Supplementary Materials and Methods.

Alignment and categorization of WGS sequence traces.

We aligned ~26 million single WGS sequence traces (average length = 799 nt, minimal Phred quality score > 20, cumulative length = ~18 billion nt), deposited recently at the National Center for Biotechnology Information (NCBI) trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?) (Wade and Daly 2005), against the reference C57 genome assembly (build36/mm8, release Feb. 2006) using GMAP (Fig. 1, Supplementary Fig. 1). This program was developed to map cDNA sequences to genomic locations and is well-suited to break contiguous sequences into exons and introns (Wu and Watanabe 2005). However, its use in mapping genomic sequence traces and finding insertions and deletions (indels) has not been reported previously (Stephens et al. 2008). Our use of GMAP sped alignments by a factor of 10 to 100-fold over other computational methods such as Blat (R. M. Stephens, N. Volfovsky, unpublished observations). Some of the WGS traces initially were assembled into an amalgam of mixed strains' genomic sequences (previously available by subscription from Celera), but this "fusion assembly" was not used here.

An "alignment score", assigned to each single sequence trace based on its quality of alignment, was defined as:

alignment score = length of coverage x % identity.

Each trace was aligned by GMAP against all portions of the reference sequence, starting in chromosome 1, and the highest alignment scores were tallied. The GMAP algorithms and further details are available upon request.

A custom Perl script was developed to filter GMAP output files. Based in part on the maximum alignment score(s), each trace first was categorized according to whether or not GMAP aligned it to a unique C57 locus (Supplementary Fig. 1, Supplementary Table 2). Sequence traces that failed to align to unique genomic locations were not mapped further. We subsequently found that such "nopath" traces either aligned with approximately equal Blat scores (Hinrichs et al. 2006) at numerous genomic locations, or in rare cases did not align to any locations. Such nopath traces were examined using RepeatMasker (Smit et al. 2007), showing they are overwhelmingly repetitive element sequences of both simple and other known repeat classes. GMAP identified another category of traces that aligned well at several discrete genomic loci. Approximately one-third of these traces showed a single best alignment (with one alignment score at least 5% better than those for other loci). Using this best alignment, such traces were mapped and categorized further (Supplementary Fig. 1). The remaining two-thirds of this category showed multiple nearly identical alignments that could not be distinguished further ("nobestpath"). These latter traces subsequently were found by RepeatMasker to comprise lower copy number genomic repeats such as segmental duplications. Traces that could not be uniquely mapped to the genome were set aside.

Remaining traces were categorized into additional groups depending on the number of well-aligned patches of contiguous sequence (Supplementary Fig. 1). For those traces that mapped as a single high scoring pair (HSP), i.e. over a single contiguous patch of alignment, two distinct categories were defined. Approximately 73% of all 26 million traces aligned well over essentially their entire length (*i.e*., "traces with minimal variation"). Others (approximately 8%) aligned well over only part of their length ("polymorphism in strain X"). Such traces were filtered into strong and weak alignment groups based on their alignment scores and several other criteria. Traces predicting polymorphisms in strain X were analyzed further as described below.

Almost all variants described here were identified from unassembled strain WGS traces that aligned to the reference genome with 90% identity and at least 200 nt at both (5' and 3') ends. We focused upon 100 nt – 10 kb variants present in the C57 genome but absent from other strain(s) (Fig. 1 and Supplementary Fig. 1). Such variants were distinguished from "weakvariants" based on the presence of two aligned patches (HSPs) within a sequence trace, the distance separating the HSPs (> 100 nt), lengths of each HSP segment (> 200 nt), and alignment scores. Traces containing more than 2 HSPs were identified but not analyzed further here; a majority contains another class of highly repetitive sequences. Some variants were confirmed subsequently by Blat alignment (Hinrichs et al. 2006) against the reference genome.

Categorized trace information was loaded into relational databases using Mysql v. 5.05. Comparing traces' genomic coordinates to those of repetitive elements (Smit et al. 2007), we found another ~ 5% of the traces aligned entirely within repeat elements, i.e. mainly L1 or LTR elements. These traces were excluded from subsequent analysis because they could not be mapped by themselves to a unique genomic locus.

For genomic sequencing, DNA fragments had been narrowly size fractionated and cloned into library plasmids for bidirectional sequencing, facilitating pairing of most sequence traces with AS trace “mates” separated by several kb or more. Therefore, in limited cases we used trace mate-pair information (downloaded from NCBI tracedb archive) to validate single trace alignments by GMAP. Frequently, a trace could be mapped uniquely even when its mate could not. Well-aligned mates accompany approximately 80-85% of all traces, regardless of categorization after GMAP alignments (Supplementary Table 2). However, it must be emphasized that all of the variants described in this study were identified by single sequence trace alignments without mate-pair information. We incorporated mate-pairs as an optional track in PolyBrowse (http://polybrowse.abcc.ncifcrf.gov/) (Stephens et al. 2008).

Merging overlapping traces.

WGS traces may align redundantly to the same genomic locus. To avoid overcounting genomic features or variants, we performed two merging procedures (Fig. 1 and Supplementary Fig. 1). In the first, overlapping traces identifying a common variant in the reference genome were merged into a “unique insertion polymorphism in C57”. Each trace’s two HSPs were represented by coordinates (i_1, i_2, i_3, i_4), where i_1 = chromosomal coordinate at alignment start (mm8, Feb. 2006 release); i_2 = coordinate at variant start; i_3 =

coordinate at variant end; and i_4 = coordinate at alignment end. A merged variant was defined using the group of all aligned traces where any pair of aligned traces i and j satisfies $| i_2 - j_2| < 100$; its final coordinates were defined as:

$$c_1 = \min (i_1);\ c_2 = \max (i_2);\ c_3 = \min (i_3);\ \text{and } c_4 = \max (i_4).$$

Using this procedure, we identified unique variants, thereby significantly reducing the number of features analyzed (Supplementary Table 1).

In a second procedure, overlapping traces with minimal variation were merged with others separated by < 50 nt, forming merged contigs (Fig. 1 and Supplementary Fig. 1). We assigned reference genome coordinates (i_1, i_2) to each well-aligned trace. Two such traces i and j, where $i_1 < j_1$, were merged if $j_1 – i_2 < 50$ or $i_2 - j_1 > 0$. If two neighboring alignments satisfy these criteria, the final coordinates of the merged contig are (i_1, j_2).

Statistical analysis of coincident L1 polymorphisms and SNP-dense regions.

To calculate SNP densities, reference genome sequences were divided into 100 kb blocks. For each block, the number of SNPs from pairwise strain comparisons was determined. A total of 1,233,499 SNPs was downloaded from dbSNP (NCBI) for A/J vs. C57, and 1,111,548 SNPs for DBA/2J vs. C57. High SNP density and low SNP density regions were defined using a threshold of >100 SNPs per 100 kb. High-SNP blocks comprise 20.1% of the genome (comparing A/J vs. C57), and 17.1% (DBA/2J vs. C57), respectively(Wade et al. 2002). Of L1 polymorphisms in C57, 51.4% occurred in high-SNP blocks (A/J vs. C57), significantly higher than their genome-wide frequency (p-value < 1.95E-13) by the binomial test(Wiltshire et al. 2003). Similarly, the rate for DBA/2J was 46.1%, significantly higher than its genome-wide rate (p-value <1E-10).

Structural analysis of transposons.

TSDs (short direct repeats) generally flank retrotransposon insertions and are hallmarks of recent transposition events. Therefore, we examined genomic L1s using a custom Perl program, modified from Szak et al. (Szak et al. 2002), to characterize potential TSDs flanking each element. We limited analysis to elements longer than 100 nt and less than 10 kb. Elements with 5' inversions, or 5' or 3' transductions were also excluded. Fragmented L1s (based upon RepeatMasker annotation) were merged. To identify TSDs between 9 nt and 200 nt long, flanking genomic sequences (200 nt upstream and downstream of reference L1s) were analyzed using bl2seq (with parameters: -g F –W 9 –F F –S 1 –d 3000 –e 1000.0). Right and left genomic flank sequences were scanned for repetitive oligonucleotide "words" starting at 20 nt and decreasing until repeated words were or were not identified at length 7. In the latter case, no TSD sequence was tabulated.

Candidate poly(A) tails were identified in the 200 nt downstream of the 3' UTR of candidate transposon integrants, based upon a minimum length of 6 nt and at least 73% As. Tails containing more than two adjacent non-A bases were disallowed. These criteria were relaxed if patterned repeats were detected in a poly(A) tract (e.g. AAT).

The genomic contexts of candidate transposons were examined. To determine GC nucleotide content flanking transposons, 5 kb genomic sequences upstream and downstream of integrants were analyzed. GC content was calculated as the total number of G or C bases in a sequence, divided by the total number of non-N (called) bases in the same sequence. Distances and orientations of variants were determined relative to RefSeq genes, expressed sequence tags (ESTs), and CpG islands. RefSeq genes (19,557 genes), knownGenes (31,863), MGC genes (17,255), and ESTs (4,415,882) were used to identify the nearest gene within 100 kb of each L1.

To validate predicted empty target site loci, unrepeated TSD sequences were sought in corresponding trace sequences. This search continued until candidate sequences flanking insertion polymorphisms were or were not identified at a minimum length of seven nt. In the latter case, no TSD sequence was tabulated. Output files were generated, along with a simulated alignment of the entire region, for statistical analysis.

Re-classification of L1 subtypes (A, T_F, and G_F).

Occasionally, RepeatMasker breaks one continuous mouse L1 sequence into two separate fragments with the same or different subfamily designation (Chen et al. 2006; Smit et al. 2007). We merged two or more L1 retrotransposon fragments if they satisfied two criteria: (1) the two fragments must be adjacent, virtually contiguous sequences (overlapping, juxtaposed or separated by a gap < 10 nt); and (2) the difference of their nucleotide substitution rates (compared with consensus subfamily sequences) was less than 5%. The L1 subtype of the merged sequence was defined to be the subtype of the larger fragment.

To reclassify candidate young L1 family members (A, T_F and G_F), merged genomic sequences (> 100 nt) were aligned against L1Md_A2 for L1Md_A, L1Md_G_F62 for L1Md_G_F, and L1spa for L1Md_T_F, respectively, using Cross_match (downloaded from http://www.phrap.org). L1s were re-classified to the subfamily whose representative member gave the best alignment score. After reclassification of RepeatMasker output, our counts of L1 subfamily members genome-wide corroborate several prior estimates (Supplementary Table 3).

Chromosomal analysis.

Chromosomal sequences (mm8, Feb. 2006 assembly from UCSC browser) (Hinrichs et al. 2006) was divided into non-overlapping, contiguous 100 kb windows. Densities of transposons (annotated by RepeatMasker) and exons (RefSeq) were calculated for each 100 kb window.

Identification of fusion transcripts containing transposons.

Reference genome coordinates of transposons, RefSeq genes and ESTs (UCSC annotation) were determined. Transcripts with exons mapping inside or within 50 nt outside of transposons were tabulated.

To detect fusion transcripts containing L1T_F fragments (or other young L1 family members), a double-stranded DNA probe for mouse L1spa ORF1 was generated by PCR using primers DES1167 (5' ACTAAAACAGGAACCAAGACCAC 3') and DES1168 (5' GTTCATTTCCATCACCTGTTTGTATG 3'). A mouse L1spa ORF2 probe was generated using DES1165 (5’ CAATACAAGAACGGGAACAAC 3’) and DES1166 (5’ ACCTTTGATGAGAATGAAGTGTC 3’). PCR amplification was performed for 30 cycles at 94°C for 30s, 58°C for 30s, and 72°C for 1 min, and products were gel purified prior to radiolabelling with alpha-32P dCTP (random nonamer labeling kit, GE Amersham).

Commercial phage libraries representing mouse testis large-insert cDNAs (Clontech) or thymus cDNAs (Stratagene) were plated at about 50,000 plaques per 150 x 15 mm Petri dish and transferred to Hybond-N filters (Amersham). Filters were hybridized in 1x Denhardt’s solution (pH 7.5), containing 1 mM EDTA, pH 8.0, 0.2mg/ml sheared and denatured fish sperm DNA (Roche) and 1% SDS, with radiolabeled ORF2 probe, washed at 65°C in 0.1X SSC and 0.1% SDS, and exposed to film. Filters were subsequently rehybridized with ORF1 probe and autoradiographed again. Those plaques that hybridized to the ORF1 probe, but not with ORF2 (so-called ORF1^{+}ORF2^{-} cDNAs) (Li et al. 2008), were enriched by screening until they were purified to homogeneity. Positives were converted to plasmids following the library manufacturers’ directions.

DNA sequencing and analysis.

Plasmids containing cDNA or genomic DNA inserts were purified using miniprep columns. Sequencing was performed using standard fluorescent dideoxy terminator chemistry (Big Dye v. 3.1, Applied Biosystems) on a 96 capillary electrophoresis instrument (Spectrumedix) using oligonucleotides for the testis cDNA library plasmid’s 5’ and 3’ ends, respectively, i.e. DES886 (5’-AAGCGCGCCATTGTGTTGG 3’) and

DES837 (5’ TAATACGACTCACTATAGGG 3’), M13R and M13F oligonucleotides for thymus cDNA and genomic DNA clones, and custom oligonucleotides for additional sequencing by primer walking (sequences available upon request).

Supplementary information.

Resolution of conflicting traces.

A single discrepancy was observed in validating predicted variants by PCR. A “weakvariant” trace predicted an L1 insertion polymorphism in the C57 genome and its absence from A/J. However, another trace with minimal variation, also from A/J, indicated its presence in A/J. PCR unequivocally demonstrated the presence of this L1 integrant in A/J. Explanations for this discrepancy include an error in trace identification; contamination of strains or DNA specimens prior to sequencing or our PCR assays, including unexpected heterozygosity at the locus; an unexpected copy number variant or segmental duplication of the locus; or an independent, concurrent L1 insertion into the identical location (which would be extremely unlikely).

Polymorphic AS L1s in genes.

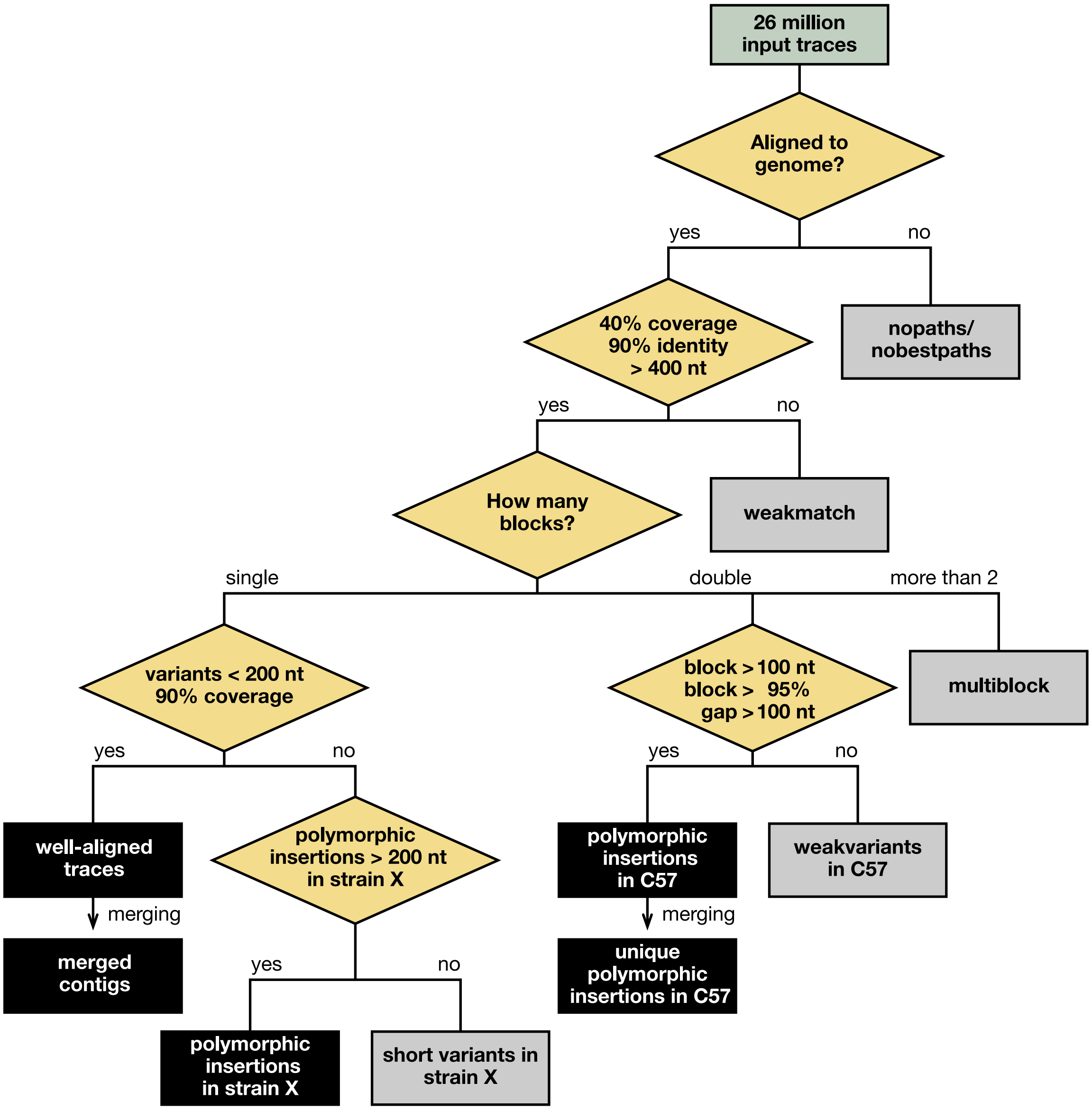
Almost all 336 L1 variants that are full-length (*i.e.* they contain at least one 5’ UTR monomer) and occur within annotated genes in the reference genome are intronic. Approximately 60% are oriented AS to flanking ORFs (Supplementary Table 9) and are members of the T_F, A and G_F subfamilies. These represent 9% of all polymorphic L1 T_F, A and G_F elements (N = 3,974; Supplementary Table 3) and 20% of full-length L1 polymorphisms genome-wide (N = 1,714; Supplementary Table 6a).

Supplementary Figure Legends.

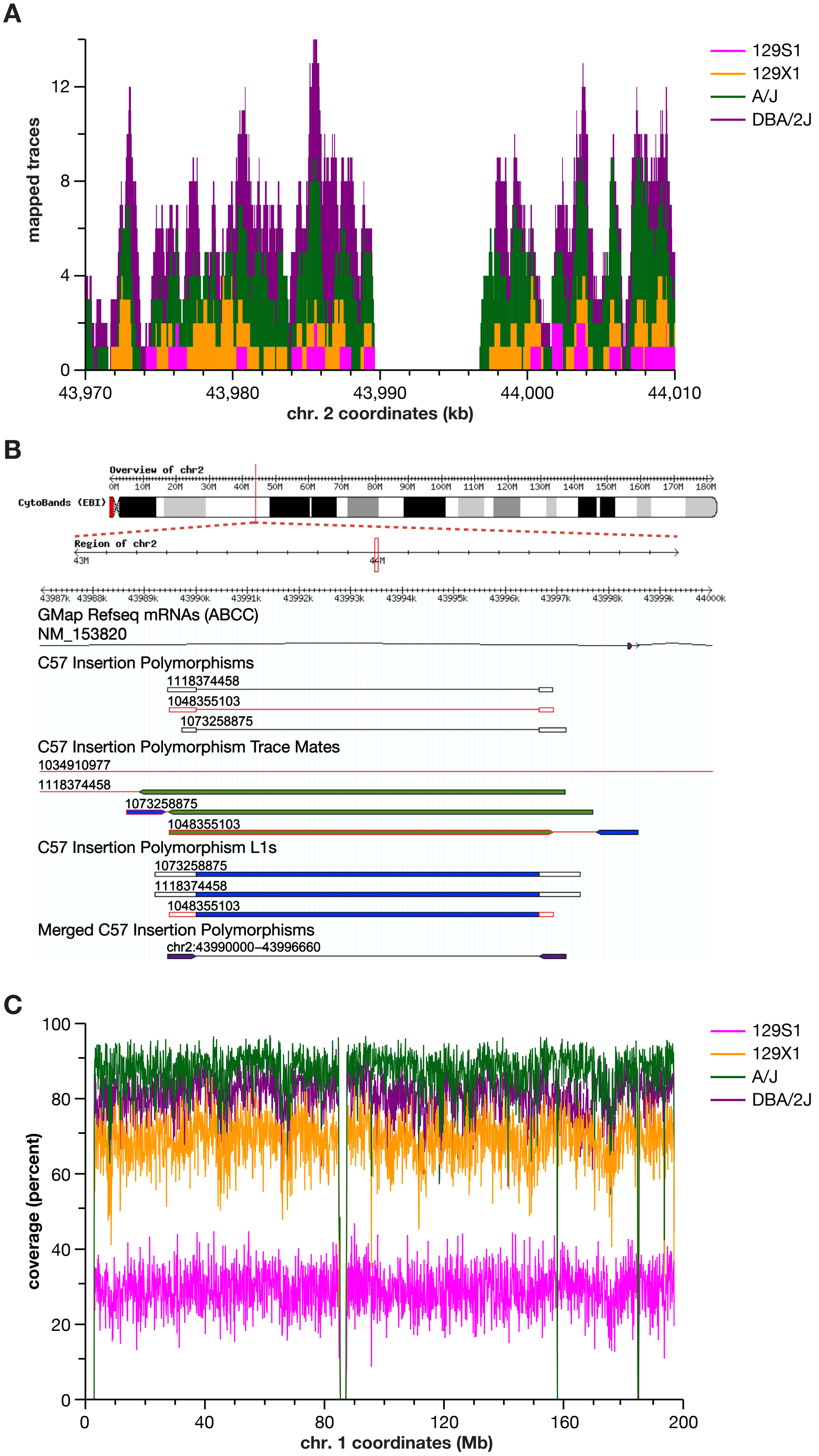
Supplementary Fig. 1. **Categories of WGS sequence trace aligments.** To discover polymorphisms between inbred mouse strains, available WGS sequence traces were aligned to the C57 reference genome using GMAP. Resulting categories of trace alignments included “well-aligned”, “insertion polymorphism in C57” and “insertion polymorphism in strain X”. Overlapping well-aligned traces were merged, forming larger contigs. Similarly overlapping traces identifying insertion polymorphisms were merged to define unique indels.

Supplementary Fig. 2. **Detection and display of mouse strain polymorphisms.** (a) Individual WGS sequence traces from four alternative strains aligned to the C57 reference genome at *ArhGAP15.* The number of aligned traces (vertical axis) at each chromosomal position (horizontal axis) is related to the overall sequence coverage density (Table S1). The coverage abruptly drops to zero at the genomic boundary of a L1 polymorphism, indicating its likely absence in all four unassembled strains and presence in the C57 reference genome. (b) PolyBrowse, a web-based display and query browser, can display annotated genes, other reference genome features and strain variants including single nucleotide polymorphisms (SNPs) and indels. Individual traces from unassembled strains are aligned to C57 reference sequence at the *ArhGAP15* locus on chr. 2 (Stephens et al. 2008). (c) Coverage density of aligned, merged WGS traces (number of nucleotides aligned per 100 kb windows) is displayed as a histogram along chromosome 1 for each of the four analyzed strains: purple, DBA/2J; green, A/J; orange, 129X1; pink, 129S1.

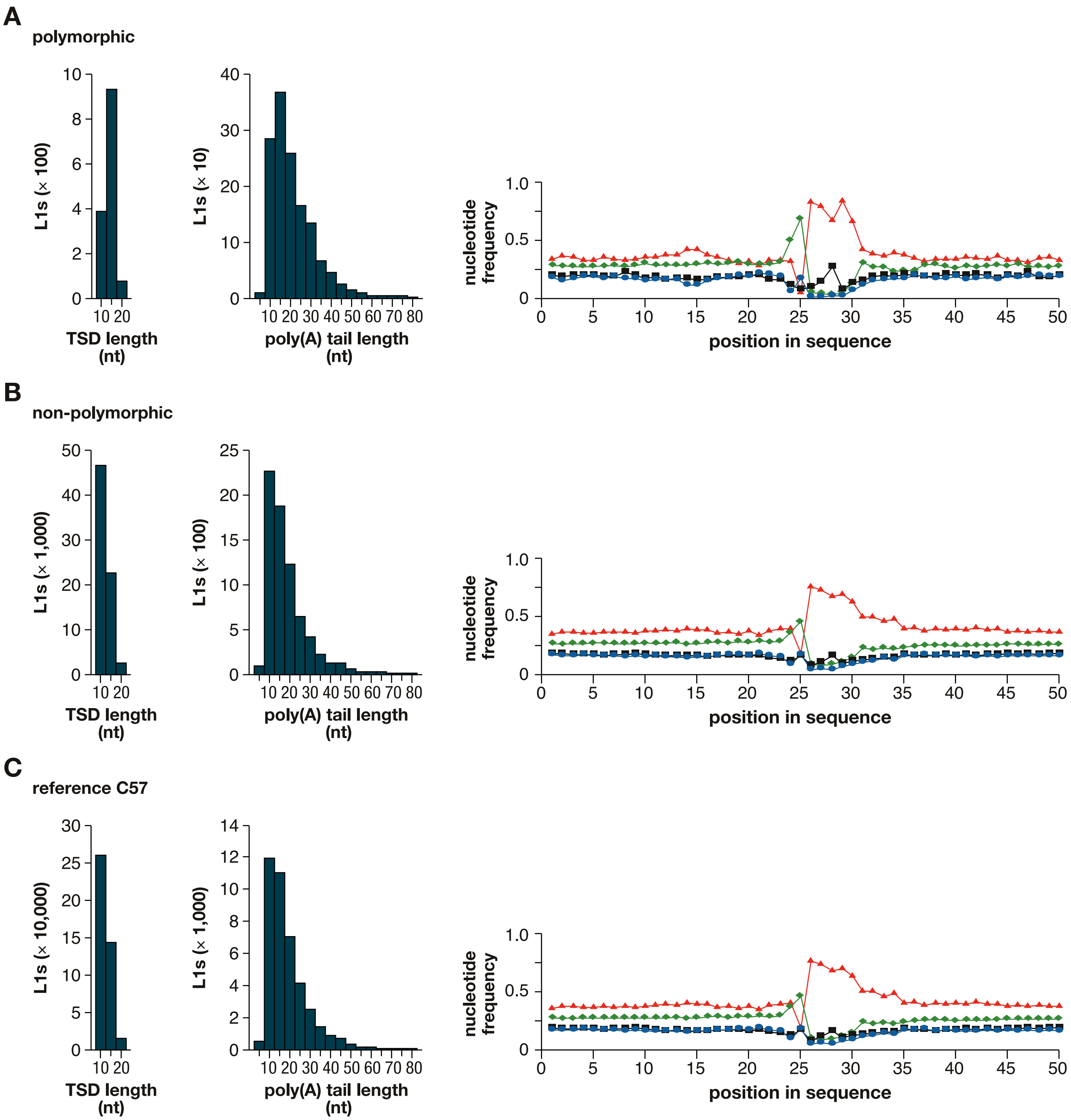
Supplementary Fig. 3. **Polymorphic L1s are recent genomic integrants.** (a) L1 polymorphisms in the reference genome (insertion polymorphisms in C57, absent from at least one of the unassembled strains); (b) non-polymorphic L1s; and (c) reference L1s. Histograms indicate counts of elements with indicated features: (*Left*) count of L1s with target site duplications of indicated lengths (nt); (*middle*) count of L1s with poly(A) tails of indicated lengths (nt); (*right*) target nick site nucleotide frequency (Symer et al. 2002). Recent L1 retrotransposition integrants are expected to include more full-length elements, longer target site duplications (TSDs), longer poly(A) tails, and canonical target nick site at TT^AAAA (Symer et al. 2002). See Fig. 4.


26 million input traces
Aligned to genome?
yes
no
40% coverage 90% identity > 400 nt
nopaths/ nobestpaths
yes
no
How many blocks?
weakmatch
single
double
more than 2
variants < 200 nt 90% coverage
block > 100 nt block > 95% gap > 100 nt
multiblock
yes
no
yes
no
well-aligned traces
polymorphic insertions > 200 nt in strain X
polymorphic insertions in C57
weakvariants in C57
merging
merging
merged contigs
unique polymorphic insertions in C57
yes
no
polymorphic insertions in strain X
short variants in strain X

Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3

Supplementary Table 1. **Reduction in overlapping trace counts by merging.**

Type of	129S1		129X1		A/J		DBA/2J	
alignment	traces	contigs	traces	contigs	traces	contigs	traces	contigs
min.variation	1,116,193	758,235	4,022,174	1,063,807	7,934,277	688,930	5,801,176	897,164
insert in C57	4,628	2,132	15,253	5,928	33,252	10,238	24,694	8,449

Traces with minimal variation and those identifying insertion polymorphisms in C57 are counted before and after merging into contigs for each unassembled strain.

Supplementary Table 2. **Categories of trace alignments.** Trace alignment categories, organized by strain of origin, repeat content, co-aligment with mate traces, and average size. Regardless of alignment category, most traces co-aligned with nearby mates at similar frequencies, suggesting alignment procedures correctly assigned most traces to reference genome loci. (See Fig. 1 and Supp. Fig. 1 for more description about alignment categories.)

Traces with minimal variation by strain

	traces		mates		repeats (some traces have multiple repeats)		average size of single repeat, nt		traces with one or more repeats	
strain	number	percent	number	percent	number	percent	trace size	repeat size	number	percent
129S1_SVIMJ_CRA	1,118,796	5.62%	935,330	83.60%	1,292,561	115.53%	795.73	193.13	766,523	68.51%
129X1_SVJ_CRA	4,031,241	20.25%	3,241,169	80.40%	4,533,969	112.47%	798.11	191.59	2,717,591	67.41%
A_J_CRA	7,953,029	39.95%	6,835,140	85.94%	9,189,658	115.55%	800.59	193.08	5,445,332	68.47%
C57BL6_J_HPGC	322,757	1.62%	0	0.00%	381,387	118.17%	890.47	158.14	210,727	65.29%
C57BL_6J_BCM	668,329	3.36%	283,142	42.37%	856,205	128.11%	919.03	187.30	476,776	71.34%
DBA_2J_CRA	5,812,166	29.20%	4,837,970	83.24%	6,703,650	115.34%	799.71	194.27	3,971,577	68.33%
total	19,906,318	100.00%	16,132,751	81.04%	22,957,430	115.33%	833.94	186.25	13,588,526	68.26%

Traces with minimal variation, by repeat size range

repeat size range, nt	repeats		traces with repeats	
	number	percent	number	percent
0-200	16,315,292	71.07%	5,802,662	42.70%
201-400	3,592,010	15.65%	3,336,480	24.55%
401-600	1,279,707	5.57%	1,822,477	13.41%
601-800	1,293,686	5.64%	2,020,124	14.87%
801-	476,735	2.08%	606,783	4.47%
total	22,957,430	100.00%	13,588,526	100.00%

LINE/L1: 22.74%, Simple repeat: 22.25%, SINE/Alu: 10.25%, LTRs: 17.96%

Traces identifying insertion polymorphisms in C57

Variant traces by strain

	traces		mates		repeats		average size, nt		traces with repeats	
strain	number	percent	number	percent	number	percent	trace size	repeat size	number	percent
129S1_SVIMJ_CRA	2,376	5.95%	2,024	85.19%	3,182	133.92%	793.62	191.35	1,767	74.37%
129X1_SVJ_CRA	7,731	19.36%	6,321	81.76%	10,335	133.68%	795.71	189.19	5,798	75.00%
A_J_CRA	16,742	41.92%	14,502	86.62%	23,138	138.20%	799.35	190.76	12,789	76.39%
C57BL6_J_HPGC	202	0.51%	0	0.00%	281	139.11%	902.35	162.13	149	73.76%
C57BL_6J_BCM	421	1.05%	158	37.53%	640	152.02%	909.54	177.44	335	79.57%
DBA_2J_CRA	12,465	31.21%	10,404	83.47%	17,065	136.90%	797.93	188.80	9,373	75.19%
total	39,937	100.00%	33,409	83.65%	54,641	136.82%	833.08	183.28	30,211	75.65%

Variant traces by repeat size range

repeat size range, nt	repeats		traces with repeats	
	number	percent	number	percent
0-200	37,864	69.30%	10,894	36.06%
201-400	9,896	18.11%	8,322	27.55%
401-600	4,058	7.43%	5,493	18.18%
601-800	2,152	3.94%	4,397	14.55%
801-	671	1.23%	1,105	3.66%
Total	54,641	100.00%	30,211	100.00%

LINE/L1: 23.90%, Simple repeat: 19.68%, SINE/Alu: 11.24%, LTRs: 20.5%

Nopaths: no alignment hit at all

nopaths by strain

	traces		mates		repeats		average size, nt		traces with repeats	
							trace	repeat		
Strain	number	percent	number	percent	number	percent	size	size	number	percent
129S1_SVIMJ_CRA	6,257	2.56%	5,148	82.28%	9,863	157.63%	808.16	200.53	4636	74.09%
129X1_SVJ_CRA	25,687	10.50%	19,982	77.79%	36,032	140.27%	811.33	189.48	18425	71.73%
A_J_CRA	53,920	22.04%	44,633	82.78%	79,686	147.79%	815.83	181.53	38193	70.83%
C57BL6_J_HPGC	91,509	37.40%	0	0.00%	17,547	19.18%	1,017.91	117.77	11650	12.73%
C57BL_6J_BCM	32,107	13.12%	424	1.32%	4,352	13.55%	850.34	123.16	2875	8.95%
DBA_2J_CRA	35,204	14.39%	28,348	80.52%	53,686	152.50%	812.05	214.23	26394	74.97%
total	244,684	100.00%	98,535	40.27%	201,166	82.21%	852.60	171.11	102,173	41.76%

nopaths by repeat size range

repeat size range, nt	repeats		traces with repeats	
	number	percent	number	percent
0-200	153,977	76.54%	39,968	39.12%
201-400	18,811	9.35%	20,219	19.79%
401-600	8,041	4.00%	14,717	14.40%
601-800	15,443	7.68%	21,527	21.07%
801-	4,894	2.43%	5,742	5.62%
Total	201,166	100.00%	102,173	100.00%

Simple repeat: 33.93%, low complexity: 10.49%, LINE/L1: 11.65%, SINE/Alu: 10.49%, LTRs: 8.72%. Highlighted in yellow: poorly aligned, nopath traces also lacking mate pair alignments that were not used further in analysis.

Nobestpaths: too many hits, with no single best alignment

nobestpaths by strain

	traces		mates		repeats		average size, nt		traces with repeats	
							trace	Repeat		
Strain	number	percent	number	percent	number	percent	size	size	number	percent
129S1_SVIMJ_CRA	118,684	5.97%	100,335	84.54%	136,323	114.86%	797.49	477.85	103,218	86.97%
129X1_SVJ_CRA	382,732	19.25%	306,824	80.17%	441,810	115.44%	800.83	449.82	327,406	85.54%
A_J_CRA	866,789	43.60%	743,279	85.75%	1,012,537	116.81%	803.70	467.93	757,399	87.38%
C57BL6_J_HPGC	12,494	0.63%	0	0.00%	14,834	118.73%	896.37	489.29	10,909	87.31%
C57BL_6J_BCM	22,442	1.13%	8,794	39.19%	25,260	112.56%	900.97	549.88	19,410	86.49%
DBA_2J_CRA	585,072	29.43%	488,310	83.46%	669,858	114.49%	802.03	506.55	518,372	88.60%
total	1,988,213	100.00%	1,647,542	82.87%	2,300,622	115.71%	833.57	490.22	1,736,714	87.35%

nobestpaths by repeat size range

repeat size range, nt	repeats		traces with repeats	
	number	percent	number	percent
0-200	722,719	31.41%	191,093	11.00%
201-400	291,817	12.68%	153,034	8.81%
401-600	184,087	8.00%	170,499	9.82%
601-800	735,919	31.99%	827,730	47.66%
801-	366,080	15.91%	394,358	22.71%
Total	2,300,622	100.00%	1,736,714	100.00%

LINE/L1: 28.06%, Satellite: 21.95%, LTRs: 28.26%, Simple repeat: 9.31%

Insertion polymorphisms in unassembled strain X

Variant traces by strain

	traces		mates		repeats		average size, nt		traces with repeats	
							trace	repeat		
strain	number	percent	number	percent	number	percent	size	size	number	percent
129S1_SVIMJ_CRA	65,828	2.97%	53,049	80.59%	92,908	141.14%	806.07	159.21	50,572	76.82%
129X1_SVJ_CRA	613,820	27.71%	482,631	78.63%	731,257	119.13%	808.16	161.66	430,549	70.14%
A_J_CRA	1,059,161	47.82%	876,137	82.72%	1,308,666	123.56%	817.79	163.69	752,184	71.02%
C57BL6_J_HPGC	55,266	2.50%	0	0.00%	71,513	129.40%	915.85	137.31	37,114	67.16%
C57BL_6J_BCM	26,209	1.18%	9,871	37.66%	35,880	136.90%	947.33	172.56	19,451	74.21%
DBA_2J_CRA	394,577	17.81%	317,184	80.39%	544,222	137.93%	810.69	161.73	300,563	76.17%
Total	2,214,861	100.00%	1,738,872	78.51%	2,784,446	125.72%	850.98	159.36	1,590,433	71.81%

Variant traces by repeat size range

repeat size range, nt	repeats		traces with repeats	
	number	percent	number	percent
0-200	2,097,387	75.33%	719,456	45.24%
201-400	405,829	14.57%	432,293	27.18%
401-600	169,000	6.07%	255,351	16.06%
601-800	103,810	3.73%	170,353	10.71%
801-	8,420	0.30%	12,980	0.82%
total	2,784,446	100.00%	1,590,433	100.00%

Simple repeat: 26.87%, LINE/L1: 21.15%, low complexity: 17.28%, SINE/Alu: 8.29%, LTRs: 13.98%

L1 polymorphisms (estimated number): 16,658

Multiexons: two or more insertion/deletion variants identified by one sequence trace

multiexons by strain

	traces		mates		repeats		average size, nt		traces with repeats	
							trace	repeat		
strain	number	Percent	number	percent	number	percent	size	size	number	percent
129S1_SVIMJ_CRA	16,381	6.29%	13,781	84.13%	20,918	127.70%	808.06	541.07	15,896	97.04%
129X1_SVJ_CRA	43,549	16.72%	34,892	80.12%	61,131	140.37%	812.62	446.54	41,666	95.68%
A_J_CRA	108,470	41.65%	92,611	85.38%	135,784	125.18%	814.55	480.47	94,999	87.58%
C57BL6_J_HPGC	1,496	0.57%	0	0.00%	2,393	159.96%	1,019.26	152.19	1,144	76.47%
C57BL_6J_BCM	1,029	0.40%	359	34.89%	1,857	180.47%	919.12	220.87	901	87.56%
DBA_2J_CRA	89,520	34.37%	74,187	82.87%	114,667	128.09%	814.08	538.57	86,780	96.94%
Total	260,445	100.00%	215,830	82.87%	336,750	129.30%	864.61	396.62	241,386	92.68%

multiexons by repeat size range

repeat size range, nt	repeats		traces with repeats	
	number	percent	number	percent
0-200	120,404	35.75%	15,569	6.45%
201-400	22,139	6.57%	16,392	6.79%
401-600	12,039	3.58%	15,646	6.48%
601-800	106,049	31.49%	115,697	47.93%
801-	76,119	22.60%	78,082	32.35%
Total	336,750	100.00%	241,386	100.00%

Satellite: 53.46%, simple repeat: 15.55%, LINE/L1: 9.06%

Supplementary Table 3. **Mouse L1 retrotransposon subfamilies in reference C57 genome.**

Counts of L1 families identified by RepeatMasker. Reference L1 elements were categorized as polymorphic (absent in at least one unassembled strain(s)); non-polymorphic (present in all five strains); or undetermined. Members of various L1 subfamilies were determined by RepeatMasker, except members of T_F, A and G_F (*yellow*) subfamilies, which were reclassified using canonical sequences and CrossMatch.

L1 class	non-polymorphic		unknown		polymorphic		total reference	
	count	%	count	%	count	%	count	%
HAL1	624	0.49%	2,654	0.50%	6	0.09%	3,284	0.49%
HAL1b	110	0.09%	392	0.07%	1	0.01%	503	0.08%
L1_Mm	1,298	1.02%	5,225	0.98%	78	1.16%	6,601	0.99%
L1_Mur1	1,236	0.97%	4,927	0.93%	38	0.57%	6,201	0.93%
L1_Mur2	2,553	2.00%	10,587	1.99%	86	1.28%	13,226	1.98%
L1_Mur3	4,363	3.41%	16,928	3.18%	97	1.44%	21,388	3.21%
L1_Mus1	4,393	3.44%	16,834	3.17%	96	1.43%	21,323	3.20%
L1_Mus2	2,822	2.21%	10,960	2.06%	64	0.95%	13,846	2.08%
L1_Mus3	3,505	2.74%	14,529	2.73%	76	1.13%	18,110	2.72%
L1_Mus4	1,713	1.34%	6,704	1.26%	35	0.52%	8,452	1.27%
L1_Rod	1,718	1.34%	7,115	1.34%	18	0.27%	8,851	1.33%
L1M	301	0.24%	1,345	0.25%	0	0.00%	1,646	0.25%
L1M1	289	0.23%	1,181	0.22%	0	0.00%	1,470	0.22%
L1M2	2,781	2.18%	11,201	2.11%	40	0.59%	14,022	2.10%
L1M2a	40	0.03%	135	0.03%	0	0.00%	175	0.03%
L1M2a1	0	0.00%	2	0.00%	0	0.00%	2	0.00%
L1M2b	0	0.00%	12	0.00%	3	0.04%	15	0.00%
L1M2c	69	0.05%	228	0.04%	0	0.00%	297	0.04%
L1M3	1,252	0.98%	5,427	1.02%	12	0.18%	6,691	1.00%
L1M3a	14	0.01%	47	0.01%	0	0.00%	61	0.01%
L1M3b	17	0.01%	64	0.01%	1	0.01%	82	0.01%
L1M3c	34	0.03%	200	0.04%	0	0.00%	234	0.04%
L1M3d	15	0.01%	99	0.02%	0	0.00%	114	0.02%
L1M3de	29	0.02%	90	0.02%	0	0.00%	119	0.02%
L1M3e	82	0.06%	340	0.06%	5	0.07%	427	0.06%
L1M3f	4	0.00%	40	0.01%	0	0.00%	44	0.01%
L1M4	1,924	1.51%	8,640	1.62%	32	0.48%	10,596	1.59%
L1M4b	441	0.35%	1,932	0.36%	3	0.04%	2,376	0.36%
L1M4c	570	0.45%	2,303	0.43%	9	0.13%	2,882	0.43%
L1M5	3,016	2.36%	12,692	2.39%	38	0.57%	15,746	2.36%
L1MA10	135	0.11%	603	0.11%	3	0.04%	741	0.11%
L1MA4	1,124	0.88%	4,314	0.81%	12	0.18%	5,450	0.82%
L1MA4A	387	0.30%	1,660	0.31%	2	0.03%	2,049	0.31%
L1MA5	764	0.60%	3,047	0.57%	8	0.12%	3,819	0.57%

L1MA5A	205	0.16%	720	0.14%	6	0.09%	931	0.14%
L1MA6	1,100	0.86%	4,512	0.85%	19	0.28%	5,631	0.85%
L1MA7	525	0.41%	2,008	0.38%	6	0.09%	2,539	0.38%
L1MA8	663	0.52%	2,469	0.46%	8	0.12%	3,140	0.47%
L1MA9	723	0.57%	2,850	0.54%	8	0.12%	3,581	0.54%
L1MB1	394	0.31%	1,503	0.28%	3	0.04%	1,900	0.29%
L1MB2	450	0.35%	1,794	0.34%	9	0.13%	2,253	0.34%
L1MB3	912	0.71%	3,251	0.61%	14	0.21%	4,177	0.63%
L1MB4	403	0.32%	1,784	0.34%	8	0.12%	2,195	0.33%
L1MB5	581	0.45%	2,280	0.43%	3	0.04%	2,864	0.43%
L1MB7	1,148	0.90%	4,309	0.81%	12	0.18%	5,469	0.82%
L1MB8	970	0.76%	3,698	0.70%	23	0.34%	4,691	0.70%
L1MC	1,176	0.92%	4,723	0.89%	21	0.31%	5,920	0.89%
L1MC1	1,034	0.81%	3,502	0.66%	10	0.15%	4,546	0.68%
L1MC2	380	0.30%	1,405	0.26%	9	0.13%	1,794	0.27%
L1MC3	931	0.73%	3,539	0.67%	6	0.09%	4,476	0.67%
L1MC4	835	0.65%	3,256	0.61%	9	0.13%	4,100	0.62%
L1MC4_3	144	0.11%	580	0.11%	1	0.01%	725	0.11%
L1MC4a	506	0.40%	2,089	0.39%	4	0.06%	2,599	0.39%
L1MC5	392	0.31%	1,757	0.33%	3	0.04%	2,152	0.32%
L1MCa	608	0.48%	2,446	0.46%	12	0.18%	3,066	0.46%
L1MCb	135	0.11%	561	0.11%	4	0.06%	700	0.11%
L1MCc	95	0.07%	473	0.09%	2	0.03%	570	0.09%
L1MD	1,002	0.78%	4,715	0.89%	13	0.19%	5,730	0.86%
L1Md_A	1,292	1.01%	9,298	1.75%	1,570	23.35%	12,160	1.82%
L1Md_F	901	0.70%	4,308	0.81%	75	1.12%	5,284	0.79%
L1Md_F2	6,899	5.40%	37,535	7.06%	531	7.90%	44,965	6.75%
L1Md_F3	1,376	1.08%	9,207	1.73%	133	1.98%	10,716	1.61%
L1Md_G$_F$	330	0.26%	3,116	0.59%	535	7.96%	3,981	0.60%
L1Md_T$_F$	840	0.66%	7,589	1.43%	1,869	27.80%	10,298	1.55%
L1MD1	400	0.31%	1,495	0.28%	1	0.01%	1,896	0.28%
L1MD2	479	0.37%	2,044	0.38%	2	0.03%	2,525	0.38%
L1MD3	448	0.35%	2,008	0.38%	5	0.07%	2,461	0.37%
L1MDa	606	0.47%	2,827	0.53%	8	0.12%	3,441	0.52%
L1MDb	110	0.09%	519	0.10%	2	0.03%	631	0.09%
L1ME1	870	0.68%	3,267	0.61%	8	0.12%	4,145	0.62%
L1ME2	563	0.44%	2,037	0.38%	5	0.07%	2,605	0.39%
L1ME3	209	0.16%	846	0.16%	1	0.01%	1,056	0.16%
L1ME3A	342	0.27%	1,426	0.27%	5	0.07%	1,773	0.27%
L1ME3B	265	0.21%	1,026	0.19%	5	0.07%	1,296	0.19%
L1ME4a	259	0.20%	1,097	0.21%	0	0.00%	1,356	0.20%
L1MEa	50	0.04%	195	0.04%	0	0.00%	245	0.04%
L1MEb	119	0.09%	459	0.09%	2	0.03%	580	0.09%
L1MEc	1,067	0.83%	4,063	0.76%	5	0.07%	5,135	0.77%
L1MEd	198	0.15%	825	0.16%	0	0.00%	1,023	0.15%
L1MEe	236	0.18%	1,175	0.22%	3	0.04%	1,414	0.21%

L1P5	40	0.03%	211	0.04%	0	0.00%	251	0.04%
L1PB4	27	0.02%	125	0.02%	0	0.00%	152	0.02%
L1VL1	437	0.34%	1,986	0.37%	13	0.19%	2,436	0.37%
L1VL2	623	0.49%	2,362	0.44%	19	0.28%	3,004	0.45%
L1VL4	886	0.69%	2,873	0.54%	14	0.21%	3,773	0.57%
Lx	3,874	3.03%	15,544	2.92%	75	1.12%	19,493	2.93%
Lx2	2,669	2.09%	10,585	1.99%	54	0.80%	13,308	2.00%
Lx2A	177	0.14%	844	0.16%	1	0.01%	1,022	0.15%
Lx2A1	409	0.32%	1,794	0.34%	4	0.06%	2,207	0.33%
Lx2B	2,044	1.60%	8,058	1.52%	44	0.65%	10,146	1.52%
Lx3_Mus	1,605	1.26%	6,890	1.30%	42	0.62%	8,537	1.28%
Lx3A	1,123	0.88%	4,840	0.91%	19	0.28%	5,982	0.90%
Lx3B	837	0.65%	3,683	0.69%	19	0.28%	4,539	0.68%
Lx3C	1,353	1.06%	5,173	0.97%	24	0.36%	6,550	0.98%
Lx4A	1,742	1.36%	6,732	1.27%	28	0.42%	8,502	1.28%
Lx4B	1,684	1.32%	6,943	1.31%	23	0.34%	8,650	1.30%
Lx5	4,880	3.82%	17,778	3.34%	83	1.23%	22,741	3.41%
Lx6	4,676	3.66%	18,661	3.51%	80	1.19%	23,417	3.51%
Lx7	6,392	5.00%	25,359	4.77%	122	1.81%	31,873	4.78%
Lx8	11,614	9.09%	42,969	8.08%	161	2.39%	54,744	8.22%
Lx9	9,019	7.06%	35,864	6.74%	135	2.01%	45,018	6.76%
MusHAL1	873	0.68%	3,348	0.63%	22	0.33%	4,243	0.64%
N/D	25	0.02%	137	0.03%	0	0.00%	162	0.02%
total	127,803	100.00%	531,802	100.00%	6,723	100.00%	666,328	100.00%

(b) **Reclassification of young L1s in the reference C57 genome.**

L1 class	RepeatMasker default output	After Cross_match re-classification
A	16,656	12,160
T_F	18,211	10,298
G_F	2,913	3,981
Total	37,780	26,439

(c) Predicted and observed numbers of full-length and active young mouse L1s.

L1 class	# predicted FL L1s	# predicted active L1s	# reference genome FL L1s
A	6,500(Saxton and Martin 1998)	900(Goodier et al. 2001)	3,514
T_F	3,000 – 4,800(DeBerardinis et al. 1998)	1,800-3,000(DeBerardinis et al. 1998)	3,438
G_F	1,500(Goodier et al. 2001)	400(Goodier et al. 2001)	704

The number of full-length (FL) L1 elements in the reference C57 genome is defined by a length > 5,000 nt and the presence of at least one 5' UTR monomer.

Supplementary Table 4. **Pairwise comparison of polymorphic L1s in unassembled strains.**

		129S1/SVIMJ		129X1/SVJ		A/J		DBA/2J	
		A	P	A	P	A	P	A	P
129S1/SVIMJ	A	550	0	422	128	232	318	234	316
		29.6%	0.0%	22.7%	6.9%	12.5%	17.1%	12.6%	17.0%
	P	0	1311	141	1170	767	544	737	574
		0.0%	70.4%	7.6%	62.9%	41.2%	29.2%	39.6%	30.8%
129X1/SVJ	A			563	0	233	330	230	333
				30.3%	0.0%	12.5%	17.7%	12.4%	17.9%
	P			0	1298	766	532	741	557
				0.0%	69.7%	41.2%	28.6%	39.8%	29.9%
A/J	A					999	0	422	577
						53.7%	0.0%	22.7%	31.0%
	P					0	862	549	313
						0.0%	46.3%	29.5%	16.8%
DBA/2J	A							971	0
								52.2%	0.0%
	P							0	890
								0.0%	47.8%

The absence or presence (A/P) status of L1 polymorphisms was counted only when known for all five strains. Each element must be present in the C57 reference genome and absent from at least one of the four unassembled strains.

Supplementary Table 5. **Validation of selected polymorphisms in mouse strains and subspecies – additional information.** The data are arranged in the same layout as Table 2. Note that A/J " weakvariant" trace 1045646480 and its mate 1045646822 align to chr. 14: ~ 39450000, as do 1018921985 and mate 1018922407.

traceid	L1 name in reference database; Genbank accession nos.	length	absent from strains as per mm8 reference	No. sup	rel orient	full gene name
1042722486	1_5810074_L1Md_A	5,605	DBA2J	2		
Rd7donorL1	4_21741132_L1Md_T	7,243	129X1, AJ, DBA2J	4		
1090553782	10_10513926_L1Md_A	979	DBA2J	3	AS	Metabotropic glutamate receptor 1 precursor (mGluR1)
1091069131	10_10520620_L1Md_F2	487	DBA2J	2	sense	Metabotropic glutamate receptor 1 precursor (mGluR1)
1100631474	10_13352437_L1Md_T	881	DBA2J	3	sense	androgen-induced 1
1097537874	10_13602567_L1Md_A	1,082	none	2		
1097610660	10_13964068_L1Md_F3	551	DBA2J	3		
1098874361	10_102162706_L1Md_T	452	DBA2J	2		
1099621093	10_102279770_L1Md_T	380	AJ, DBA2J	5		
1035280108	10_102320921_L1Md_T	1,072	129X1, AJ	7		
1083301601	10_105033229_L1Md_A	829	AJ, DBA2J	3		
1043213053	10_107383036_L1Md_A	2,280	AJ	3		
1047671029	10_110582693_L1Md_A	567	DBA2J	1	AS	oxysterol-binding protein-like protein 8 isoform
1038614451	11_14815876_L1Md_T	6,275	129S1,129X1,AJ,DBA2J	6		
1030700574	12_35318651_L1Md_T	6,519	129S1,129X1,A/J,DBA/2J	8	AS	histone deacetylase 9
1019767717	13_92202618_L1Md_A	7,053	AJ, DBA/2J	3	AS	AK048302
1042769406	14_39455316_L1Md_T	7,584	DBA/2J	2	AS	
1083543712	15_20476076_L1Md_A	6,421	129S1, 129X1, A/J	3	AS	
1072826857	16_59113811_L1Md_T	7,095	129S1, 129X1	4		
1018878835	17_39025598_L1Md_T	5,835	A/J	2		
1030552568	18_19954996_L1Md_A	6,194	129S1, DBA/2J	4		
1097301560	19_13837558_L1Md_T	6,676	DBA/2J	2		
4ASIII2_1	ND; EF591881		C57BL/6J; expect insertion trace(s)			
8AS1_1	ND; EF591882		C57BL/6J; expect insertion trace(s)			poly (ADP-ribose) polymerase family, member 8
11ASII1	ND; EF591883		C57BL/6J; expect insertion trace(s)			Drosha, RNaseIII
1ASII1	ND; EF591880		C57BL/6J; expect insertion trace(s)			
7ASIII4_2	2_43990002_L1Md_T	6,648	129X1, A/J	4	AS	Rho GTPase activating protein 15 isoform 1
9AS1_1	16_95119881_L1Md_T	6,209	none			
5ASII	2_66106851_L1Md_F2	6,362	none; non-polymorphic		AS	sodium channel 3
7ASIII2_1A	12_78323702_L1Md_T	6,439	none; non-polymorphic		AS	fucosyltransferase 8
7ASIII2_1B	12_78302247_L1Md_T	6,269	none; non-polymorphic		AS	fucosyltransferase 8

Supplementary Table 6. **Structural features of L1 polymorphisms.**

a) Y**oung, active L1 retrotransposons by length and presence in strains.**

		L1 subfamily			
presence in 5 strains		**A**	**G_F**	**T_F**	**young L1 subfamilies**
non-polymorphic	full-length	471	48	222	741
	N	1,292	330	840	2,462
	percent	36.46%	14.55%	26.43%	30.10%
undetermined	full-length	2,838	567	2,499	5,904
	N	9,298	3,116	7,589	20,003
	percent	30.52%	18.20%	32.93%	29.52%
polymorphic (insert in C57)	full-length	651	124	939	1,714
	N	1,570	535	1,869	3,974
	percent	41.46%	23.18%	50.24%	43.13%
total	full-length	3,960	739	3,660	8,359
	N	12,160	3,981	10,298	26,439
	percent	32.57%	18.56%	35.54%	31.62%
polymorphic vs. non-polymorphic L1s	significance (p-value)	0.006	4.99E-31	0.002	1.30E-25

The number and percentage of full-length vs. total (any length) L1 elements comprising each young active subfamily (L1 A, T_F and G_F) are indicated for non-polymorphic, polymorphic and unknown categories. For members of young, active L1 subfamilies, L1 polymorphisms are significantly more full-length than are non-polymorphic L1s. The p-values were calculated by chi square tests.

b) **TSDs and poly(A) tails in L1 subfamilies.**

				L1 subfamily		
presence in 5 strains	**TSD type**		**A**	**G_F**	**T_F**	**all L1s**
non-polymorphic	non-TSD	count	369	132	223	84,282
		percent	28.56%	40.00%	26.55%	65.95%
	TSD-only	count	657	147	409	36,385
		percent	50.85%	44.55%	48.69%	28.47%
	TSD and poly(A) tail	count	266	51	208	7,136
		percent	20.59%	15.45%	24.76%	5.58%
	total	count	1,292	330	840	127,803
undetermined	non-TSD	count	2,654	1,211	2,053	344,123
		percent	28.54%	38.86%	27.05%	64.73%
	TSD-only	count	4,829	1,422	3,688	154,391
		percent	51.94%	45.64%	48.60%	29.04%
	poly(A)-tailed TSD	count	1,815	483	1,848	33,288
		percent	19.52%	15.50%	24.35%	6.26%
	total	count	9,298	3,116	7,589	531,602
polymorphic (insertion in C57)	non-TSD	count	261	138	300	2,268
		percent	16.62%	25.79%	16.05%	33.73%
	TSD-only	count	903	272	955	3,082
		percent	57.52%	50.84%	51.10%	45.84%
	poly(A)-tailed TSD	count	406	125	614	1,373
		percent	25.86%	23.36%	32.85%	20.42%
	total	count	1,570	535	1,869	6,723
total	non-TSD	count	3,284	1,481	2,576	430,673
		percent	27.01%	37.20%	25.01%	64.63%
	TSD-only	count	6,389	1,841	5,052	193,858
		percent	52.54%	46.24%	49.06%	29.09%
	poly(A)-tailed TSD	count	2,487	659	2,670	41,797
		percent	20.45%	16.55%	25.93%	6.27%
	total	count	12,160	3,981	10,298	666,328
polymorphic vs. non-polymorphic L1s		p-values	7.38E-12	2.94E-05	3.11E-11	0.000

The presence of TSDs and poly(A) tails adjacent to L1 T_F, A, and G_F integrants is indicated for non-polymorphic, non-polymorphic and unknown categories. Members of each of the young active L1 subfamilies are significantly more likely to have both TSDs and poly(A) tails than to lack them. Significance values of chi-square tests were computed for 2x2 tables. The test compared the frequencies of L1 without TSD to the frequencies of L1 with poly(A)-tailed TSD in two groups: polymorphic L1s and non-polymorphic L1s (highlighted in *yellow*).

c) **Poly(A) tail lengths in young, active L1 subfamilies.**

poly(A) tail length

		L1 subfamily members with poly(A) tails			
L1 integrant status in 5 strains		**A**	**G_F**	**T_F**	**all L1s**
non-polymorphic	mean	18.91	18.61	19.54	18.72
	std. dev.	8.495	7.571	8.567	10.985
	N	266	51	208	7,136
undetermined	mean	19.86	20.61	20.51	19.61
	std. dev.	0.551	10.965	10.715	11.895
	N	1,815	483	1,848	33,288
polymorphic (insertion in C57)	mean	20.39	21.50	21.54	21.02
	std. dev.	8.810	11.175	11.451	10.588
	N	406	125	614	1,373
total	mean	19.84	20.62	20.67	19.51
	std. dev.	9.330	10.790	10.749	11.711
	N	2,487	659	2,670	41,797
t-test		**A**	**G_F**	**T_F**	**all L1s**
polymorphic vs. non-polymorphic L1s	significance (p-value)	0.031	0.049	0.008	9.94E-13

The lengths of poly(A) tails at the 3' end of L1 T_F, A, and G_F elements are indicated for non-polymorphic, polymorphic and unknown categories. A two-sided t-test was performed to compare significance of difference between polymorphic vs. non-polymorphic L1s (*yellow*).

d) **Nucleotide substitution rates in L1 subfamilies.**

		substitution rate			
		L1 subfamily			
L1 integrant status in 5 strains		**A**	**G_F**	**T_F**	**all L1s**
non-polymorphic	mean (%)	3.62	6.58	3.75	20.98
	std. dev.	4.207	3.787	2.112	9.163
	N	1,292	330	840	127,803
undetermined	mean (%)	3.43	5.77	3.50	20.19
	std. dev.	3.702	3.118	1.928	9.492
	N	9,298	3,116	7,589	531,802
polymorphic (insertion in C57)	mean (%)	1.45	5.08	2.76	8.41
	std. dev.	1.635	1.748	1.386	9.593
	N	1,570	535	1,869	6,723
total	mean (%)	3.20	5.74	3.39	20.22
	std. dev.	3.627	3.053	1.882	9.511
	N	12,160	3,981	10,298	666,328

t-test		**A**	**G_F**	**T_F**	**all L1s**
polymorphic vs. non-polymorphic L1s	significance (p-value)	3.23E-63	4.50E-11	1.24E-33	0.00

Substitution rates as a function of L1 class and polymorphism are shown. Polymorphic L1 members of young, active subfamilies are significantly more likely to have decreased nucleotide substitution (when aligned to consensus subfamily elements) compared to non-polymorphic or undetermined L1s. A two-sided t-test was performed to compared significance of difference between polymorphic vs. non-polymorphic L1s (*yellow*).

Supplementary Table 7. **Gene orientation and GC content near L1 retrotransposons.**

L1 integrant status in 5 strains		non-polymorphic		unknown		polymorphic		total	
target type	orientation	count	percent	count	percent	count	percent	count	percent
no gene	NA	59,977	46.93%	251,197	47.24%	3,358	49.95%	314,532	47.20%
Inside	antisense	17,899	68.12%	71,093	66.86%	785	58.11%	89,777	67.02%
	sense	8,377	31.88%	35,243	33.14%	566	41.89%	44,186	32.98%
	total	26,276	20.56%	106,336	20.00%	1,351	20.10%	133,963	20.10%
3’	antisense	10,166	52.26%	42,263	52.09%	523	55.05%	52,952	52.15%
	sense	9,285	47.74%	38,879	47.91%	427	44.95%	48,591	47.85%
	total	19,451	15.22%	81,142	15.26%	950	14.13%	101,543	15.24%
5’	antisense	10,882	49.24%	45,986	49.38%	500	46.99%	57,368	49.33%
	sense	11,217	50.76%	47,141	50.62%	564	53.01%	58,922	50.67%
	total	22,099	17.29%	93,127	17.51%	1,064	15.83%	116,290	17.45%
Total		127,803	100.00%	531,802	100.00%	6,723	100.00%	666,328	100.00%
percent GC content		39.81				39.68			

Supplementary Table 8. **Primary sequences and NCBI GenBank accession numbers for fusion L1-gene transcripts and polymorphic L1 integrant loci.** Internal L1 sequences have not been determined fully and are estimated by the size of spanning PCR amplicons.

Clone ID	GenBank accession no.	Gene name
7ASIII4-2	EF591871	*Arghap15*
5ASII	EF591872	*Scn1a*
1ASII-1	EF591873	previously unannotated
7ASIII2-1B	EF591874	*Fut8*
7ASIII2-1A	EF591875	*Fut8*
4ASIII2-1	EF591876	AK129128
8AS1-1	EF591877	*Parp8*
11ASII-1	EF591878	*Drosha (Rnasen)*
9AS1-1	EF591879	previously unannotated
1ASII-1int	EF591880	previously unannotated
4ASIII2-1/4ASI-1int	EF591881	AK129128
8AS1-1int	EF591882	*Parp8*
11ASII-1/2ASII-1int	EF591883	*Drosha (Rnasen)*

Supplementary Table 9. **Variable expression of fusion transcripts from full-length, polymorphic, AS L1s in reference genes.**

chr coordinates	strains	129S1	129X1	A/J	DBA/2J	GenBank acc.	gene name	L1 subfamily	location in gene	dist to exon
chr1:8678199-8685197	1	0	0	1	0	NM_027671	Sntg1	A	intron6	6236
chr1:9048084-9054491	2	0	2	1	0	NM_027671	Sntg1	T_F	intron2	58901
chr1:9165332-9171910	2	0	2	2	0	NM_027671	Sntg1	A	intron2	21202
chr1:11699606-11705927	1	0	0	1	0	NM_177173	A830018L16Rik	T_F	intron7	27251
chr1:11708434-11715071	2	1	0	1	0	NM_177173	A830018L16Rik	A	intron7	18107
chr1:11762011-11768993	3	0	2	1	1	NM_177173	A830018L16Rik	T_F	intron8	14701
chr1:20417355-20424773	1	0	0	1	0	NM_153179	Pkhd1	T_F	intron36	10598
chr1:20559648-20566214	1	0	0	2	0	NM_153179	Pkhd1	A	intron11	2996
chr1:24450608-24457191	1	0	0	0	1	NM_007733	Col19a1	A	intron11	5267
chr1:25535375-25541866	2	0	0	1	1	NM_175642	Bai3	T_F	intron2	30964
chr1:25606603-25613866	1	0	1	0	0	NM_175642	Bai3	T_F	intron2	102192
chr1:32316912-32323281	1	0	0	1	0	NM_133235	Khdrbs2	A	intron3	36135
chr1:44546645-44553802	3	0	2	2	2	NM_027506	Gulp1	A	intron1	50227
chr1:66204276-66210957	3	0	1	3	2	NM_001039934	Mtap2	G_F	intron3	19356
chr1:106790167-106797160	2	0	2	2	0	NM_011800	Cdh20	A	intron7	1207
chr1:107395580-107401974	1	0	0	4	0	NM_013784	Pign	T_F	intron22	1783
chr1:128619754-128625904	3	0	2	1	1	NM_176957	D130011D22Rik	T_F	intron1	31958
chr1:148366273-148372566	1	0	0	0	1	NM_153539	B830045N13Rik	T_F	intron2	72227
chr1:162257998-162264984	1	0	0	0	1	NM_013862	Rabgap1l	A	Intron17	13515
chr1:163900252-163906855	1	0	0	0	1	NM_001038619	Dnm3	T_F	intron17	5596
chr1:164175158-164182062	1	0	0	1	0	NM_001038619	Dnm3	G_F	intron4	8132
chr1:175669610-175676103	1	0	1	0	0	NM_008329	Ifi204	T_F	intron3	66708
chr1:176544520-176550723	3	0	2	2	2	NM_019445	Fmn2	A	intron14	5796
chr1:178729985-178736580	2	0	2	1	0	NM_029756	Sdccag8	A	intron12	15204
chr1:188763441-188770121	1	0	0	1	0	NM_028848	Spata17	T_F	intron9	14164
chr1:190036535-190043492	2	0	1	0	3	NM_021408	Ush2a	G_F	intron3	8332
chr10:10645980-10652380	3	0	1	2	1	NM_016976	Grm1	T_F	intron2	19212
chr10:10665554-10672065	2	1	0	1	0	NM_016976	Grm1	T_F	intron2	38786
chr10:26944089-26953707	1	0	0	1	0	NM_008481	Lama2	T_F	intron14	626
chr10:42063175-42070425	1	0	1	0	0	NM_145743	Lace1	A	intron7	14513
chr10:48914197-48919077	1	0	1	0	0	NM_010349	Grik2	T_F	intron13	9916
chr10:113939581-113946592	2	1	0	0	1	NM_146241	Trhde	T_F	intron6	17275
chr10:122893619-122900243	2	0	0	1	1	NM_182807	AI851790	A	intron2	59636
chr11:26361459-26367854	2	0	0	1	2	NM_025923	Fancl	A	intron8	480
chr11:36185856-36192275	2	0	0	1	1	NM_011856	Odz2	T_F	intron4	36266
chr11:46602963-46610042	2	0	2	2	0	NM_134248	Havcr1	T_F	intron3	21
chr11:62784528-62790896	1	0	0	1	0	NM_025496	Cdrt4	A	intron1	16952
chr11:81247063-81253478	1	0	0	2	0	NM_001034013	Accn1	A	intron1	459254
chr11:104465308-104471679	1	0	0	1	0	NM_016780	Itgb3	A	intron10	3133

chr11:107921941-107928609	1	0	0	0	3	NM_011101	Prkca	T_F	intron3	47978
chr12:37665034-37671832	2	0	0	2	1	NM_008584	Meox2	T_F	intron1	5749
chr12:38548289-38554585	1	0	0	1	0	NM_178681	Dgkb	G_F	intron2	39963
chr12:41647307-41653691	1	0	0	0	1	NM_053122	Immp2l	A	intron4	25941
chr12:53031427 - 53038069	2	2	0	2	0	NM_029760	Nubpl	T_F	Intron4	4412
chr12:83552473-83559347	1	0	0	0	3	NM_015812	Rgs6	T_F	intron1	11309
chr12:83702324-83708519	1	0	0	0	1	NM_015812	Rgs6	G_F	intron2	131564
chr12:91569522-91576320	1	0	0	0	2	NM_181815	4930534B04Rik	A	intron20	42801
chr12:91872982-91880525	2	0	0	2	1	NM_011648	Tshr	T_F	intron1	18335
chr12:116983009-116989624	1	0	0	3	0	NM_011215	Ptprn2	A	intron1	32783
chr12:116999285-117005661	2	0	1	0	1	NM_011215	Ptprn2	A	intron1	16746
chr12:117289453-117295493	1	0	0	1	0	NM_011215	Ptprn2	T_F	intron5	5358
chr13:4064764-4075288	2	0	2	2	0	NM_134072	Akr1c14	T_F	intron3	77
chr13:19066726-19073232	1	1	0	0	0	NM_175007	Amph	T_F	intron2	11563
chr13:59111100-59117574	1	0	0	1	0	NM_001025074	Ntrk2	T_F	intron16	18481
chr14:9402036-9408477	1	1	0	0	0	NM_010210	Fhit	T_F	intron4	114823
chr14:34028687-34035616	1	0	1	0	0	NM_008166	Grid1	G_F	intron4	45169
chr14:34226630-34233061	2	1	2	0	0	NM_008166	Grid1	T_F	intron8	40244
chr14:34356295-34362958	1	0	2	0	0	NM_008166	Grid1	A	intron13	6319
chr14:39753081-39759875	1	0	0	0	2	NM_177816	Sh2d4b	A	intron4	1778
chr14:75804807-75811814	1	0	2	0	0	NM_175369	Ccdc122	A	intron3	1703
chr14:85772172-85778665	1	0	0	2	0	NM_019670	Diap3	T_F	intron17	18523
chr14:122547162-122555311	1	0	0	1	0	NM_177393	Vgcnl1	A	Intron12	2149
chr14:123674576-123681530	2	0	2	2	0	NM_207667	Fgf14	T_F	intron1	178207
chr15:21280148-21287315	1	0	1	0	0	NM_001008420	Cdh12	T_F	intron2	16132
chr15:42971456-42978185	1	0	1	0	0	NM_172815	Rspo2	T_F	intron2	21619
chr15:43347033-43353840	2	0	1	1	0	NM_025736	4921531G14Rik	A	intron10	3110
chr15:44299187-44306422	2	0	1	1	0	NM_138674	Pkhd1l1	A	intron4	1203
chr15:54274396-54280862	2	0	0	1	3	NM_173422	Colec10	T_F	intron3	5082
chr15:71480024-71486814	1	0	1	0	0	NM_177819	A830008O07	T_F	intron1	28258
chr16:11773215-11780072	1	0	0	0	4	NM_146067	C530044N13Rik	T_F	intron2	21083
chr16:28415494-28422500	1	0	0	1	0	NM_010199	Fgf12	A	intron2	101873
chr16:45484189-45490573	1	0	0	1	0	NM_198106	Slc9a10	A	intron13	2086
chr16:59836222-59842995	2	0	1	0	1	NM_007938	Epha6	T_F	intron11	15044
chr16:64514474-64521088	1	0	0	0	3	NM_173861	Ckt2	A	intron1	28565
chr16:66672539-66679189	1	0	1	0	0	NM_178721	Igsf4d	A	intron8	6902
chr16:70434987-70441772	1	1	0	0	0	NM_028803	Gbe1	A	intron14	1931
chr16:74088508-74094963	1	0	0	0	1	NM_175549	Robo2	T_F	intron2	140267
chr17:10831296-10836431	1	1	0	0	0	NM_016694	Park2	T_F	intron1	74055
chr17:11336892-11343759	3	2	0	3	2	NM_016694	Park2	T_F	intron6	740
chr17:11597852-11603945	1	0	0	0	1	NM_016694	Park2	T_F	intron7	78014
chr17:11853643-11862696	1	0	0	0	1	NM_016694	Park2	A	intron10	6157
chr17:17258719-17265933	2	0	1	3	0	NM_172827	Lnpep	T_F	intron9	779
chr17:21248130-21256413	3	0	1	1	2	NM_144515	Zfp52	T_F	intron2	3268
chr17:44566936-44574510	2	1	0	2	0	NM_178652	Supt3h	A	intron11	8213
chr17:57053720-57060073	1	0	1	0	0	NM_010130	Emr1	T_F	intron2	6651
chr18:37157787-37164265	3	0	2	1	2	NM_138663	Pcdha12	T_F	intron1	9190
chr18:53685854-53692113	1	0	0	1	0	NM_001033281	Prdm6	T_F	intron5	1411
chr18:79023515-79030028	2	0	2	3	0	NM_053099	Setbp1	T_F	intron2	1505

chr19:11660086-11666881	1	0	0	4	0	NM_022431	Ms4a11	A	intron3	120
chr19:22711341-22718778	3	0	2	1	1	NM_177341	Trpm3	A	intron1	45740
chr19:32099385-32105758	1	1	0	0	0	NM_018830	Asah2	G_F	intron10	3719
chr19:39465743-39473138	1	0	0	0	1	NM_010002	Cyp2c38	A	intron5	7630
chr19:48279126-48285597	2	0	0	1	2	NM_025696	Sorcs3	T_F	intron1	19018
chr19:50604387-50611198	2	0	0	1	1	NM_021377	Sorcs1	A	intron1	75624
chr2:22216730-22222947	2	0	0	1	1	NM_148413	Myo3a	A	intron8	1258
chr2:22281003-22287199	2	0	0	1	1	NM_148413	Myo3a	F3	intron15	1106
chr2:41895503-41901678	1	1	0	0	0	NM_053011	Lrp1b	T_F	intron2	112716
chr2:41905457-41912335	1	0	0	4	0	NM_053011	Lrp1b	T_F	intron2	122670
chr2:42391707-42397997	1	0	0	0	2	NM_053011	Lrp1b	T_F	intron1	76746
chr2:43990000-43996660	2	0	2	1	0	NM_153820	Arhgap15	T_F	intron10	1719
chr2:54408963-54416013	3	1	1	0	1	NM_173030	Galnt13	A	intron3	76896
chr2:62661888-62668147	1	0	1	0	0	NM_133207	Kcnh7	T_F	intron3	9893
chr2:76251404-76258000	2	0	1	0	1	NM_145525	Osbpl6	A	intron1	42095
chr2:102403734-102410545	1	0	0	2	0	NM_173749	E430002G05Rik	A	intron4	155
chr2:105743245-105749843	2	0	0	1	2	NM_028260	Immp1l	A	intron5	5288
chr2:139592079-139600649	1	0	0	0	1	NM_175225	Tasp1	A	intron11	1324
chr2:143316681-143323281	1	0	0	0	1	NM_008792	Pcsk2	A	intron2	51693
chr2:144927733-144934117	2	0	1	0	2	NM_053195	Slc24a3	A	exon1	-1
chr2:161837475-161844326	2	1	0	1	0	NM_021464	Ptprt	G_F	intron6	10529
chr3:26953018-26959292	1	0	0	0	2	NM_027583	Spata16	G_F	intron3	9285
chr3:32352816-32359464	3	1	0	1	1	NM_028231	Kcnmb2	A	intron3	5769
chr3:58940075-58946886	1	0	0	1	0	NM_153384	Ush3a	A	intron3	5695
chr3:65277296-65285455	2	1	0	0	2	NM_010597	Kcnab1	T_F	intron1	68964
chr3:76172800-76179209	2	0	0	1	1	NM_178673	Fstl5	A	intron1	11899
chr3:80828350-80835035	2	0	2	3	0	NM_001039195	Gria2	T_F	intron2	1225
chr3:86415222-86422346	1	0	0	2	0	NM_030695	Lrba	T_F	intron22	939
chr3:100254609-100261524	2	0	0	3	1	NM_028892	Spag17	T_F	intron47	106
chr3:102961710-102968400	1	0	0	1	0	NM_011516	Sycp1	T_F	intron30	1522
chr3:125894814-125900994	1	0	0	3	0	NM_011674	Ugt8a	T_F	intron2	5665
chr3:141541027-141548015	2	0	0	2	1	NM_009472	Unc5c	T_F	intron1	67358
chr3:141697348-141704009	2	0	2	2	0	NM_009472	Unc5c	T_F	intron5	1821
chr3:156660245-156667112	3	0	1	1	1	NM_177274	Negr1	A	intron1	127617
chr4:8118890-8125130	1	0	0	1	0	NM_007592	Car8	T_F	intron3	2351
chr4:9512604-9519865	2	0	1	0	1	NM_023066	Asph	A	intron11	1601
chr4:12880644-12887119	2	0	0	1	1	NM_173746	C130086A10	T_F	intron2	2972
chr4:35231502-35237723	1	0	0	0	1	NM_178061	Mobkl2b	T_F	intron2	34933
chr4:65841374-65848487	1	0	0	2	0	NM_019514	Astn2	T_F	intron1	41756
chr4:69826506-69832701	2	1	0	0	1	NM_145990	Cdk5rap2	A	intron11	2586
chr4:90045770-90052266	1	0	1	0	0	NM_153096	Zfp353	T_F	intron4	327887
chr4:94283592-94290200	2	0	0	1	2	NM_013690	Tek	A	intron2	428
chr4:94404625-94411613	2	0	2	1	0	NM_027089	4930579C15Rik	T_F	intron7	327
chr4:101251511-101257860	1	0	0	2	0	NM_010704	Lepr	A	intron7	1374
chr5:3987531-3994570	3	2	0	2	1	NM_194462	Akap9	T_F	intron14	311
chr5:12548667-12555109	2	1	1	0	0	NM_028882	Sema3d	A	intron10	10
chr5:21725851-21732250	2	1	1	0	0	NM_011261	Reln	A	intron3	7098
chr5:27689680-27696387	1	0	0	2	0	NM_010075	Dpp6	G_F	intron1	17154
chr5:72257979-72264056	1	0	0	0	2	NM_008069	Gabrb1	T_F	intron4	44905

chr5:82646162-82652399	1	0	1	0	0	NM_198702	Lphn3	A	intron10	10658
chr5:98229019-98236145	1	0	0	0	2	NM_133738	Antxr2	T_F	intron7	7383
chr5:102885436-102892267	1	0	0	0	1	NM_146161	Arhgap24	A	intron1	12117
chr5:105262401-105268818	1	0	0	0	5	NM_029509	5830443L24Rik	T_F	intron4	2312
chr5:128164138-128172293	1	0	0	0	1	NM_172885	Tmem132d	G_F	intron4	10197
chr6:9089107-9095538	2	0	0	1	2	NM_008751	Nxph1	T_F	intron1	101546
chr6:33589143-33595591	1	0	1	0	0	NM_009148	Exoc4	T_F	intron10	78907
chr6:36154436-36160835	1	0	0	3	0	NM_001033377	9330158H04Rik	A	Intron4	12070
chr6:45348678-45355085	1	0	0	2	0	NM_001004357	Cntnap2	A	intron1	228511
chr6:45409274-45415970	1	0	0	2	0	NM_001004357	Cntnap2	T_F	intron1	167626
chr6:63509387-63515893	1	0	1	0	0	NM_008167	Grid2	A	intron2	76362
chr6:64307605-64313847	1	0	0	0	3	NM_008167	Grid2	G_F	intron11	10924
chr6:95623527-95630225	2	0	0	2	2	NM_011507	Suclg2	T_F	intron1	2465
chr6:96887385-96893788	1	0	0	1	0	NM_177233	C130034I18Rik	T_F	intron3	58007
chr6:103540759-103548025	2	1	1	0	0	NM_007697	Chl1	T_F	intron2	14871
chr6:108231651-108238996	1	0	0	1	0	NM_010585	Itpr1	G_F	intron4	24540
chr6:111381312-111387899	1	0	0	1	0	NM_177328	Grm7	T_F	intron8	56455
chr6:112618450-112625448	1	0	0	0	2	NM_021385	Rad18	T_F	intron9	2836
chr6:114725043-114731214	1	0	0	0	2	NM_028835	Atg7	A	intron17	11361
chr6:129822478-129829560	1	0	2	0	0	NM_133203	Klra17	T_F	intron6	1798
chr6:146875952-146882231	1	0	0	1	0	NM_026221	Ppfibp1	T_F	intron1	13844
chr7:16280016-16286317	1	1	0	0	0	NM_007676	Psg16	G_F	intron5	2779
chr7:49922040-49928999	1	0	0	0	1	NM_001037906	Nell1	T_F	intron2	1639
chr7:50009969-50017405	1	0	1	0	0	NM_001037906	Nell1	A	intron4	21248
chr7:50063997-50070316	1	0	0	2	0	NM_001037906	Nell1	G_F	intron5	17286
chr7:55002138-55009040	1	0	0	0	1	NM_178705	Luzp2	G_F	intron5	26178
chr7:55943092-55949795	2	1	0	1	0	NM_010418	Herc2	T_F	intron4	2643
chr7:56819560-56826130	2	0	1	0	1	NM_008074	Gabrg3	A	intron5	24513
chr7:59162384-59168137	1	0	0	0	1	NM_001033962	Ube3a	T_F	intron9	3399
chr7:64490291-64496709	1	0	0	3	0	NM_007461	Apba2	A	intron2	17271
chr7:67152876-67159412	1	0	1	0	0	NM_001033713	Mef2a	T_F	intron4	5526
chr7:89558468-89564188	2	0	0	4	1	NM_181407	Me3	T_F	intron1	47631
chr7:89626117-89634603	1	0	0	0	1	NM_181407	Me3	A	intron3	10969
chr7:91616784-91623619	2	0	0	4	1	NM_011807	Dlgh2	A	intron3	8448
chr7:96066696-96073721	1	0	0	0	1	NM_011858	Odz4	T_F	intron1	20011
chr7:97016144-97022588	2	0	2	1	0	NM_010248	Gab2	T_F	intron1	59046
chr7:133938788-133945823	1	0	0	2	0	NM_007400	Adam12	A	intron3	27238
chr8:16216361-16222863	1	0	0	0	1	NM_053171	Csmd1	T_F	intron17	337
chr8:16361110-16367113	1	0	2	0	0	NM_053171	Csmd1	A	intron8	2681
chr8:39022462-39028961	2	0	2	4	0	NM_145841	Sgcz	G_F	intron4	2696
chr8:39040809-39047167	1	0	0	1	0	NM_145841	Sgcz	T_F	intron3	9029
chr8:39095784-39102293	2	1	0	1	0	NM_145841	Sgcz	T_F	intron3	6156
chr8:39133214-39139435	2	0	1	1	0	NM_145841	Sgcz	A	intron2	24677
chr8:39313950-39320299	1	0	0	0	1	NM_145841	Sgcz	G_F	intron1	101080
chr8:65960989-65968051	2	0	0	1	2	NM_023689	Spock3	T_F	intron3	37681
chr8:68633078-68641000	1	0	0	1	0	NM_175188	March1.	T_F	intron1	85326
chr8:68759481-68766057	1	0	0	3	0	NM_175188	March1.	A	intron1	42261
chr8:71498913-71505117	2	0	0	4	1	NM_172753	4732435N03Rik	A	intron2	12804
chr8:75864022-75870567	2	0	1	0	5	NM_010687	Large	T_F	intron6	55790

chr8:84850325-84857221	2	0	0	1	1	NM_001024617	Inpp4b	G_F	intron11	1786
chr8:84930266-84936585	3	1	1	1	0	NM_001024617	Inpp4b	F2	intron17	15
chr8:88803986-88810627	2	0	1	1	0	NM_199446	Phkb	A	intron7	1604
chr8:94307055-94313971	1	1	0	0	0	NM_011936	Fto	G_F	intron3	7109
chr9:53757309-53763536	3	0	1	2	1	NM_177769	Elmod1	T_F	intron1	9625
chr9:55764431-55770807	4	1	3	1	1	NM_183111	4930563M21Rik	A	intron7	1019
chr9:59843505-59849814	2	0	1	0	1	NM_172444	Thsd4	T_F	intron2	5042
chr9:70702943-70709466	1	0	1	0	0	NM_008280	Lipc	T_F	intron1	24226
chr9:106776523-106782837	1	0	0	1	0	NM_153413	Dock3	G_F	intron41	3757
chrX:44110999-44118357	3	1	0	3	1	NM_053123	Smarca1	T_F	intron9	424
chrX:65890227-65897096	1	0	1	0	0	NM_008032	Aff2	T_F	intron3	57512
chrX:68865579-68872496	1	0	0	1	0	NM_008067	Gabra3	G_F	intron1	36230
chrX:94807349-94812337	1	0	1	0	0	NM_052976	Ophn1	A	intron19	7455
chrX:96204427-96212062	2	0	1	1	0	NM_010099	Eda	A	intron1	25532
chrX:110650043-110656468	1	0	0	1	0	NM_172781	Klhl4	A	intron5	2864
chrX:136807474-136815063	1	0	0	3	0	NM_007736	Col4a5	T_F	intron5	2267
chrX:138109216-138115539	1	0	0	1	0	NM_019496	Ammecr1	T_F	intron3	3877
chrX:145955068-145961465	1	0	0	1	0	NM_008824	Pfkfb1	A	intron5	19
chrX:152627074-152633781	1	0	0	1	0	NM_011077	Phex	T_F	intron12	14475

205 full-length AS L1 polymorphisms, absent from at least one of the unassembled strains, were identified within the coding sequence of RefSeq genes in the reference genome. Almost all are intronic. The number of strains with the L1 polymorphism absent is presented along with the number of WGS indel traces for each strain. Zeroes indicate no sequence traces supporting an indel, i.e. either well-aligned traces or insufficient sequence coverage. RT-PCR assays for both L1 fusion transcripts and native transcripts were performed on total RNA isolated from adult testes for each of the five mouse strains (*yellow highlights*). Fusion transcripts were detected in adult testes only from strains containing the L1 polymorphisms (*green*). At *Arhgap15*, a fusion L1-gene transcript was detected in a different screen (*blue*; see Table 2).

Supplementary Table 10. **Significant changes in non-polymorphic L1s or reference L1s within genes in various ontological categories.**

Annotated genes containing 26,104 distinct intronic L1 non-polymorphisms (integrants present in all five strains) or 132,849 reference L1s were identified, respectively. They were assigned to top-level ontological categories (including biological processes and molecular functions) using Gene Ontology (GO) Panther software. Because more than one ontological category can be assigned to a given gene, a total of 44,715 or 226,409 assignments were made for these non-polymorphic or reference L1s, respectively. Only statistically significant differences in ontological categories (corrected p-values < 0.01) are listed here. Enrichments or exclusions of polymorphic L1s and additional information about methods are presented in Table 4.

(A) - (D) *In silico* simulations resulted in 2,045,793 random “integrants”. Their annotated (A) and (B) biological processes and (C) and (D) molecular functions were determined for intronic (A) and (C) non-polymorphic or (B) and (D) reference L1 integrants.

(E) and (F) We compared the ontological categories of non-polymorphic L1 genes against reference L1 genes, identifying their annotated (E) biological processes and (F) molecular functions (Mi et al. 2005).

(A) Non-polymorphic L1s vs random simulation: biological processes

biological process	intronic non-polymorphic L1, %	random simulation, %	fold-change	p-value
Neuronal activities	9.46	7.07	1.34	0.00E+00
Nucleoside, nucleotide and nucleic acid metabolism	10.78	13.42	0.80	1.03E-36
Sensory perception	1.75	2.32	0.75	3.83E-09
Cell adhesion	7.22	6.37	1.13	5.33E-07
Biological process unclassified	30.04	28.62	1.05	7.78E-06
Oncogenesis	2.11	2.57	0.82	2.98E-05
Other metabolism	1.95	2.38	0.82	3.05E-05
Transport	9.3	8.6	1.08	1.28E-03
Signal transduction	24.5	23.47	1.04	1.60E-03
Amino acid metabolism	0.77	0.98	0.79	8.07E-03

(B) Reference L1s vs random simulation: biological processes

biological process	intronic reference L1, %	random simulation, %	fold-change	p-value
Neuronal activities	9.46	7.07	1.34	0.00E+00
Cell adhesion	7.12	6.37	1.12	0.00E+00
Transport	9.29	8.6	1.08	0.00E+00
Signal transduction	24.58	23.47	1.05	0.00E+00
Lipid fatty acid and steroid metabolism	4.5	4.05	1.11	0.00E+00
Nucleoside, nucleotide and nucleic acid metabolism	11.13	13.42	0.83	1.18E-137
Sensory perception	1.85	2.32	0.80	3.08E-31
Cell structure and motility	6.46	7.21	0.90	1.72E-25
Cell proliferation and differentiation	4.05	4.52	0.90	1.57E-15
Amino acid metabolism	0.77	0.98	0.79	1.72E-13
Apoptosis	2.21	2.53	0.87	2.06E-13
Cell cycle	3.96	4.38	0.90	3.65E-13
Muscle contraction	1.35	1.58	0.85	1.49E-10
Other metabolism	2.12	2.38	0.89	1.94E-09
Protein metabolism and modification	14.29	14.91	0.96	2.86E-09
Biological process unclassified	29.39	28.62	1.03	4.33E-09
Electron transport	1.17	1.01	1.16	8.21E-08
Blood circulation and gas exchange	0.18	0.25	0.72	1.50E-05
Oncogenesis	2.37	2.57	0.92	5.57E-05
Intracellular protein traffic	5.54	5.83	0.95	6.90E-05
Protein targeting and localization	1.39	1.27	1.09	5.85E-04

(C) Non-polymorphic L1s vs random simulation: molecular functions

molecular function	**intronic non-polymorphic L1, %**	**random simulation, %**	**fold-change**	**p-value**
Receptor	12.06	10.71	1.13	0.00E+00
Ion channel	4.83	3.96	1.22	0.00E+00
Transcription factor	6.23	8.35	0.75	5.58E-37
Nucleic acid binding	7.91	10.1	0.78	7.20E-33
Isomerase	0.9	0.67	1.34	2.63E-04
Molecular function unclassified	27.31	26.35	1.04	6.76E-03

(D) Reference L1s vs random simulation: molecular functions

molecular function	**intronic reference L1, %**	**random simulation, %**	**fold-change**	**p-value**
Receptor	12.28	10.71	1.15	0.00E+00
Nucleic acid binding	8	10.1	0.79	2.03E-151
Transcription factor	7.01	8.35	0.84	1.01E-72
Isomerase	0.83	0.67	1.24	2.83E-09
Ion channel	4.9	3.96	1.24	3.34E-09
Cell adhesion molecule	4.26	3.94	1.08	2.75E-08
Chaperone	0.37	0.45	0.82	2.54E-04
Molecular function unclassified	26.86	26.35	1.02	2.87E-04
Protease	2.87	2.69	1.07	7.80E-04
Phosphatase	1.93	2.08	0.93	8.42E-04
Transfer/carrier protein	1.4	1.52	0.92	5.10E-03

(E) Non-polymorphic L1s vs reference L1s: biological processes

biological process	**intronic non-polymorphic L1, %**	**reference L1, %**	**fold-change**	**p-value**
Protein metabolism and modification	15.05	14.29	1.05	8.20E-03

(F) Non-polymorphic L1s vs reference L1s: molecular functions

molecular function	**intronic non-polymorphic L1, %**	**reference L1, %**	**fold-change**	**p-value**
Transcription factor	6.23	7.01	0.89	9.03E-06

Supplementary References.

Chen, J., A. Rattner, and J. Nathans. 2006. Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements. *Hum Mol Genet* **15:** 2146-2156.

DeBerardinis, R.J., J.L. Goodier, E.M. Ostertag, and H.H. Kazazian, Jr. 1998. Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. *Nat Genet* **20:** 288-290.

Goodier, J.L., E.M. Ostertag, K. Du, and H.H. Kazazian, Jr. 2001. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res* **11:** 1677-1685.

Hinrichs, A.S., D. Karolchik, R. Baertsch, G.P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T.S. Furey, R.A. Harte, F. Hsu et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34:** D590-598.

Li, J., M. Kannan, and D.E. Symer. 2008. Diverse fusion transcripts are initiated by a novel antisense promoter in mouse L1 retrotransposons. *Submitted for publication.*

Mi, H., B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M.J. Campbell et al. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33:** D284-288.

Saxton, J.A. and S.L. Martin. 1998. Recombination between subtypes creates a mosaic lineage of LINE-1 that is expressed and actively retrotransposing in the mouse genome. *J Mol Biol* **280:** 611-622.

Smit, A.F.A., R. Hubley, and P. Green. 2007. RepeatMasker.

Stephens, R.M., K. Akagi, J.R. Collins, B. Neelam, D. McCullough, N. Volfovsky, and D.E. Symer. 2008. PolyBrowse: An interface to access, query and display mouse genomic variation. *Submitted for publication*.

Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Szak, S.T., O.K. Pickeral, W. Makalowski, M.S. Boguski, D. Landsman, and J.D. Boeke. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052.

Wade, C.M. and M.J. Daly. 2005. Genetic variation in laboratory mice. *Nat Genet* **37:** 1175-1180.

Wade, C.M., E.J. Kulbokas, 3rd, A.W. Kirby, M.C. Zody, J.C. Mullikin, E.S. Lander, K. Lindblad-Toh, and M.J. Daly. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature* **420:** 574-578.

Wiltshire, T., M.T. Pletcher, S. Batalov, S.W. Barnes, L.M. Tarantino, M.P. Cooke, H. Wu, K. Smylie, A. Santrosyan, N.G. Copeland et al. 2003. Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc Natl Acad Sci U S A* **100:** 3380-3385.

Wu, T.D. and C.K. Watanabe. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21:** 1859-1875.