table S25. Top GO enrichments for exaptations with similar sequence composition, assuming a uniform null over genes**

clique	p-value	GO term
clique2	0.000108018	specific RNA polymerase II transcription factor activity
clique4	0.000430383	susceptibility to natural killer cell mediated cytotoxicity
clique2	0.000580352	tryptophan-tRNA ligase activity
clique2	0.000580352	tryptophanyl-tRNA aminoacylation
clique4	0.00124042	BRE binding
clique4	0.00124042	positive regulation of cell killing
clique4	0.00124042	positive regulation of immune cell mediated cytotoxicity
clique4	0.00124042	positive regulation of natural killer cell mediated cytotoxicity
clique4	0.00124042	translation repressor activity, nucleic acid binding
clique5	0.00141058	carbamoyl-phosphate synthase (ammonia) activity
clique5	0.00141058	susceptibility to natural killer cell mediated cytotoxicity
clique4	0.00185005	RNA interference
clique4	0.00185005	regulation of natural killer cell mediated cytotoxicity
clique3	0.0020314	regulation of transcription, DNA-dependent
clique3	0.00222354	transcription, DNA-dependent
clique2	0.00244544	ligand-regulated transcription factor activity
clique3	0.00243173	regulation of transcription
clique4	0.00231542	DNA topoisomerase type I activity
clique4	0.00231542	RNA-mediated gene silencing
clique4	0.00231542	RNA-mediated posttranscriptional gene silencing

Table S23 shows the exaptations that make up each clique. We see no strong enrichments using the GO annotation of nearby genes and the enrichment test that assumes a uniform null over genes.

**Table S26. Top pathway enrichments for exaptations with similar sequence composition, assuming a uniform null over bases**

clique	p-value	pathway
clique3	0.00341244	GATA3 participate in activating the Th2 cytokine genes expression pathway
clique1	0.00507357	GATA3 participate in activating the Th2 cytokine genes expression pathway
clique1	0.0184214	Glycogen Metabolism
clique2	0.0244434	Nuclear Receptors
clique5	0.0244385	Role of MEF2D in T-cell Apoptosis pathway
clique2	0.0273532	Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa alpha pathway
clique5	0.030254	GATA3 participate in activating the Th2 cytokine genes expression pathway
clique5	0.0332134	Effects of calcineurin in Keratinocyte Differentiation pathway
clique5	0.0341383	BCR Signaling Pathway pathway
clique5	0.0354481	Calcium Channels
clique5	0.0355802	Neuropeptides VIP and PACAP inhibit the apoptosis of activated T cells pathway
clique5	0.0355844	fMLP induced chemokine gene expression in HMC-1 cells pathway
clique5	0.0455133	Fc Epsilon Receptor I Signaling in Mast Cells pathway

Table S23 shows the exaptations that make up each clique. We see no strong enrichments using the pathway annotation of nearby genes and the enrichment test that assumes a uniform null over bases.

**Table S27. Top pathway enrichments for exaptations with similar sequence composition, assuming a uniform null over genes**

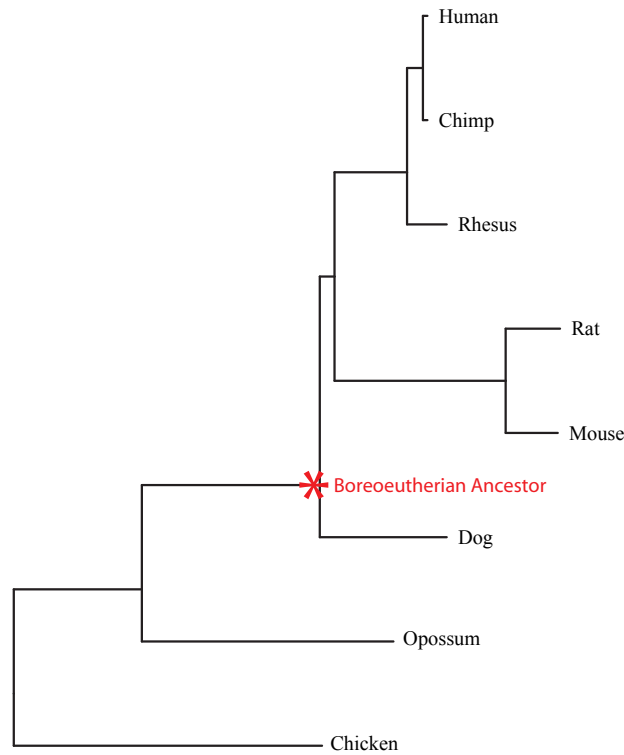
clique	p-value	pathway
clique3	0.00237504	GATA3 participate in activating the Th2 cytokine genes expression pathway
clique1	0.00334285	GATA3 participate in activating the Th2 cytokine genes expression pathway
clique1	0.0124343	Glycogen Metabolism
clique2	0.0131214	Nuclear Receptors
clique5	0.0181044	GATA3 participate in activating the Th2 cytokine genes expression pathway
clique5	0.0203433	Effects of calcineurin in Keratinocyte Differentiation pathway
clique5	0.0203433	Role of MEF2D in T-cell Apoptosis pathway
clique2	0.0205513	Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa alpha pathway
clique5	0.0235241	Calcium Channels
clique5	0.0281455	Neuropeptides VIP and PACAP inhibit the apoptosis of activated T cells pathway
clique5	0.0323825	Control of skeletal myogenesis by HDAC and calcium calmodulin-dependent kinase CaMK pathway
clique5	0.033853	BCR Signaling Pathway pathway
clique5	0.038057	Signaling Pathway from G-Protein Families pathway
clique5	0.038057	fMLP induced chemokine gene expression in HMC-1 cells pathway
clique5	0.0351543	Fc Epsilon Receptor I Signaling in Mast Cells pathway
clique5	0.0402543	Chondroitin / Heparan sulfate biosynthesis
clique5	0.0437443	T Cell Receptor Signaling Pathway pathway

Table S23 shows the exaptations that make up each clique. We see no strong enrichments using the pathway annotation of nearby genes and the enrichment test that assumes a uniform null over genes.

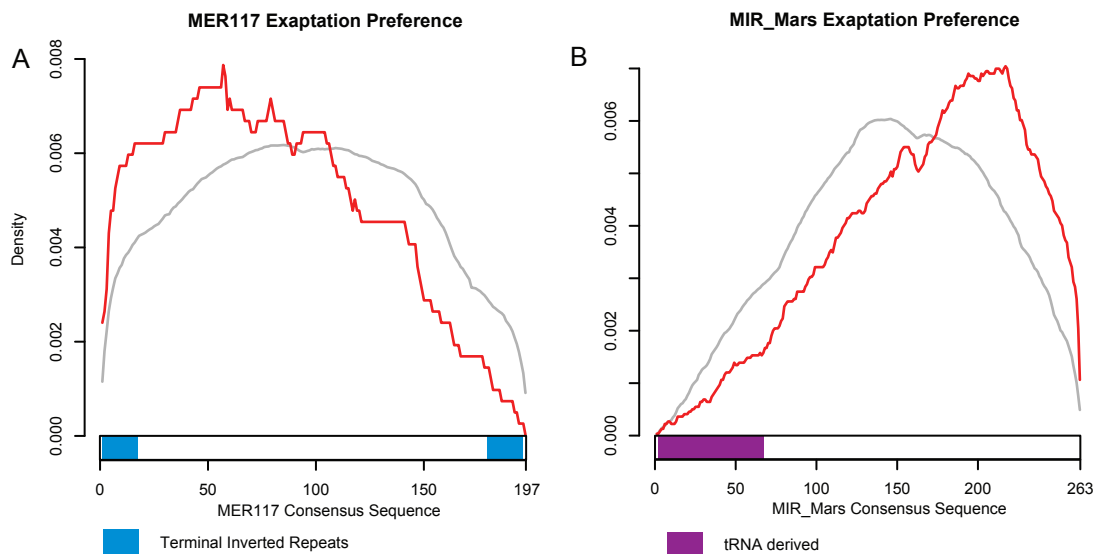
**Table S28. Entropy of exaptation clouds**

nearby gene	cloud entropy	entropy p-value	cloud relative entropy	relative entropy p-value
ODZ4	4.22394	0.071512	0.800085	0.31728
HNT	3.78582	0.757998	0.759496	0.434821
BRUNOL4	3.61931	0.886686	0.680547	0.740578
EPHB1	4.10043	0.167666	1.74524	2e-06
Unknown	3.68071	0.680444	1.18375	0.080675
FCMD	3.70311	0.523548	0.956758	0.476443
CDK5RAP2	3.63825	0.719776	0.790847	0.692403
DIAPH3	3.30954	0.974321	1.21726	0.072967
CRTAC1	3.88485	0.381596	1.03452	0.167772

The clusters, or clouds, of exaptation as defined in Table S16 and shown in Figure 4 were analyzed for their entropy and relative entropy to see if genes were exapting multiple copies of the same element, or trying to acquire one of each. The p-values for getting an entropy that high with the given number of elements in the cloud are based on a simulation. We could not see a general trend across the largest clouds (clusters) of exaptations.



**Fig. S1. Location of the boreoeutherian ancestor** We have highlighted the location of the boreoeutherian ancestor on a species tree with a red asterisk. We insist that all our conserved elements be present in human, chimp, rhesus, rat, mouse and dog so the elements pre-date the boreoeutherian ancestor. By similar logic, we only include repeat subfamilies that have copies in the same six species, and therefore were active before, or during, the boreoeutherian ancestor. We term these subfamilies to be “pan-boreoeutherian.”

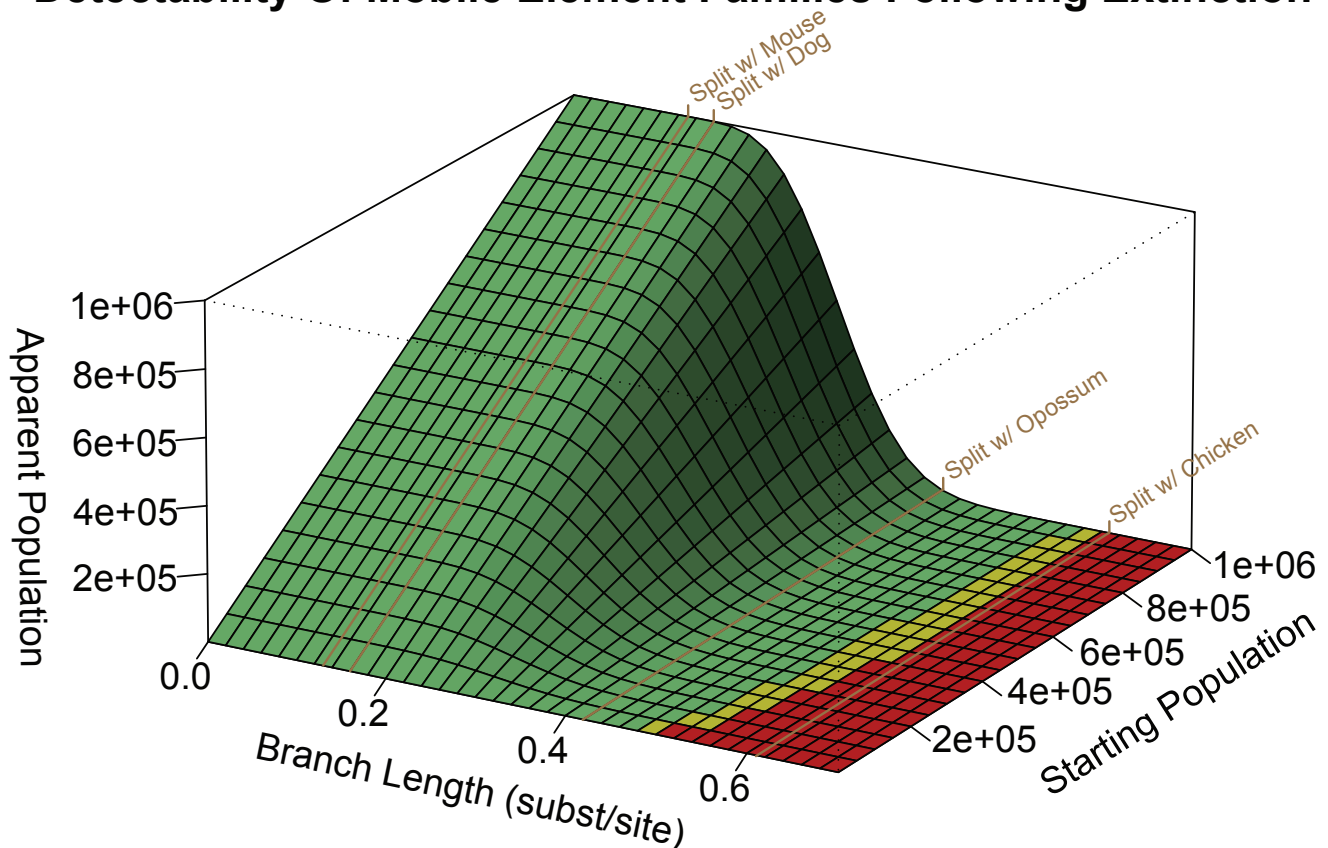


**Fig. S2. Preferential exaptation of specific portions of mobile elements.** For each base in the mobile element consensus (x-axis) the relative abundance is plotted (y-axis). The abundance throughout the entire genome is shown in gray and the abundance that has come under strong purifying selection for a nonexonic function is in red. (A) MER117 is a putative non-autonomous DNA transposon [1] that shows preferential 5' propensity to CNE exaptation. (B) MIR\_Mars is a SINE [1] that shows preferential 3' CNE exaptation, while its tRNA derived 5'-end, responsible for Pol III transcription initiation in SINEs [12], is depleted.



**Fig. S3. Putative engrailed (En-1) binding sites near all known cellular response genes in the reelin-signaling pathway.** All four genes known to play a cellular role in the reelin receptor pathway have an insertion of a MIRb element near them, which has subsequently come under strong purifying selection. These four MIRb elements contain multiple predicted binding sites for En-1, Oct-1 (Pou2f1), SRY, v-Myb, and YY1, each orthologously conserved back to dog. We have aligned together orthologous copies of each of the four paralogs, from six different species, anchored to the consensus sequence of the progenitor MIRb (top). Each genomic sequence is labeled with the gene it is thought to regulate as well as the species it is from. We show a section of the alignment that is rich in potential binding sites for En-1. Each En-1 binding site is shown in dark blue and orthologously conserved instances have a light blue rectangle connecting them. Each paralog appears to conserve multiple binding sites from the original sequence, but not necessarily the same ones. The cases of Oct-1, SRY, v-Myb, and YY1 are qualitatively similar.

## Detectability Of Mobile Element Families Following Extinction



**Fig. S4. Evolutionary distance needed for an extinct repeat family to vanish.** We investigated how long a family can be detected after ceasing to replicate itself. The apparent population size of the repeat in the extant human genome will depend upon how large the family was when it stopped replicating and how long it has been since replication stopped. If the death of the element happened at the speciation of human and dog, then almost all elements will clearly align to each other in the extant genome. If the element stopped jumping at the speciation of human and opossum, then the element will be less obvious, but the numbers should be large enough to notice that a repeat once existed and it may be reconstructed. If the death was before the speciation of human and chicken then the element can most likely not be detected in the present human genome. The surface is colored green when over 100 elements will significantly align to a member of the family, yellow when over 35 elements will align, and red when less than 35 elements will align.