nature genetics

The regulated retrotransposon transcriptome of mammalian cells

Geoffrey J Faulkner[1], Yasumasa Kimura[2], Carsten O Daub[2], Shivangi Wani[1], Charles Plessy[2], Katharine M Irvine[3], Kate Schroder[3], Nicole Cloonan[1], Anita L Steptoe[1], Timo Lassmann[2], Kazunori Waki[2], Nadine Hornig[4,5], Takahiro Arakawa[2], Hazuki Takahashi[2], Jun Kawai[2], Alistair R R Forrest[2,6], Harukazu Suzuki[2], Yoshihide Hayashizaki[2], David A Hume[7], Valerio Orlando[4,5], Sean M Grimmond[1] & Piero Carninci[2]

Although repetitive elements pervade mammalian genomes, their overall contribution to transcriptional activity is poorly defined. Here, as part of the FANTOM4 project, we report that 6–30% of cap-selected mouse and human RNA transcripts initiate within repetitive elements. Analysis of approximately 250,000 retrotransposon-derived transcription start sites shows that the associated transcripts are generally tissue specific, coincide with gene-dense regions and form pronounced clusters when aligned to full-length retrotransposon sequences. Retrotransposons located immediately 5′ of protein-coding loci frequently function as alternative promoters and/or express noncoding RNAs. More than a quarter of RefSeqs possess a retrotransposon in their 3′ UTR, with strong evidence for the reduced expression of these transcripts relative to retrotransposon-free transcripts. Finally, a genome-wide screen identifies 23,000 candidate regulatory regions derived from retrotransposons, in addition to more than 2,000 examples of bidirectional transcription. We conclude that retrotransposon transcription has a key influence upon the transcriptional output of the mammalian genome.

Repetitive elements comprise 30–50% of mammalian genomes[1,2]. The vast majority of this sequence content originates from retrotransposition of SINE (short interspersed nuclear element), LINE (long interspersed nuclear element) and LTR (long terminal repeat) transposable element superfamilies, as well as direct transposition of genomic DNA. Retrotransposons facilitate genome evolution, support genome structure and provide alternative promoters, exons, terminators and splice junctions to protein-coding loci[3–8]. Retrotransposon insertions can disrupt gene expression[9,10] and cause numerous diseases[11], hence the notion that their expression and mobility are generally under strict control, including by endogenous RNA interference (RNAi) in human cells[12].

Retrotransposons were first characterized in maize as "controlling elements" of neighboring genes[13]. Since then, anecdotal findings have hinted that this control is due to noncoding RNA (ncRNA) transcription from retrotransposon promoters[14]. Noncoding RNAs have been shown to contribute to epigenetic regulation in eukaryotes through mechanisms such as RNAi[15,16], transcriptional interference[17,18] and antisilencing[19]. A recent study also noted the widespread production of short double-stranded RNAs (dsRNAs) from retrotransposons[20]. Thus, the insertion and transcription of retrotransposons proximal to protein-coding loci may create new 'transcriptional landscapes' throughout evolution[21].

Mutations and truncations have rendered most mammalian retrotransposons transpositionally incompetent. However, this does not preclude transcription initiation from promoters present within immobile elements[22]. For instance, active copies of LINE-1 (L1) contain a canonical 5′ promoter necessary for full-length transcription[23], as well as a downstream antisense promoter (ASP)[8]. The ASPs of transpositionally incompetent L1s have been shown to act as alternative promoters for more than 40 human protein-coding genes[8,24,25]. A global analysis of retrotransposon transcription would therefore likely expose additional promoters in immobile retrotransposons, such as the L1 ASP.

Unfortunately, genome-wide studies of retrotransposon transcription using array technologies have in the past been hampered by cross-hybridization[26]. Sequence tag technologies, by contrast, can detect single-base-pair differences between retrotransposon copies and thereby enable their discrimination. Cap Analysis

[1]Expression Genomics Laboratory, Institute for Molecular Bioscience, University of Queensland, Queensland, Australia. [2]Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa, Japan. [3]Cooperative Research Centre for Chronic Inflammatory Diseases and Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. [4]Dulbecco Telethon Institute, IRCCS Fondazione Santa Lucia at EBRI, Rome, Italy. [5]Dulbecco Telethon Institute, IGB CNR, Naples, Italy. [6]The Eskitis Institute for Cell and Molecular Therapies, Griffith University, Australia. [7]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Scotland, UK. Correspondence should be addressed to V.O. (vorlando@dti.telethon.it), S.M.G. (s.grimmond@imb.uq.edu.au) or P.C. (carninci@riken.jp).

Received 8 October 2008; accepted 5 March 2009; published online 19 April 2009; doi:10.1038/ng.368

Table 1 Description of CAGE dataset and mapping results

Species	Libraries	Clustered tissues	Tags	Mapped (%)	
Human	80	12	65,090,084	42,771,372	(65.7%)
Mouse	96	13	18,547,165	11,981,567	(64.6%)

Gene Expression (CAGE) is a tagging technology that functions through the cleavage of 20- to 21-nt tags from the extreme 5′ end of full-length cDNAs followed by high-throughput sequencing[27]. When mapped to a reference genome, CAGE tags survey transcription start site (TSS) activity on a massive scale[28]. A previous application of this technology produced an atlas of promoter usage in mouse and human and showed that a large proportion of the identified promoters lacked a precise TSS, but instead initiated transcription from multiple pyrimidine–purine dinucleotides within a 30-to 100-bp window[29].

The results of the current study, based on CAGE data obtained from the third and fourth stages of the Functional Annotation of Mouse (FANTOM) project, are essentially divided into two main themes. In the first, we observe and characterize the global transcription of repetitive elements throughout the mouse and human genomes across a comprehensive panel of cell types and tissues. We show that the fraction of retrotransposon-initiated RNA transcription varies greatly between cell and tissue types and that the expression of different families of retrotransposons is correlated with cell and tissue type. We also identify nearly 250,000 previously unknown retrotransposon-driven TSSs in mouse and human and analyze their consensus initiator sequence, tissue specificity and positions with respect to full-length retrotransposable elements.

In the second theme, we examine the functional consequences of retrotransposon transcription. These include the widespread potential of retrotransposons to provide alternative promoters to protein-coding genes, the putative *cis* regulation of protein-coding genes by nearby retrotransposon ncRNAs and the correlation between 3′-UTR retrotransposon insertions and reduced full-length transcription of protein-coding genes. Our findings highlight the global impact of retrotransposon transcription on the evolution and functional output of the mammalian transcriptome.

RESULTS

Repetitive elements are widely expressed

We first mapped 65 million human CAGE tags and 18.5 million mouse CAGE tags obtained from 176 CAGE libraries to their corresponding genomes (**Table 1** and **Supplementary Table 1** online). More than 80% of these tags were produced for the current study, with the remainder generated by FANTOM3. Once mapped, overlapping tags were merged into tag clusters and libraries were grouped into general tissue categories (**Supplementary Data 1** online). We then assigned tag clusters to repetitive elements defined by RepeatMasker (see URLs section in Methods and **Supplementary Note** online) and tested for corroborating mRNA and EST support (**Table 2**). In this analysis, a tag cluster containing two or more CAGE tags was considered to represent a single reliable TSS, on the basis of previous calculations that CAGE tags overlapped on the genome to form tag clusters far more frequently than random expectation, thus demonstrating that the vast majority of tag clusters represented true-positive TSSs[29].

We found that 44,264 and 275,185 TSSs occurred within repetitive elements in mouse and human, respectively, representing 18.1% and 31.4% of the total TSSs detected in each species. Despite being abundant, retrotransposon TSSs were found to be less expressed on average than nonrepeat TSSs. Only 2.8% and 5.2% of TSSs with >100 CAGE tags were attributed to retrotransposons in human and mouse, an observation perhaps explained by the strong nuclear subcellular localization of the associated transcripts (**Supplementary Note**). Finally, retrotransposon TSSs were corroborated by independent mRNA and EST support at one-sixth the rate observed for nonrepeat TSSs (**Table 2**), an unsurprising result considering that the technologies used to generate these data would deplete repeat-containing sequences[30].

Our analysis dealt only with tags that could be mapped unequivocally to their reference genome or with very high confidence via a strategy previously derived for multimap tags[31]. It should be noted that the proportion of multimap CAGE tags resolved by the latter method and mapped to repetitive elements was 28%, compared with

Table 2 CAGE-based discovery of repetitive element–associated transcription start sites and supporting evidence

		Mouse			Human		
	Tags in cluster	Tag clusters	mRNA support (%)	EST support (%)	Tag clusters	mRNA support (%)	EST support (%)
Retrotransposon	1	107,110	0.3	0.3	590,318	0.1	0.2
	2–10	24,231	0.8	1.1	198,925	0.2	0.6
	11–100	2,862	2.9	5.1	13,738	1.2	3.5
	>100	340	14.5	12.3	1,055	11.0	18.6
Satellite	1	483	0.8	0.0	12,670	0.1	0.4
	2–10	186	1.1	0.5	7,979	0.4	0.7
	11–100	32	4.5	4.5	1,770	0.8	1.9
	>100	5	100.0	50.0	119	11.3	20.8
Simple	1	19,560	2.6	5.6	39,967	1.9	5.4
	2–10	10,287	7.1	13.7	34,557	4.5	11.8
	11–100	3,637	31.4	40.5	11,257	16.0	29.9
	>100	2,684	81.2	82.5	5,785	67.9	76.0
Nonrepeat	1	444,873	2.5	10.2	1,112,611	1.0	6.0
	2–10	163,196	6.4	20.4	522,785	2.3	12.0
	11–100	27,955	26.5	42.2	65,475	10.8	29.4
	>100	8,947	82.0	83.7	13,526	63.5	72.9

mRNA support and EST support equates to having the 5′ end of a GenBank mRNA or EST overlapping the boundaries of a given tag cluster.

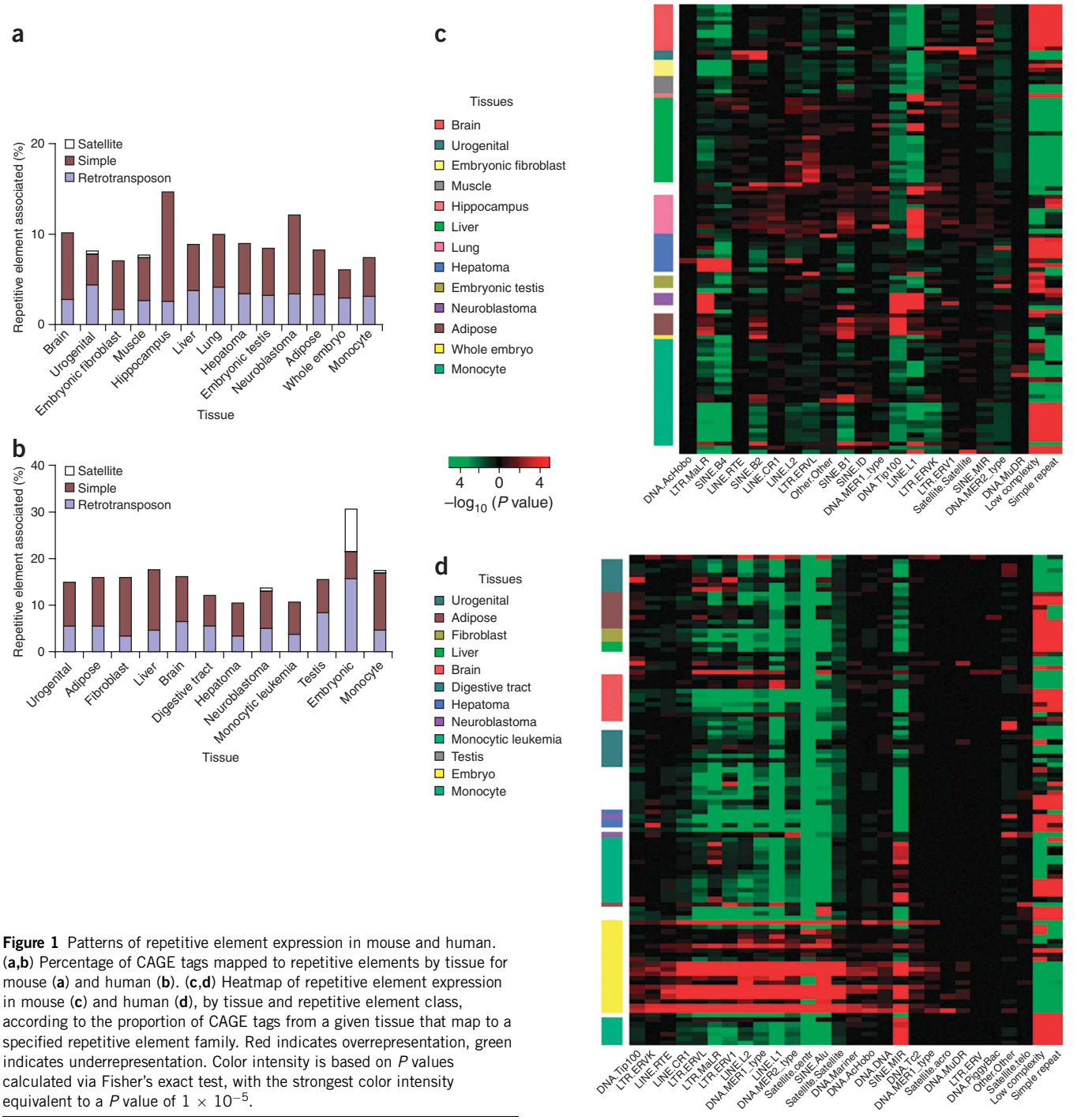

Figure 1 Patterns of repetitive element expression in mouse and human. (**a**,**b**) Percentage of CAGE tags mapped to repetitive elements by tissue for mouse (**a**) and human (**b**). (**c**,**d**) Heatmap of repetitive element expression in mouse (**c**) and human (**d**), by tissue and repetitive element class, according to the proportion of CAGE tags from a given tissue that map to a specified repetitive element family. Red indicates overrepresentation, green indicates underrepresentation. Color intensity is based on *P* values calculated via Fisher's exact test, with the strongest color intensity equivalent to a *P* value of 1×10^{-5}.

15% for other regions. Removal of multimap CAGE tags from the overall set reduced the number of retrotransposon TSSs by only ~20% (**Supplementary Table 2** online).

Tissue-specific patterns of repetitive element expression

The expression of repetitive elements varied considerably between tissues and activation states (**Fig. 1**). In one of the clearest patterns (**Fig. 1b**), 30% of all CAGE tags derived from human embryonic tissues were associated with repetitive elements (16% retrotransposon, 10% satellite, 5% simple repeat). Consequently, human embryonic tag clusters were strongly overrepresented ($P < 1 \times 10^{-5}$, Fisher's exact test) in numerous retrotransposon families and particularly in LINE subfamilies (**Fig. 1d**; for a complete list, see **Supplementary Data 1**).

Otherwise, the specificity of retrotransposon expression (**Fig. 1c**,**d**) was often poorly conserved between species and in some cases between similar samples represented by a common general tissue category. This suggested highly specific spatiotemporal expression of retrotransposons but likely also reflected differences between mouse and human copies of a retrotransposon family. Simple repetitive elements were strongly overrepresented in half of tissues, and strongly underrepresented in the remaining half. These elements, associated with CpG-rich promoters, also showed the strongest conservation of

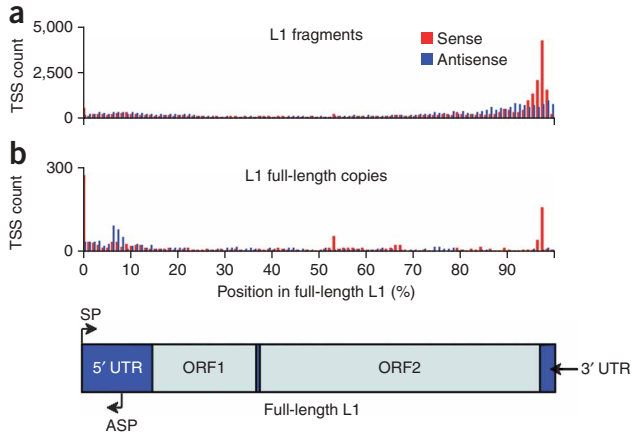

Figure 2 Transcription start site positions in human L1 fragments and full-length copies aligned to a consensus L1 sequence. (**a**) Human L1 fragments. (**b**) Full-length copies. Note the occurrence and relative abundance of 5′ and 3′ sense direction peaks in **a** and **b** and the occurrence of an antisense peak from the 6th to 8th percentile in **b**. The 5′ peak corresponds to the canonical L1 promoter[23], and the antisense peak corresponds to a previously characterized antisense promoter[8]. The 3′ promoter has not previously been observed. These distributions were calculated using RepeatMasker annotations that record the position of L1 fragments relative to a full-length element. Also note that as most L1 copies on the genome are 5′ truncated the vast majority of TSSs in each distribution, but in particular in **a**, occur in L1 fragments where the canonical 5′ promoter is not present.

expression between species. Overall, approximately 35% of all retrotransposon-associated TSSs showed spatially or temporally restricted expression, in contrast to the 17% observed for other TSSs (**Supplementary Fig. 1** online).

As a detailed example, expression of the VL30 subfamily of LTR retrotransposons, which is represented 100–200 times in the mouse genome and is divided into four subgroups[32], was clearly tissue restricted and almost completely excluded from brain, hypothalamus, neuroblastoma and embryonic tissues (**Supplementary Fig. 2** online). When we examined the activity of individual VL30 TSSs, we confirmed that distinct VL30 copies have acquired tissue-specific activities. An analysis of the ten most highly expressed copies of VL30 (**Supplementary Table 3** online) indicated that four were tissue specific.

Retrotransposons mainly use sharp transcription initiation

Previous studies using CAGE revealed two general mammalian promoter architectures: sharp and broad[29]. Sharp distribution promoters are preferentially associated with tissue-specific expression from a single, well-defined initiation site, and a substantial minority are controlled by an upstream TATA box[33]. Broad distribution promoters use pyrimidine/purine (Py/Pu) dinucleotides as TSSs within a 30- to 100-bp window, are rarely tissue specific or associated with TATA-box activity[33] and provide the majority of the known transcription initiation sites in mammals. We found that the overall and simple repeat promoter sets were predominantly broad (**Supplementary Fig. 3** online) and used the Py/Pu consensus initiator (**Supplementary Fig. 4** online).

In contrast, most retrotransposon promoters were sharp (**Supplementary Fig. 3**). Furthermore, retrotransposon-derived transcription initiation involved a strong preference for a guanine at position +1 as part of a degenerate Py/Pu dinucleotide (**Supplementary Fig. 5** online; for results by family, see **Supplementary Figs. 6** and **7** online). We showed previously that Py/Pu dinucleotides, a simplification of the Inr element, are associated with abundant transcripts[29]. However, the strongest initiation dinucleotide (CG) was only associated with highly expressed, broad promoters and was rare among the mainly sharp retrotransposon promoters presented here. This finding was congruent with previous results that suggested that the CG dinucleotide was far less common in sharp promoters than broad promoters[29]. An alternative hypothesis is that CG dinucleotides in retrotransposon promoters are usually silenced by DNA methylation. Instead of a CG, retrotransposon promoters were found to use the weaker AG, GG and TG initiation dinucleotides.

Few retrotransposon promoters conformed to the classical description of a sharp promoter, as only ~5% presented a TATA box 28–34 bp upstream of the main TSS, compared to ~20% for nonrepeat promoters[33]. Transcription-factor binding-site analyses of the regions surrounding retrotransposon promoters did not reveal any significantly overrepresented motifs within 200 bp to replace the TATA box.

Novel promoters dominate retrotransposon transcription

To address whether the observed retrotransposon promoters were known, we translated the positions of TSSs occurring in retrotransposon fragments to the equivalent positions in ancestral full-length elements (**Fig. 2** and **Supplementary Figs. 8** and **9** online). Overall, these results indicated that although the canonical 5′ promoters of elements such as human L1 and mouse SINE B2 were active in retrotransposon fragments, the vast majority of retrotransposon transcription initiated in previously unidentified sense and antisense promoters (**Supplementary Table 4** online). In some retrotransposon families (human L1, L2, Alu and MIR, mouse L1, B1, B2 and B4), the TSSs formed strong peaks that may indicate conserved promoters. The peaks were not caused by the relative frequency of retrotransposon fragments on the genome (**Supplementary Figs. 10** and **11** online).

We further expanded upon these observations by undertaking an in-depth analysis of promoters in fragments, full-length copies and active copies of the most highly expressed human retrotransposon, L1 (**Supplementary Table 4**). Although expression of the canonical 5′ L1 promoter was observed in L1 fragments, the bulk of L1 transcription initiation occurred at a distinct 3′ promoter (**Fig. 2a**) also present in mouse L1 (**Supplementary Fig. 9**). Even though most L1 copies on the human genome were found to be 5′ truncated (**Supplementary Fig. 10**), this did not account for the strong 3′ peak observed at the 3 end (**Fig. 2a**). This 3′ promoter was expressed in many tissues overall but was tissue specific in 42% of individual cases (using TSSs with >30 tags; **Supplementary Fig. 12a** online).

Next, we considered only TSSs occurring in full-length L1 copies that were >98% the consensus L1 length (**Fig. 2b**). In this case, the canonical 5′ promoter and ASP were far more apparent than in L1 fragments and were expressed in a tissue-restricted fashion (**Supplementary Fig. 12b**). Finally, we mapped the entire human CAGE set to an active L1 sequence, regardless of multimapping on the genome, and found a dominant canonical 5′ promoter expressed primarily in developmental and cancerous tissues (**Supplementary Fig. 13** online).

Retrotransposons generate alternative mRNAs and ncRNAs

Pervasive, tissue-specific retrotransposon transcription is likely to have functional consequences on the protein-coding transcriptome. To explore this possibility, we first compared the genomic coordinates of all RefSeq transcripts against the RepeatMasker coordinates of expressed repetitive elements (**Fig. 3a**,**b**). As expected[1], expressed retrotransposons were underrepresented in exonic sequences, although a number of families were actually overrepresented in

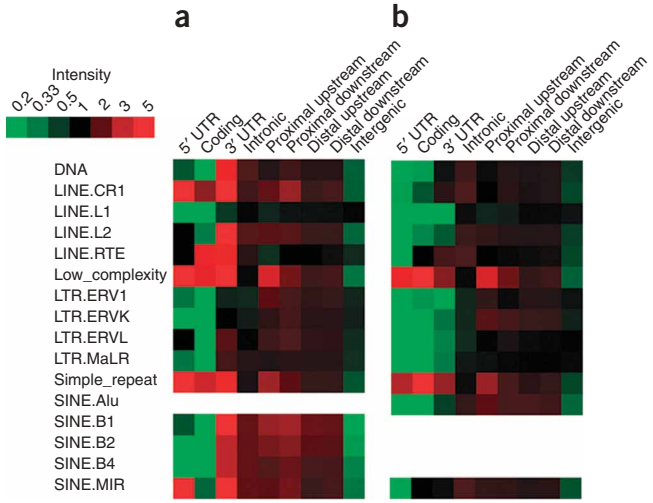

Figure 3 Abundance of expressed repetitive elements proximal to known transcripts. (**a**) Mouse. (**b**) Human. Heatmap intensities are based on the ratio of observed repetitive element frequencies divided by expected repetitive element frequencies. Maximum intensities correspond to fivefold changes. Cells with a *P* value >0.001, including those representing a fold change of 1.0, are automatically colored in black.

3′ UTRs. In both mouse and human, SINEs were very abundant in intronic and flanking sequences, and LTRs were overrepresented in flanking regions and LINEs in introns. These trends were generally much stronger than what was observed for retrotransposons irrespective of expression (**Supplementary Fig. 14** online). Expressed retrotransposons other than L1 were generally overrepresented close to protein-coding loci and underrepresented away from protein-coding loci, suggesting that retrotransposons are intrinsic components of the transcriptional forest regions of the genome[28].

Protein-coding genes are known to utilize nearby intergenic and retrotransposon sequences as alternative promoters[34]. With this in mind, we identified 15,518 and 117,165 retrotransposon TSSs for mouse and human protein-coding loci, respectively, which were transcribed in the same direction as, but did not overlap, a RefSeq transcript within 100 kb (**Supplementary Data 2** online). We confirmed alternative promoter activity for 154 mouse and 579 human retrotransposon TSSs by identifying ESTs that initiated in a retrotransposon promoter and terminated in a downstream RefSeq exon (**Supplementary Data 3** online).

To examine the remaining candidate alternative promoters, we selected five retrotransposon promoters expressed during the differentiation of human monocytic leukemia (THP-1) cells with phorbol myristate acetate (PMA) (see **Supplementary Table 5** online for a summary of results and Methods for experimental design). Quantitative real-time PCR (qRT-PCR) corroborated the CAGE signals in all five targets and three were confirmed as alternative promoters based upon generation of an amplicon containing the retrotransposon exon and a downstream RefSeq exon. These included a SINE MIR in intron 10 of *CSF1R* (**Fig. 4**), a LINE L2 in intron 1 of *GSN* (**Supplementary Fig. 15a** online) and an ERV1 LTR 20 kb upstream of the canonical TSS of *SLC2A5* (**Supplementary Fig. 15b**). cDNA sequencing confirmed the presence of transcripts that spanned the alternative promoter and downstream gene in each example (**Supplementary Table 5**). CAGE signals from the fourth target and fifth targets, a SINE Alu intronic to *GCLC* (**Supplementary Fig. 15c**) and a LINE L2 upstream of *MYBL2* (**Supplementary Fig. 15d**), were confirmed by qRT-PCR but produced RNAs that were independent of the tested coding regions.

In a subsequent experiment, we undertook 5′ RACE coupled with 454 sequencing[35] on 24 target RefSeq transcripts, each with at least one upstream candidate retrotransposon alternative promoter expressed in human hepatocarcinoma (HEPG2) or THP-1 cells (Methods and **Supplementary Table 6** online). Fifteen RefSeq transcripts were found to have at least one 'deep-RACE'–defined upstream retrotransposon alternative promoter (**Supplementary Table 7** online). In eight of these examples, sequence reads that spanned an alternative promoter and corresponding RefSeq transcript were detected, further supporting their association. When combined with the 800 mouse and human RefSeq transcripts known to initiate within a retrotransposon (**Supplementary Data 4** online), the qRT-PCR, cDNA sequencing, deep-RACE, EST and CAGE data presented here provide exceptionally strong evidence that retrotransposons frequently promote transcription of protein-coding genes.

Putative retrotransposon regulation of nearby genes

Noncoding transcription from intergenic regions has been shown to regulate nearby protein-coding genes[14,17,18,36]. We correlated the expression of numerous retrotransposon TSSs with the expression of the nearest upstream or downstream RefSeq transcript and found that

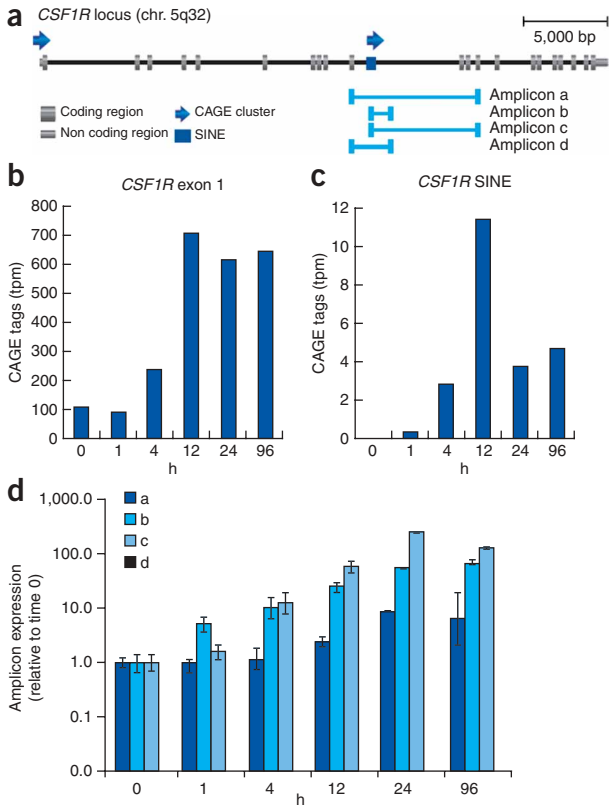

Figure 4 Validation of a SINE-derived alternative promoter in human *CSF1R*. (**a**) Schematic of *CSF1R* locus, SINE MIR and qRT-PCR amplicons. *CSF1R* exons and introns are not to scale. (**b**,**c**) CAGE expression profile of *CSF1R* (**b**) and CAGE expression profile of SINE MIR (**c**) in *CSF1R* in THP-1 cells over a timecourse of PMA treatment. (**d**) qRT-PCR quantification of amplicons a, b, c and d in PMA (phorbol myristate acetate) treated THP-1 cells (normalized to 0 hr). Error bars, ± s.d.; three technical replicates. The SINE MIR was located at chr5:149426871..149427084.

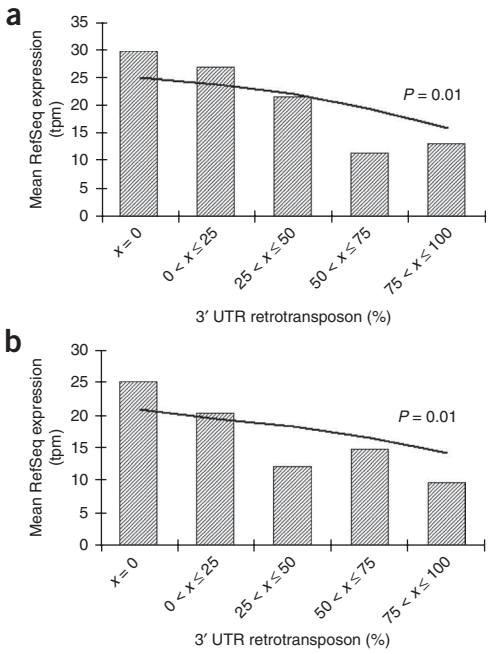

Figure 5 Inverse correlation between mean RefSeq CAGE expression and 3′ UTR retrotransposon percentage (*x*). (**a**) Human. (**b**) Mouse. Significance of underexpression is demarked by a line corresponding to $P = 0.01$ (permutation test).

3,445 mouse and 19,633 human tag clusters were correlated at >0.5 (based on both Spearman rank and Pearson correlations) to the nearest RefSeq transcript (within 100 kb). On the other hand, fewer than 20 were correlated at less than –0.5. This analysis included 68.1% sense–sense pairs overall and 78.7% positively correlated sense–sense pairs, suggesting an enrichment for positive correlations between sense–sense pairs. Examples of strong positive correlations included an L1 immediately downstream of *AZU1* (**Supplementary Fig. 15e**, also correlated with seven other protein-coding loci nearby), a MIR SINE downstream of *ORC6L* (**Supplementary Fig. 15f**) and a MIR SINE antisense to *SLC12A8* (**Supplementary Fig. 15g**). These results suggest that retrotransposon transcription infrequently produces transcriptional interference[18].

Extensive bidirectional transcription

Antisense transcription is a pervasive[37] and well-characterized mechanism by which ncRNAs are known modulate the epigenome in *cis*. Retrotransposons have also recently been shown to produce short dsRNAs from bidirectional transcription[20]. With this in mind, we identified sense–antisense CAGE tag cluster pairs separated by less than 300 bp where at least one of the transcripts initiated within a retrotransposon (illustrated in **Supplementary Fig. 16** online). Strongly correlated pairs (>0.5) were then identified as coexpressed: 2,081 human and 50 mouse pairs fit these criteria (**Supplementary Data 5** online). Of 1,797 human pairs expressed in THP-1 cells, 52 produced 21- to 25-nt short RNAs prepared from equivalent samples (**Supplementary Data 6** and **Supplementary Methods** online).

Bidirectional retrotransposon transcription can also delineate chromatin boundaries. As an example, a SINE B2 acts as a boundary element and maintains euchromatin around the mouse growth hormone locus via bidirectional transcription[14]. In this case, the action of transcription is the most important factor in establishing the chromatin boundary, rather than the structure of the transcripts involved. We carried out a genome-wide screen for similar elements by filtering the aforementioned sense–antisense pairs for those correlated at >0.5 with the nearest upstream or downstream RefSeq (within 100 kb). In this manner, we identified 333 human and 3 mouse putative boundary elements (see **Supplementary Data 5**).

3′ UTR retrotransposon insertions reduce mRNA expression

The 3′ UTRs of many protein-coding genes are known to contain active promoters[29] and, as noted earlier, we found that expressed retrotransposons were overrepresented in 3′ UTRs (**Fig. 3**). Subsequently, we investigated whether retrotransposons contributed disproportionately to 3′ UTR promoter activity and found no evidence that 3′ UTR promoters were associated specifically with retrotransposons. However, we noted that 27.7% of mouse and 28.5% of human RefSeqs contained at least one retrotransposon in their 3′ UTR (**Supplementary Data 4** and **Supplementary Table 8** online). When we plotted CAGE-quantified RefSeq expression as a function of 3′ UTR retrotransposon percentage (a statistic independent of 3′ UTR length), we discovered that RefSeq expression decreased as the percentage of 3′ UTR sequence contributed by retrotransposons increased (**Fig. 5**). This trend was found for most retrotransposon families (**Supplementary Figs. 17** and **18** online).

In order to validate this observation, we analyzed publicly available massively parallel signature sequencing (MPSS) of mouse tissues and observed the same global trend for transcripts known to contain 3′ UTR retrotransposons (**Supplementary Fig. 19** online). Mouse and human RefSeq transcripts with 3′ UTRs composed of >75% retrotransposon sequence presented mean CAGE tags-per-million (tpm) values 60% lower than RefSeqs lacking retrotransposons in their 3′ UTRs, a significant underexpression ($P < 0.01$, permutation test).

DISCUSSION

This study presents the most comprehensive survey of the repetitive element transcriptome achieved to date. We propose a summary model of a retrotransposon promoter: sharp in distribution, tissue specific, using a subclass of the mammalian Py/Pu consensus initiator dinucleotide, located in a newly identified position of an immobilized retrotransposon fragment and driving transcription of RNA predominantly localized in the nucleus. The last point suggests that these transcripts often perform functions distinct from those of transpositionally competent retrotransposon transcripts, which are known to be primarily exported to the cytoplasm[6]. Given the abundance of transcribed retrotransposons proximal to protein-coding genes, the common provision of alternative promoters to protein-coding genes by retrotransposons, the multiple regulatory mechanisms associated with ncRNAs and the finding that many retrotransposon transcripts are retained in the nucleus, we suggest that retrotransposons are multifaceted regulators of the functional output of the mammalian transcriptome.

The central principle underlying this work is that CAGE can accurately detect TSS activity in individual repetitive element promoters. More than two-thirds of CAGE tags successfully mapped to the genome, through an exact string matching algorithm (see Methods), aligned either perfectly or with a single G-addition error (**Supplementary Table 1**). Such high-quality sequence tag data enables single-nucleotide resolution and therefore the differentiation of TSSs that diverge by at least one nucleotide. Furthermore, we have previously demonstrated the capacity of CAGE to detect true 5′ transcript ends[28,29]. The extensive retrotransposon transcription observed in this study was also confirmed by experimental validation, frequent

coexpression of retrotransposon tag clusters and proximal protein-coding genes and by consistent repetitive element expression for similar tissues, such as the various human embryonic CAGE libraries. Every CAGE signal tested by qRT-PCR was validated, confirming the transcription of RNA from a specific retrotransposon TSS.

It is important to note that the retrotransposon TSSs observed here were as likely to be derived from truncated and subsequently capped RNAs[38] as those TSSs associated with protein-coding genes. The latter have previously been demonstrated by CAGE and numerous other techniques to predominantly correspond to the 5′ end of known genes, with a minority mapping to internal exons[29]. Although we cannot exclude that some of these retrotransposon TSSs are produced by novel types of capping[38], the promoters found in immobile retrotransposons conformed to a highly specific architecture involving a single dominant TSS and near universal use of a subtype of the consensus mammalian initiator dinucleotide.

Despite being accurately detected, retrotransposon transcripts appeared to be less expressed on average than protein-coding mRNAs. Retrotransposon expression may have been underestimated because their transcripts were preferentially localized in the nucleus, and the CAGE libraries presented here were almost entirely prepared from whole cells and tissue lysates, which are strongly biased toward cytoplasmic RNA (**Supplementary Note**). Second, the proportion of retrotransposon CAGE tags that mapped to multiple locations and were excluded from the analysis was twofold higher than the proportion observed for protein-coding mRNAs. Third, retrotransposon promoters were far more tissue-restricted in their expression than other promoters, which further hampered their discovery. Even if these arguments are disregarded and we conclude that most active retrotransposon promoters are poorly expressed, it could still be pointed out that transcript abundance is not necessarily important for function, as noted for ncRNA epigenetic regulators[36,39].

Rather than advocating universal utility for retrotransposons, we instead suggest that they contain active promoters and that at least some of these are functional. The high frequency of retrotransposons in mammalian genomes ensures the concurrence of many thousands of well-expressed retrotransposon promoters proximal to protein-coding genes, with widespread effects upon the regulation and evolution of those genes.

One of the most intriguing results of this study was the poor expression of canonical 5′ promoters in the immobilized forms of active and autonomous[6] retrotransposons, such as L1 and mouse SINE B2. Our results indicate that previously unknown promoters elsewhere in these elements supplant the activity of their canonical 5′ promoters post-transposition, most likely as a result of 5′ truncations. The newly identified promoters are individually tissue-specific but are expressed in a wide range of tissues if considered in total.

As an example, the canonical 5′ promoter of human L1 was highly expressed in active and full-length copies, was restricted in expression to developmental and cancer cells (confirming the literature on L1 activity[40–42]) and was far less expressed in L1 fragments on the genome than a strong, ubiquitously expressed 3′ promoter. Such disparate tissue expression patterns between the canonical 5′ and novel 3′ L1 promoters indicate that the 3′ L1 transcripts are likely to derive from true TSSs in L1 elements. Furthermore, the 3′ promoter would escape the endogenous RNAi that inhibits full-length L1 transcripts *in vivo*[12]. These findings complement previous discoveries of novel retrotransposon promoters[8,43] and suggest that the canonical and novel promoters of autonomous retrotransposons have distinct functional roles. The ultimate function, perhaps after further processing, of transcripts associated with novel retrotransposon promoters deserves future study.

The presence of retrotransposon sequences in the 3′ UTR of >25% of protein-coding mRNAs was clearly associated with a reduction in their expression. This was especially intriguing as 3′ UTRs are the site of intense transcriptional activity and transcriptional regulation, such as the identification of 3′ UTR, capped RNA transcripts[29] and termini-associated short RNAs (TASRs) in approximately half of protein-coding genes[44]. TASR expression correlated with elevated full-length transcription, with these short RNAs postulated to maintain euchromatin around genes.

Retrotransposon insertions in 3′ UTRs could interfere with 3′ UTR transcripts or other 3′ UTR *cis* regulatory elements, introduce miRNA binding sites or promote RNA editing[45], all of which could repress transcription from the affected gene. Another possibility is that polyadenylation signals contained within retrotransposon sequences truncate the 3′ UTRs of full-length transcripts by providing an alternative terminator[9]. Yet another possibility is that the mRNAs containing 3′ UTR retrotransposons are degraded in *trans* by the widespread expression of other retrotransposons, owing to sequence complementarity. This hypothesis is corroborated by the identification of endogenous siRNAs derived by sense-antisense transcription of pseudogenes and retrotransposons[20,46].

Regardless of mechanism, the suppressive effects of 3′ UTR retrotransposon insertions affect more than one quarter of the protein-coding genes in human and mouse. These events are only moderately selected against in comparison to what is observed for retrotransposon insertions in coding exons and may provide a gradual mechanism of evolution by which retrotransposons alter the expression profile of the genome.

With the use of powerful, high-throughput methodologies, we were able to elucidate in depth the extent and character of repetitive element transcription in mammalian cells. We anticipate extensive biological analyses as a consequence of this work, such as the phenotypic effects of large-scale inhibition of retrotransposon expression, as well as an increasing inventory of functional retrotransposons proximal to protein-coding genes. The availability of CAGE, coupled with next-generation sequencing, will also enable surveys of retrotransposon transcription in other organisms, such as chicken and *Drosophila*, where retrotransposon content differs greatly from that of human and mouse. Ultimately, repetitive elements are a pervasive source of transcription and transcriptional regulation and therefore must be considered in future studies of the genome as a 'transcription machine'.

METHODS

Nomenclature. Throughout the text and figures we group and denote SINE, LINE, LTR and DNA superfamilies as retrotransposons for simplicity. SINEs, LINEs and LTRs provide the vast majority of the elements studied. The term 'simple repeat' refers to short repetitive elements defined by RepeatMasker as low-complexity or simple-repeat elements. Lastly, 'satellite' refers to the major tandem repeats found predominantly in centromeric and telomeric DNA.

Sequence tag mapping. Mouse and human CAGE libraries were generated via RISA[47] and multiplex pyrosequencing[48] for previous and concurrent works[28] and mapped to MM9 and Hg18[49], respectively, using Nexalign (unpublished data), a suffix array-based alignment method. Nexalign supports alignments with up to three mismatches or one insertion or deletion and is guaranteed to find all full-length matches of a query to a target database. Here, we applied Nexalign in a hierarchical fashion by aligning exactly matching tags first, then those tags that mapped with a G-addition error[29], then allowing up to one error and finally allowing a G-addition error as well as one other error. A CAGE tag mapping to a single repetitive element thus indicates that the particular subsequence of the repetitive element copy is unique on the genome.

The alignment providing the highest number of matched bases for any given tag was defined as the 'best alignment'. If that tag aligned equally well to more

than one genomic location it was defined as a 'multimapper'. Multimapping tags were then processed using a previously described method for their resolution[31], although these represented the minority of mapped tags. Finally, tags were flagged as a product of ribosomal RNA contamination in the CAGE libraries and assigned a count of 0 if a match as good as, or better than, the best genomic alignment was recorded against the mouse and human rDNA (rRNA precursor) sequences. Massively parallel signature sequencing tags were obtained from NCBI GEO and mapped to mouse RefSeq transcripts using Nexalign and a maximum of one error per tag.

CAGE library clustering. Tag clusters (TCs) were generated for all CAGE libraries using a previously published method[28], where an overlap of at least one base was necessary for clustering, and normalized to tags per million (tpm). Libraries containing fewer than 10,000 tags were omitted to reduce noise. Tag clusters that contained at least 40 tpm in one or more libraries were then used to cluster libraries via average linkage[50]. Libraries of a similar origin that clustered together were finally grouped into 12 and 13 cell types for human and mouse, respectively (**Supplementary Data 1**).

Heatmap preparation. The heatmaps presented in **Figure 1c,d** were created using Fisher's exact test. Specifically, the 2 × 2 contingency table for this test was comprised of: (1,1) library X tpm value for repeat class Y, (1,2) sum of tpm values for library X across repeats other than Y, (2,1) sum of tpm values for libraries other than X for repeat class Y and (2,2) sum of tpm values for libraries other than X and repeats other than Y. One-tailed Fisher's exact tests were used to calculate over (red) / under (green) representation and two-tailed Fisher's exact tests were used to calculate *P* values (color intensity).

The heatmaps presented in **Supplementary Figure 14** are based on observed values divided by expected values for the frequency of each repetitive element family relative to the nearest RefSeq transcripts. RefSeq exons were divided into 5′ UTR, protein-coding sequence and 3′ UTR. 'Distal Upstream' was defined as –100,000, –10,000 in relation to the 5′ end of a RefSeq transcript. 'Proximal Upstream' was defined as –10,000, –1 in relation to the 5′ end of a RefSeq transcript. 'Distal Downstream' and 'Proximal Downstream' were +10,000, +100,000 and +1, +10,000 of the 3′ end of a RefSeq transcript, respectively. 'Intergenic' refers to any repetitive element more than 100 kb from the nearest RefSeq transcript. Any given repetitive element could have relationships to multiple RefSeqs; for example, an element could be intronic to RefSeq A and in the 5′ UTR or RefSeq B. Significance values were determined by calculating the normal distribution of expected repetitive element frequencies by using five iterations of randomized RefSeq positions on the genome to provide a mean and s.d. for the normal distribution. During this process only the location of the RefSeq was randomized; the internal spacing between exons was not changed. Cells with a *P* value greater than 1×10^{-2} were set to black in each heatmap. **Figure 3** was prepared in the same way as **Supplementary Figure 14** except only repetitive elements containing at least one CAGE tag cluster of ≥2 tags were analyzed.

Promoter architecture. We classified a sharp promoter distribution as a CAGE tag cluster of at least 40 tags that also included a single position containing >80% of the total tags for that cluster. All other tag clusters with at least 40 tags were denoted as broad.

Pairwise correlations. Expression profiles were assigned to RefSeq transcripts by assigning all CAGE tags within –300,+100 of a RefSeq 5′ end to that RefSeq promoter. Those RefSeq promoters containing ≥2 CAGE tags were then compared to nearby retrotransposon CAGE tag clusters, which also were required to contain at least two CAGE tags. Both Spearman-ranked and Pearson correlation coefficients were used for robust pairwise correlations owing to sparse expression profiles (CAGE counts were often zero for most tissues in each pairwise comparison) and in order for a expression comparison to exceed a correlation threshold (for example, 0.5) both the Spearman-ranked and Pearson correlations had to exceed the threshold.

RNA isolation. A THP-1 cell line was subcloned and one clone (5) was selected for ability to differentiate relatively homogeneously in response to PMA. THP-1.5 was used for all subsequent experiments. THP-1.5 cells were cultured in RPMI, 10% FBS, penicillin/streptomycin, 10 mM HEPES, 1 mM sodium pyruvate, 50 μM 2-mercaptoethanol. THP-1.5 cells were treated with 30 ng/ml PMA (Sigma) over a time course of 96 h (2×10^7 THP-1 per T225 flask in 100 ml). Total RNA was treated with DNase and purified on Qiagen RNeasy columns according to the manufacturer's instructions.

Cytoplasmic RNA was extracted from 5×10^7 HEPG2 cells. After centrifugation, cells were resuspended in 2 ml of sterile PBS buffer. Added to this was 5 ml of pre-cooled Cell Lysis Solution (100 mM NaCl, 5 mM $MgCl_2$, 50 mM Tris-HCl (pH 7.5), 0.5% IGEPAL CA-630) premixed with 250 μL of Eppendorf RNase inhibitor. After incubation on ice for 3 min, nuclei were centrifuged for 1–3 min at 7,500 rpm at 4 °C. The nuclear pellet was washed with PBS and was subjected to standard RNA extraction with Trizol protocol. The supernatant was used to extract the cytoplasmic RNA. Cytosolic RNA was then precipitated by adding 15.5 ml of CTAB buffer (1% CTAB, 4 M urea, 50 mM Tris-HCl (pH 7.0), 1 mM EDTA (pH 8.0)) and 970 μl 5 M NaCl and centrifuged for 10 min at 7,500 rpm (9,500*g*). Cytoplasmic RNA was resuspended with 2 ml of 7 M guanidinium chloride, phenol-chloroform and chrororform extracted and finally precipitated with 1 volume of isopropanol. The cytosolic RNA pellet was washed with 70% before redissolving it in RNAse free water.

qRT-PCR. Quantitative RT-PCR was performed using an ABI prism 7900HT Fast Real-Time PCR system with SYBR-green PCR mastermix (Applied Biosystems), 2.5 ng of total RNA (reverse transcribed using oligo dT) per reaction and 0.5 nM forward and reverse primer concentrations. Cycling was done using the default program (50 °C for 2 min, 95 °C for 10 min, 40 cycles of 95 °C for 15 s, 60 °C for 1 min), followed by dissociation protocol (95 °C for 15 s, 60 °C for 15 s, 95 °C for 15 s). The assay included a nontemplate control and all samples were run in triplicate.

cDNA sequencing. PCR products were gel-purified using QIAEX II Gel Extraction Kit (Qiagen). Sequencing was carried out using 10 ng of this purified PCR product and the respective forward or reverse primer on the AB3730xl 96 platform.

PCR primer design. RefSeq transcripts were targeted with a forward primer in the first exon and a reverse primer in the second exon. If the second exon was too small, and the first exon was sufficiently large, the reverse primer was placed at the 3′ end of the first exon. Amplicons were designed to be between 150 bp and 250 bp. CAGE signals occurring in retrotransposons were targeted by an amplicon immediately downstream of the tag cluster. The forward primer was placed in the tag cluster (a unique genomic region) and the reverse primer was placed in the first unique genomic region more than 150 bp downstream of the tag cluster. Alternative promoters from retrotransposons were tested by combining forward primers within retrotransposon CAGE signals with reverse primers from RefSeq transcripts. All primer sequences are listed in the **Supplementary Note**.

Deep RACE. Using a novel protocol[35] 5′ RACE products were sequenced from 24 RefSeq transcripts (**Supplementary Table 6**) using a GS FLX 454 sequencer (Roche) to achieve deep RACE. Primers were designed using Primer3Plus. Barcodes were used to simultaneously sequence RACE products from HEPG2 nuclear, HEPG2 cytoplasmic, THP-1 0 h PMA and THP-1 96 h PMA RNA samples (preparation described above). We generated 109,177 reads with a median length of 47 nt after removal of barcode, adaptor and primer sequences. These reads were then mapped to Hg18 using BLAT[49].

URLs. Data (except short RNA data), http://fantom.gsc.riken.jp/4/download/Supplemental_Materials/Faulkner_et_al_2009/; short RNA data, https://fantom.gsc.riken.jp/4/download/Supplemental_Materials/Taft_et_al_2009/; Repeat Masker, http://www.repeatmasker.org/.

Accession codes. GenBank: mouse and human rDNA (rRNA precursor) sequences, BK000964 and U13369, respectively. NCBI GEO: massively parallel signature sequencing tags were obtained from GDS868.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

G.J.F. is supported by an Australian Postgraduate Award through the Australian government Department of Education, Training and Youth Affairs (DETYA).

A.R.R.F. is funded by a CJ Martin Fellowship from the Australian NHMRC (ID 428261). K.S. and K.M.I. are members of the CRC for Chronic Inflammatory Diseases. P.C. and Y.H. are supported by the National Project on Protein Structural and Functional Analysis from MEXT and the National Project on Genome Network Analysis and the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. D.A.H. acknowledges the funding of the ARC Special Research Centre for Functional and Applied Genomics and the NHMRC. S.M.G. holds a Senior Research Fellowship with the Australian NHMRC. N.C. is supported by a UQ postdoctoral research fellowship. V.O. is supported by Telethon Foundation (TCP00094), Associazione Italiana Ricerca sul Cancro (AIRC) and Fondazione Compagnia di San Paolo, and N.H. is supported by an EMBO long-term fellowship.

AUTHOR CONTRIBUTIONS

G.J.F. led the bioinformatic analysis and drafting of the manuscript. K.M.I., K.S., N.C., A.L.S., T.L., K.W., N.H., T.A., J.K., H.S., Y.H., S.M.G. and P.C. provided data and resources. Y.K., C.O.D. and A.R.R.F. provided bioinformatic analyses. G.J.F. and P.C. designed the experiments. S.W., C.P., A.L.S., H.T. and N.C. performed validation. D.A.H., V.O., S.M.G. and P.C. interpreted data and edited the manuscript. P.C. organized the project.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
2. Margulies, E.H. *et al.* Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc. Natl. Acad. Sci. USA* **102**, 3354–3359 (2005).
3. Babushok, D.V., Ostertag, E.M. & Kazazian, H.H. Jr. Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell. Mol. Life Sci.* **64**, 542–554 (2007).
4. Hasler, J., Samuelsson, T. & Strub, K. Useful 'junk': Alu RNAs in the human transcriptome. *Cell. Mol. Life Sci.* **64**, 1793–1800 (2007).
5. Jurka, J. Evolutionary impact of human Alu repetitive elements. *Curr. Opin. Genet. Dev.* **14**, 603–608 (2004).
6. Kazazian, H.H. Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
7. Peaston, A.E. *et al.* Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
8. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* **21**, 1973–1985 (2001).
9. Han, J.S., Szak, S.T. & Boeke, J.D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268–274 (2004).
10. Ustyugova, S.V., Lebedev, Y.B. & Sverdlov, E.D. Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts. *Genetica* **128**, 261–272 (2006).
11. Batzer, M.A. & Deininger, P.L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).
12. Yang, N. & Kazazian, H.H. Jr. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.* **13**, 763–771 (2006).
13. McClintock, B. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21**, 197–216 (1956).
14. Lunyak, V.V. *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**, 248–251 (2007).
15. Lei, E.P. & Corces, V.G. RNA interference machinery influences the nuclear organization of a chromatin insulator. *Nat. Genet.* **38**, 936–941 (2006).
16. Noma, K., Cam, H.P., Maraia, R.J. & Grewal, S.I. A role for TFIIIC transcription factor complex in genome organization. *Cell* **125**, 859–872 (2006).
17. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–670 (2007).
18. Mazo, A., Hodgson, J.W., Petruk, S., Sedkov, Y. & Brock, H.W. Transcriptional interference: an unexpected layer of complexity in gene regulation. *J. Cell Sci.* **120**, 2755–2761 (2007).
19. Schmitt, S. & Paro, R. RNA at the steering wheel. *Genome Biol.* **7**, 218 (2006).
20. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539–543 (2008).
21. Whitelaw, E. & Martin, D.I. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat. Genet.* **27**, 361–365 (2001).
22. Conley, A.B., Miller, W.J. & Jordan, I.K. Human *cis* natural antisense transcripts initiated by transposable elements. *Trends Genet.* **24**, 53–56 (2008).
23. Swergold, G.D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10**, 6718–6729 (1990).
24. Matlik, K., Redik, K. & Speek, M. L1 antisense promoter drives tissue-specific transcription of human genes. *J. Biomed. Biotechnol.* **2006**, 71753 (2006).
25. Nigumann, P., Redik, K., Matlik, K. & Speek, M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628–634 (2002).
26. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
27. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
28. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
29. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
30. Carninci, P. *et al.* Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630 (2000).
31. Faulkner, G.J. *et al.* A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**, 281–288 (2008).
32. Nilsson, M. & Bohm, S. Inducible and cell type-specific expression of VL30 U3 subgroups correlate with their enhancer design. *J. Virol.* **68**, 276–288 (1994).
33. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* **8**, 424–436 (2007).
34. Denoeud, F. *et al.* Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* **17**, 746–759 (2007).
35. Olivarius, S., Plessy, C. & Carninci, P. High-throughput verification of transcriptional starting sites by Deep-RACE. *Biotechniques* **46**, 130–132 (2009).
36. Schmitt, S., Prestel, M. & Paro, R. Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev.* **19**, 697–708 (2005).
37. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
38. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
39. Seidl, C.I., Stricker, S.H. & Barlow, D.P. The imprinted *Air* ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J.* **25**, 3565–3575 (2006).
40. Babushok, D.V. & Kazazian, H.H. Jr. Progress in understanding the biology of the human mutagen LINE-1. *Hum. Mutat.* **28**, 527–539 (2007).
41. Garcia-Perez, J.L. *et al.* LINE-1 retrotransposition in human embryonic stem cells. *Hum. Mol. Genet.* **16**, 1569–1577 (2007).
42. van den Hurk, J.A. *et al.* L1 retrotransposition can occur early in human embryonic development. *Hum. Mol. Genet.* **16**, 1587–1592 (2007).
43. Ferrigno, O. *et al.* Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* **28**, 77–81 (2001).
44. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
45. Chen, L.L., DeCerbo, J.N. & Carmichael, G.G. Alu element-mediated gene silencing. *EMBO J.* **27**, 1694–1705 (2008).
46. Tam, O.H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
47. Shibata, K. *et al.* RIKEN integrated sequence analysis (RISA) system—384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res.* **10**, 1757–1771 (2000).
48. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
49. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
50. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).