Genome-wide analysis of repetitive elements associated with gene regulation

Repetitive elements and gene regulation

Lu Zeng[1, 2, *], Stephen M. Pederson[2, *], Danfeng Cao[1], Zhipeng Qu[2], Zhiqiang Hu[1], David L. Adelson[2&], and Chaochun Wei[1&]

[1]School of Life Sciences and Biotechnology

Shanghai Jiao Tong University

Shanghai

P. R. China

[2]School of Biological Sciences

The University of Adelaide

Adelaide, SA

Australia

RUNNING TITLE: Repeats and Gene Regulation

[*]These authors contributed equally to this work

[&]To whom correspondence should be addressed.

[1] Tel: (86) (21) 34204348

Fax: (86)(21)34204348

Email: ccwei@sjtu.edu.cn

[2]Tel: +61 8 83137555

Fax: +61 8 83133262

Email: david.adelson@adelaide.edu.au

ABSTRACT

Nearly half of the human genome is made up of transposable elements (TEs) and evidence supports a possible role for TEs in gene regulation. Here, we have integrated publicly available genomic, epigenetic and transcriptomic data to investigate this potential function in a genome-wide manner. Results show that although most TE classes are primarily involved in reduced gene expression, Alu elements are associated with up regulated gene expression. This is consistent with our previously published work which showed that intronic Alu elements are capable of generating alternative splice variants in protein-coding genes, and further illustrates how Alu elements can alter protein function or gene expression level. Furthermore, non-coding regions were found to have a great density of TEs within regulatory sequences, most notably in repressors. Our exhaustive analysis of recent datasets has extended and updated our understanding of TEs in terms of their global impact on gene regulation, and indicates a significant association between repetitive elements and gene regulation.

Keywords: Transposable elements/Epigenetic/Gene expression/Gene ontology/Regulatory elements.

INTRODUCTION

Repetitive elements are similar or identical DNA sequences present in multiple copies throughout the genome. The majority of the repetitive sequences in the human genome are

43 derived from transposable elements (TEs) [1, 2] that can move within the genome,
44 potentially giving rise to mutations or altering genome size and structure. Typical eukaryotic
45 genomes contain millions of copies of transposable elements (TEs) and other repetitive
46 sequences. TEs fall into two major classes: those moving/replicating via a copy and paste
47 mechanism and an RNA intermediate (retrotransposons) and those moving via direct cut
48 and paste of their DNA sequences (DNA transposons). Retrotransposons can be subdivided
49 into two groups: Those with long terminal repeats (LTRs), and those without LTRs (non-
50 LTRs). Human LTR elements are related to endogenous retroviruses (HERVs), which along
51 with similar elements account for nearly 8% of the human genome [3]. Non-LTR
52 retrotransposons include two sub-types: autonomous long interspersed elements (LINEs)
53 and non-autonomous short interspersed elements (SINEs), which are dependent on
54 autonomous elements for their replication; both LINEs and SINEs are widespread in
55 eukaryotic genomes. LINE-1 (long interspersed element 1) and Alu elements are two TEs
56 that belong to non-LTR retrotransposons, which account for approximately one-quarter of
57 the human genome [1].

58 A number of existing studies have shown that TEs can influence host genes by providing
59 novel promoters, splice sites or post-transcriptional modification to re-wire different
60 developmental regulatory and transcriptional networks [4-6]. TEs tend to regulate gene
61 expression through several mechanisms [6-9]. For example, the expression levels of protein
62 coding genes containing repetitive elements are significantly associated with the number of
63 repetitive elements in those genes in rodents [7]. L1 family repeats show a stronger negative

64 correlation with expression levels than the gene length [10], and the presence of L1
65 sequences within genes can lower transcriptional activity [11]. Moreover, TEs have been
66 shown to influence gene expression through non-coding RNAs, resulting in the reduction or
67 silencing of gene expression [12]. For example, the expression of long intergenic non-
68 coding RNAs (lincRNAs) was strongly correlated with HERVH transcriptional regulatory
69 signals [13]. Past studies have found that TEs have contributed to nearly half of the active
70 regulatory elements of the human genome [14], by altering gene promoters and creating
71 alternative promoters and enhancers to regulate gene activity [15-17]. According to previous
72 research, 60% of TEs in both human and mouse were located in intronic regions and all TE
73 families in human and mouse can exonize, supporting the view that TEs may create new
74 genes and exons by promoting the formation of novel or alternative transcripts [18, 19]. The
75 association between repetitive elements and RNAs has also been investigated. For example,
76 Alu elements in lncRNAs can lead to *STAU1* mediated mRNA decay by duplexing with
77 complementary Alu elements in the 3'UTRs of mRNAs [20], and the insertion of TEs may
78 also drive the evolution of lincRNAs and alter their biological functions [13].
79
80 In this paper, TEs in the human genome were analyzed using genome-wide datasets
81 associated with gene regulation. These datasets enabled an assessment of the association of
82 TEs with chromatin states, as marked by histone modification within six human cell lines,
83 lincRNAs, Gene Ontology (GO) enrichment, as well as overall transcriptome profiles.
84 Whilst our analysis is limited to a general comparison of repeat families, as opposed to

specific repeat elements, we found clear associations between repeat families and gene regulation, both within regulatory regions and in the generation of splice variants.

RESULTS

The distribution of repetitive elements in the human genome

Initially, we compared the distribution of repetitive elements in the human genome. We found many repetitive elements overlapped with gene models from the human RefSeq gene datasets, and their distributions with respect to components of the gene model are shown in Fig 1. Most repetitive elements were found in non-coding intervals such as 5'UTR introns, CDS introns, 3'UTR introns and intergenic regions. In regards to clade-specific repeats, these were found more often in introns and intergenic regions than in 3'UTR exons or 5'UTR exons (Fig 1).

Fig 1. Distribution of repetitive elements overlapping with different human gene regions.

Gene regions are shown on the x-axis and the y-axis shows the percentages of the genomic regions containing repetitive elements. Human-specific repeats were those annotated with "Homo sapiens" or "primates" as their origin (see Table S6 for the list of human-specific repeat classes). The remaining repetitive regions were categorized as shared repeats.

Role of transposable elements in gene regulation by chromatin states.

Based on previous studies, TE-derived sequences can provide transcription factor binding

sites, promoters and enhancers, and insulators/silencers [5, 21, 22]. To look for enrichment of TEs within regulatory elements, we looked at the proportions of nucleotides with a TE in each of the six defined regulatory elements [23] as they appear in different components of the gene model. This represents the probability of a given nucleotide within an regulatory element (RE) being from a transposable element (TE), i.e. p(TE|RE) (see Methods for details) across the set of genic regions (Fig 2A). Confidence intervals for the pairwise differences in p(TE|RE) are shown in Fig 2B, and these reveal that for all regulatory elements, TEs are more sparsely distributed across regulatory elements within CDS exons than across regulatory elements in all other genic regions. Likewise, regulatory elements in the 3'UTR were more sparsely populated with TEs in comparison to those in other regions, with the sole exception of Active Promoters in the 5'UTR, Conversely, Intergenic Polycomb Repressed Regions were enriched for TEs in comparison to these elements in other components of gene models. Finally, Intergenic Insulators were also found to be enriched for TEs in comparison to Insulators in all other components of gene models, except for those in 5'UTRs.

Fig 2. Analysis of the co-occurrence of Transposable Elements and Regulatory Elements across multiple genomic regions A) Probability estimates are those of an individual base within each region being part of a Regulatory Element, i.e. p(RE), or a Transposable Element within a Regulatory Element, i.e. p(TE|RE). Error bars indicate ±1 Std. Error as calculated on the logit-transformed values. B) $1-\alpha$ Confidence Intervals for the difference between logit-transformed probabilities p(TE|RE), adjusted for multiple comparisons at the level $\alpha = 0.05/6$ within each RE (60). Intervals highlighted in red are

126 those do not contain zeros and are indicative of a significant difference between the two
127 values.

128 **Different classes of transposable elements and their associations with chromatin state.**

129 In order to systematically characterize the role of different repeat classes within the defined
130 regulatory elements, the distribution of regulatory elements within specific classes of TEs
131 were investigated, using the estimates of p(RE|TE) (See Methods). Out of all six TE classes
132 investigated (Alu, L1, L2, LTR, MIR and DNA), L1 elements were consistently found with
133 the lowest probability of a nucleotide also belonging to a regulatory element (Fig 3A). It
134 was also clear that regulatory elements had the highest probability of being exapted as Weak
135 Enhancer and Polycomb Repressed Regions compared to the other elements. Confidence
136 intervals were used to perform pair-wise comparisons on the probability of containing an
137 RE for each TE type. No difference was found between ancestral and recent L1 elements for
138 any RE (Fig 3B), and L1s were confirmed as containing a significantly lower proportion of
139 their content as an RE in comparison to all other TEs. The notable exceptions to this were
140 Polycomb Repressed Regions, where little difference was found in their rate of occurrence
141 between any TE types, beyond comparative enrichment in MIR elements compared to L1s.
142 MIR elements were more likely to contain a strong enhancer than all other elements, except
143 L2 and DNA elements. DNA elements were also more likely to contain an Insulator than
144 Alu elements, as well as the previously mentioned L1 elements.

Fig 3. Analysis of the occurrence of Regulatory Elements within specific classes of Transposable Elements. A) Probability estimates are for an individual base within each type of element belonging to each of the regulatory elements, i.e. p(RE|TE). Error bars indicate ±1 Std. Error as calculated on the logit-transformed values. B) $1-\alpha$ Confidence Intervals for the difference between logit-transformed probabilities p(RE|TE), adjusted for multiple comparisons at the level $\alpha = 0.05/6$ within each RE (60). Intervals highlighted in red are those do not contain the zero and are indicative of a significant difference between the two values

Are Regulatory Elements containing TEs abundantly present in long intergenic non-coding RNAs?

TEs are a source of endogenous small RNAs in animals and plants, and endogenous small RNAs are considered to be functionally significant in gene regulation [24]. Furthermore, it is well known that many Alu elements have inserted into long non-coding RNAs and mRNAs, which can cause mRNA decay via short imperfect base-pairing [25]. We expanded this to see whether different classes of TEs had any significant associations with non-coding RNA, especially lincRNAs.

Unsurprisingly, we found that TEs consistently made up a lower proportion of nucleotides in CDS-exons across all regulatory elements, when compared to CDS-introns, lincRNA exons and lincRNA introns (Fig 4). An additional enrichment for TEs in Active Promoters within lincRNA introns was also observed in comparison to all other regions investigated in this stage of the analysis. Weak Promoters in both lincRNA introns and CDS introns also

showed TE enrichment compared to both types of exonic regions. The observation that >30% of nucleotides from many of the regulatory elements were derived from TEs was also quite striking. In particular, the observation that lincRNA exonic regions contained the highest RE density for Polycomb Repressed Regions (Fig 4), with a nearly a third of these nucleotides being derived from TEs, suggests that the presence of transposable elements in lincRNA exons may be strongly linked to gene regulation.

Fig 4. Analysis of the co-occurrence of Transposable Elements and Regulatory Elements across non-coding regions A) Probability estimates for an individual base within each type of non-coding region being part of a Regulatory Element p(RE) or a Transposable Element within each Regulatory Element p(TE|RE). Error bars indicate ±1 Std. Error as calculated on the logit-transformed values. B) $1-\alpha$ Confidence Intervals for the difference between logit-transformed probabilities p(TE|RE), adjusted for multiple comparisons at the level $\alpha = 0.05/6$ within each RE (60). Intervals highlighted in red show significant pairwise differences (confidence intervals do not cross the 0 difference value).

Associations of TEs with gene model features and gene expression

Next, we summarized the overall distribution of transposable elements within various components of the gene model, by finding genes containing TEs across single or multiple components (Fig S1, Table 1), and genes containing one or more types of TEs (Fig S2, Table 2). We further examined the relationship between gene length and which components of a gene contain a TE (Fig S3), as well as the relationship between gene length and the presence of a specific type of TE (Fig S4), using a Wilcoxon Test (Tables S2 & S3) in both

187 cases. We found that only genes with TEs in the 3'UTR or within multiple genic regions
188 showed a bias towards longer length, whilst for TEs exclusively within the proximal
189 promoter or 5'UTR there was a bias towards shorter genes (Fig S3; Table S2). When
190 assessing the relationship between gene length and the presence of a specific TE class, the
191 length of genes with Alu, L2 or MIR elements alone were very similar to genes with no TE,
192 whilst L1 and LTR elements showed a bias towards shorter genes, and the presence of
193 multiple elements biased towards longer genes (Fig S4; Table S3).

194 **Table 1:** Total counts of elements within each genomic region, along with the number of
195 genes with Transposable Elements in one region only.

Region	Total	Elements in Single Regions	Proportion
Proximal Promoter	9462	4835	0.511
5'UTR	4913	1528	0.311
CDS	655	131	0.200
3'UTR	5640	1766	0.313

196 **Table 2:** Total counts of each TE element, along with how many are found in isolation, i.e.
197 in genes with no other elements.

Element Type	Total	Found in Isolation	Proportion
Alu	9460	2199	0.232
L1	5111	539	0.105
L2	5863	687	0.117
LTR	4089	363	0.089
MIR	7638	1423	0.186

198

Effects on the probability of a gene being detected as expressed due to the presence of a TE across the different component of the gene model

As chromatin states are not always indicative of changes in transcriptional activity, we investigated any effects on human gene expression due to the presence of specific TE classes within each of the four regulatory regions, i.e., Proximal Promoter, 5'UTR, CDS and 3'UTR. However, as TEs are far less frequent in CDS regulatory regions with the vast majority co-occurring with other TEs (Figure S1), the subsequent analysis instead focused on the other three regions. Six human tissue transcriptome datasets (adipose, brain, kidney, liver, skeletal muscle and testes tissue) were selected from the Illumina BodyMap2 dataset for this analysis, and global patterns of gene expression were investigated based on the presence or absence of each TE within each of these three genic regions.

The weighted bootstrap method was applied to both the probability of a gene being detected as expressed (Fig 5A), and to the overall expression levels for those genes detected as expressed (Fig 5B). This revealed that Alu elements are commonly associated with a higher probability of expression when located in either the 5'UTR or the 3'UTR across the majority of tissues. In contrast to the presence of an Alu, the presence of L1 elements in the Proximal Promoter showed a negative impact on the probability of a gene being detected as expressed in 3 of the 6 tissues, with the remaining tissues being directionally consistent and quite likely to be Type II errors (Supplementary Fig S6A).

Fig 5. Effects of the presence of each TE in each genomic region. A) Confidence Intervals for the difference in the probability of a gene being detected as expressed due to the presence of each TE in each genic region. B) Confidence Intervals for the difference in mean log2(TPM) counts. For both A) and B), Confidence Intervals were obtained using the weighted bootstrap and are $1 - \alpha/m$ intervals, where $\alpha = 0.05$ and $m = 90$ as the total number of intervals presented. Dots represent the median value from the bootstrap procedure, whilst the vertical line indicates zero. Intervals which do not contain zero are coloured red, and indicate a rejection of the null hypothesis, $H_0: \Delta\theta = 0$, where θ represents the parameter of interest.

Effects on the levels of gene expression due to the presence of a TE in each component of the gene model.

Again using the weighted bootstrap approach to minimize any influence of co-occurring elements, the presence of an Alu in the 5'UTR was found to be associated with increased expression levels in five of the six tissues investigated (Fig 5B). Similarly, Alu elements in the Proximal Promoter were associated with increased expression in two of the tissues. Alu elements in 3'UTR were associated with elevated expression levels in the Kidney sample only. The presence of ion

elements showed varying degrees of reduced gene expression across the tissues when located in the 3'UTR only. It was also noted that whilst strongly controlling the family-wise Type-I error rate (FWER), the adjusted confidence intervals will result in an increase in the

Type-II error rate where true differences are not able to be detected. As such, the point at which the confidence intervals would include zero was found and taken as a proxy for the p-value. Confidence intervals based on these p values to an FDR of 0.05 are shown in Supplementary Figure S6 with the p values given in Supplementary Table S4. It is clear from this additional approach that the role of TEs such as L1 elements in Proximal Promoters and 3'UTRs, LTR elements in 5'UTRs and many of the elements in the 3'UTR may have been considerably understated in this more conservative approach.

Analysis of genes with exapted or exonized TEs

TEs may influence gene expression in different ways, thus we evaluated the possible functional effects of repetitive elements in the human genome, the six primary repeat classes were mapped to the human genome (http://www.repeatmasker.org). Genic regions (annotated using Gene Ontology) that overlapped with TEs were analyzed to assess the association of TEs with different gene functions.

The three fundamental GO categories are: cellular component, molecular function and biological process. Enrichment information for each GO category is listed in Supplementary Table S5. We discovered that for the biological process category (Fig 6, Table S5a), the predominant types of annotation were related to regulatory processes involving metabolic processes. This was consistent with the annotations for the cellular component terms, which were predominantly for intracellular/cytoplasmic structures (Fig S7, Table S5b). The molecular function terms had functions mainly associated with binding (Fig S8, Table S5c).

Using this same method, we also found that genes with protein coding exons containing Alus were enriched for the GO term “intracellular non-membrane-bounded organelle”. Interestingly, these exonization/exaptation events were found associated with splice variants when incorporating Alu sequences (Table S6 & Fig S9, S10).

Fig 6. Enrichment of GO terms of genes containing TEs in promoter, 5’UTR and 3’UTR regions.

Enrichment of GO terms of genes containing TEs in “Biological Process”. Genes containing different types of repetitive elements in the proximal promoter regions are labeled as "Promoter with Repeats", and Genes containing repetitive elements in UTR regions are labeled as "5/3UTR with Repeats". Genes named “Combined Repeats” are the combined data from 3 regions we mentioned above. The darker the color, the greater the GO term enrichment as determined by FDR.

Moreover, according to our analysis of TEs and alternative splicing data, we found that 2.98% of alternatively spliced transcripts contained TEs within protein coding exons (Table 3). Alu and MIR were more likely to be involved in alternative splicing and exonization, which is consistent with previous studies showing that exonization of SINEs occurred in primates [26]. Based on our study, LINEs may also have contributed to these splice variant activities. This shows that exonization of TEs could potentially increase the coding and regulatory versatility of the transcriptome.

Table 3. The number of repeats in protein coding regions (CDS-exon) with alternative splicing. Repeats were counted only if they overlapped CDS-exon regions by at least 25 bps.

Repeat Class	Alu	MIR	L1	L2	LTR	ERV	All repeats
TEs in CDS-exon containing AS	329	224	186	213	229	15	1,402

DISCUSSION

In this work, we have primarily analyzed the distribution of various classes of transposable elements, and their association with regulatory elements (active chromatin) and gene expression in the human genome. Based on the analysis of the TE distributions in genic regions and corresponding gene expression patterns, the presence of some TEs was found to be associated with changes in gene expression. Further, gene function as defined by GO term analysis differed depending on the TE insertion site within the gene. Finally, we looked at TEs present in ncRNAs, specifically lincRNAs, and found that repetitive elements were present at higher levels in lincRNAs than coding exons.

Considering the association between the location of TEs in genes, we found that genes had a greater proportion of sequence originating from TEs in the 5' and 3'UTRs compared to coding exons (Fig 2). This is not surprising considering the potential adverse effect of TE insertion in a protein coding sequence, but it is also relevant with respect to the known

295 regulatory functions within the UTRs [27, 28]. The repeat content for 5'UTR introns was
296 comparable to other types of introns, but this may be significant in the context of
297 transcriptional repression, where genes with shorter 5' UTR introns are expressed at higher
298 levels [29, 30].

299 Furthermore, the presence of TEs in genomic regions that can be epigenetically modified to
300 regulate transcription through active chromatin [31], was consistent with our findings that
301 TEs have a potential role as regulators of gene expression. From our results (Fig 2), we
302 found that some functional regions of active chromatin contained higher percentages of TEs,
303 especially Polycomb Repressed Regions and Weak Enhancers. This fits with the existing
304 theory of epigenetic silencing of TEs [31], since TEs in Polycomb Repressed Regions
305 would inevitably be silenced, and was also consistent with the high repeat content of
306 5'UTRs, which are also known to regulate gene expression [32]. Furthermore, it has been
307 shown that TEs in 3'UTRs are associated with lower transcript abundance [33], and with the
308 clear exception of Alu elements, we have presented further evidence of this. This suggests
309 that exaptation of repetitive elements into regulatory regions is most often associated with
310 repression of gene expression. The general theme of TEs having a role in transcriptional
311 repression was further supported with lincRNAs, which are known to regulate gene
312 expression through epigenetic mechanisms [34] and competition with transcription factors
313 [25]. In our analysis, lincRNA exons were found to be clearly enriched for Polycomb
314 Repressed Regions whilst the abundance of TEs within these regions was relatively
315 consistent with both CDS and lincRNA introns (Fig 4). However, this overall enrichment

316 for repressed regions is consistent with previous research that lincRNAs containing TEs can
317 reduce gene expression in many tissues and cell lines [13]. As different repeat classes were
318 also found to be present at different levels in active chromatin or in specific regulatory
319 regions, such as Polycomb Repressed Regions (Fig 3), this suggests a function with respect
320 to gene expression.

321 Figures 5B and S6B summarize our findings with regard to gene expression, and indicate
322 that Alu elements in 3'UTR, 5'UTR and proximal promoter regions are commonly
323 associated with increased gene expression. Taken in addition with the increased probability
324 of expression due to the presence of an Alu in the 5'UTR and 3'UTR (Fig 5A), these results
325 support previous reports showing TEs such as Alus can be exapted as transcription factor
326 binding sites [35-37], but are in contrast with reports concerning the direction of expression
327 for human genes. We also found genes containing L1 elements were associated with
328 decreasing gene expression (Fig 5), and that L1 elements were less prevalent in regulatory
329 elements or active chromatin, when compared to other repeat classes (Fig 3). This makes
330 intuitive sense as most L1 elements in the human genome are 5' truncated and lack
331 promoter content compared to Alu elements [38]. This is also consistent with a previous
332 study showing that highly and broadly expressed housekeeping genes can be distinguished
333 by their TE content, with these genes being enriched for Alus and depleted for L1s [39].
334 LTRs were found associated with repression of gene expression, which is in contrast to
335 previous work that implicated LTRs as alternative promoters [40]. Anecdotally, it has been
336 shown that an LTR in the first intron of the equine TRPM gene suppresses gene expression

by acting as an alternative poly-A site [41], and the insertion of LTRs in introns has been associated with premature termination of transcription [42], supporting the results presented here. L2 and MIR are ancient TE families conserved among mammals, and are regarded as inactive or fossil TE elements [43]. However, these TEs showed a level of association with reduced gene expression when located in 3'UTRs (Fig 5), which is also consistent with a previous finding on their ability to impact the evolution of gene 3' ends by containing cis-elements for modified polyadenylation [44].

In addition to potentially altering gene expression by insertion into regulatory elements, TEs may also be associated with specific functional characteristics of expressed protein coding genes. When we examined the functional annotation of repeat containing genes, we found that some functions were over-represented (Table S5). Perhaps the most interesting of these associations was that genes with Alu insertions were found to contribute to coding exons through alternative splice variants. One explanation of this observation is that Alu-induced alternative transcripts may result in nonsense mediated decay of alternative transcripts [45]. Two examples of alternatively spliced genes of this type with implications for human disease are *DISC1* and *NOS3* (Table S6 and Fig S9 & S10). *DISC1* alternative transcripts are known to contribute to increased risk of schizophrenia [46, 47] and *NOS3* transcript variants are associated with cardiovascular disease phenotypes [48, 49]. Based on previous research, nearly 4% of protein-coding sequences include transposable elements, and one-third of them are Alu insertions [50]. Therefore, Alu exonization in protein-coding genes may play an important role in modifying gene expression.

In conclusion, while there are many publications implicating TEs in the regulation of individual genes, our work clarifies some previous uncertainties and resolves some contradictions, confirming that this role of TEs is significant across the genome. In general, most TEs would appear to be strongly associated with repression of gene expression, either through the 5'UTR or perhaps as components of lincRNA exons. However, the presence of Alus in 3'UTR and proximal promoter regions may act to increase gene expression. These results are consistent with some previous published research [10] and provide a new understanding of how repeats are associated with epigenetic regulation of gene expression. Finally, while exapted TEs may contribute to the generation of transcripts that undergo nonsense mediated decay as part of gene regulation, we speculate that they may also provide an opportunity for alternative splicing and novel exaptation. TEs therefore are important agents of change with respect to the evolution of gene expression networks.

MATERIAL AND METHODS

Theoretical framework and methods

We constructed pipelines to analyze the distribution of repetitive elements in different parts of the human genome. Repetitive elements overlapping with protein coding regions, non-coding regions and regulatory elements were identified. GO term over-representation and expression analyses were carried out for repetitive elements overlapping with protein-coding regions. The pipelines and related materials are described below.

Tools used to develop pipelines for repetitive element analysis

379 The identification and classification of TEs from the human genome was conducted by
380 developing a pipeline with Perl, R [51], and BEDTools [52]. Perl was used to extract
381 information from different datasets. R was used to build graphs to illustrate the repeat
382 distribution in different genic regions, the identification of repetitive elements with respect
383 to functional elements, GO term over-representation analysis and expression analysis of TEs.
384 BED format file intersection was used to extract the overlapping regions between different
385 datasets, with a lower limit of 25-bps. The UCSC Genome Browser [53, 54] was used to
386 download genome sequence data and genome annotations including RefSeq genes. RSEM
387 [55] was adopted to assemble RNA-Seq reads into transcripts and estimate their abundance
388 (measured as transcripts per million (TPM)). Plots were generated using ggplot2 in R [56].

389 **Datasets**

390 **Genomes and annotations.** NCBI's Human genome and its annotation datasets (RefSeq
391 hg19) [57] were downloaded from the UCSC Genome Browser [23, 53]. A total of 37,697
392 human RefSeq transcripts were merged into 18,777 genes by taking the longest transcript(s)
393 that represented each distinct gene locus. Repetitive elements were downloaded from the
394 RepeatMasker (http://www.repeatmasker.org) track of the UCSC Genome Browser. All
395 repetitive sequence intervals were also de-duplicated to deal with potential overlapping
396 repeat annotations. Overall, there were 5,298,130 human repetitive elements which
397 represented approximately 1.467Gb in the human genome.

Regulatory element datasets from six human cell lines. The regulatory element datasets from six human cell lines were downloaded from the UCSC Genome Browser. Each cell line dataset contained the annotation of six regulatory elements: 1) Active Promoters, 2) Weak Promoters, 3) Strong Enhancers, 4) Weak Enhancers, 5) Insulators and 6) Polycomb Repressed Regions. These regulatory element annotations were derived from different chromatin states that have been marked by histone methylation, acetylation and histone variants H2AZ, PolIII, and CTCF [23].

Gene expression datasets from six human tissues. Human RNA-seq data from the Illumina bodyMap2 transcriptome (Paired End reads only) (http://www.ebi.ac.uk/ena/data/view/ERP000546) dataset was used to measure the association between TEs and the expression levels of genes containing TEs in six tissues.

The distribution of repetitive elements in the human genome. To assess how human TEs were distributed in genes, we compared different genic regions containing TEs. Based on Repbase [58, 59] annotations identified by RepeatMasker (http://www.repeatmasker.org), repeat elements in human were divided into two categories: human-specific repeats, and repeats shared with different species. Human-specific repeats were those annotated with "Homo sapiens" or "primates" as their origin (See Table S6 for the list of human-specific repeat classes), whilst those remaining were categorized as shared repeats. Intergenic regions as well as the exons and introns within 5'UTR, CDS and 3'UTR regions from RefSeq genes [60] were then compared with these different categories of repeats. Next, we

generated the summarised distributions of repetitive elements overlapping these regions by calculating the proportions of bases belonging to repetitive elements within each of the combined sets of regions, i.e.:

$$proportion = \frac{total\ length\ of\ repeats\ in\ defined\ regions\ (bps)}{total\ length\ of\ defined\ regions\ (bps)} \quad (1)$$

The code repository for the above can be found at https://github.com/UofABioinformaticsHub/RepeatElements.

The occurrence of transposable elements within regulatory regions. We further explored the association between any TE and the regulatory elements defined above, by calculating the proportion of nucleotides within each of the five sets of genic regions (5'UTR, CDS-exon, CDS-intron, 3'UTR and Intergenic) that were part of a regulatory element for each of the six human cell lines. The proportion of nucleotides that were TEs within each regulatory element were also calculated for each genic region. All proportions were subsequently transformed using the logit function for model fitting across tissues (Table S1) using the model

$$\boldsymbol{y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}} \quad (2)$$

where y_{ijk} is the logit transformed proportions representing $p(\mathrm{TE}|\mathrm{RE})$ across each genic region i, each regulatory element j and tissue k, such that μ is the overall mean, α_i is the effect due to each genic region, β_j is the effect due to each regulatory element with $(\alpha\beta)_{ij}$ representing any changes not accounted for in the first two terms. Tests for normality and homoscedasticity were performed using the Shapiro-Wilk test and Levene's test respectively. Where violations of homoscedasticity were found robust standard errors were

437 obtained using the sandwich estimator [61]. Confidence Intervals for pairwise comparisons
438 were obtained as implemented in the R package *multcomp* [62] in order to control the Type
439 I error at α=0.05 across the entire set of comparisons.
440 A specific TE analysis was then performed using six of the major human TE classes based
441 on the Repbase classification system: Alu, L1, L2, MIR, LTR and DNA. L1 elements were
442 further resolved into either ancestral (L1M, L1PB and L1PA subfamilies) or recent/clade-
443 specific (L1HS subfamily), based on their Repbase annotations. The proportions of
444 nucleotides within each TE type that were also regulatory elements were calculated giving
445 tissue-specific estimates of $p(\mathrm{RE}|\mathrm{TE})$. Proportions were again transformed using the logit
446 function, and the same analysis as above was performed.
447 **Quality control and preprocessing of the gene expression data in different human**
448 **tissues.** RNA-seq reads of six human tissues were first assessed using FastQC software
449 (www.bioinformatics.babraham.ac.uk), to provide an overview of whether the raw RNA-
450 Seq data contained any problems or biases before further analysis. Reads with poor-quality
451 bases were trimmed (based on the results of FastQC with MINLEN set to 26) for subsequent
452 data analysis. Table 4 showed the numbers of reads in raw RNA-Seq datasets and the
453 statistics after the QC process by using Trimmomatic-0.32 [63]. Then, we built transcript
454 reference sequences using rsem-prepare-reference [64] from Hg19 human RefSeq genes.
455 The references were then input to rsem-calculate-expression [64] using default parameters
456 for all 6 tissues to obtain TPM based expression values.

Table 4. Description of RNA-Seq datasets and QC results. Reads with poor-quality bases were trimmed using FastQC (with MINLEN set to 26, HEADCROP set to 13 and LEADING set to 15).

	Raw Data (Read Number)	**High Quality reads (%)**			**Low Quality reads (%)**
		Both Reads	Forward Only Reads	Reverse Only Reads	
Kidney	8,039,733,700	78.72	9.84	4.87	6.58
Liver	8,004,862,300	81.25	8.11	4.67	5.96
Brain	7,351,304,700	69.31	7.76	12.86	10.06
Testes	8,183,619,900	76.9	9.88	5.51	7.71
Adipose	7,730,007,200	79.88	9.74	4.49	5.89
Skeletal Muscle	8,211,113,900	76.06	12.12	4.87	6.94

Proximal promoter regions were defined as 1,000bp upstream of the gene transcription start sites based on the longest transcripts for each gene. Alu, MIR, L1, L2 and LTR repeat regions were then identified within the proximal promoters, 5'UTR, CDS and 3'UTR regions.

The weighted bootstrap procedure for assessing the effects of a TE in each genic region. Many genes contain multiple transposable elements, with only a minority of genes containing a single TE (Fig S2). In order to assess any effects on transcription due to the presence of a single TE, a weighted bootstrap approach was devised. For a given TE within each genic region within each individual tissue, the frequencies of co-occurring TEs and combinations of TEs were noted. Uniform sampling probabilities were then used for the set of genes containing a specific TE in a specific region, whilst sampling weights were assigned to genes lacking the specific TE based on TE composition, such that the TE content of the sampled set of reference genes matched that of the test set of genes, based on

474 the defined categories. Gene length was divided into 10 bins and these were included as an
475 additional category when defining sampling weights. This ensured that two gene sets were
476 obtained for each bootstrap iteration, which were matched in length and TE composition
477 with the sole difference being the presence of the specific TE within each specific genic
478 region (Figure S5). The mean difference in expression level, as measured by log(TPM), and
479 the difference in the proportions of genes detected as expressed were then used as the
480 variables of interest in the bootstrap procedure. The bootstrap was performed on sets of
481 1000 genes for 10,000 iterations using the proximal promoter as defined above, along with
482 5'UTR and 3'UTRs. When comparing expression levels, genes with zero read counts were
483 omitted prior to bootstrapping. In order to compensate for multiple testing considerations,
484 confidence intervals were obtained across the $m = 90$ tests at the level $1 - \alpha/m$, which is
485 equivalent to the Bonferroni correction, giving confidence intervals which controlled the
486 FWER at the level $\alpha = 0.05$. Approximate two-sided p-values were also calculated by
487 finding the point at which each confidence interval crossed zero, and additional significance
488 was determined by estimating the FDR on these sets of p-values using the Benjamini-
489 Hochberg method.

490 **Long intergenic non-coding RNAs and TEs.** Annotations for 8,196 previously described
491 putative human lincRNAs were downloaded [65] and the distribution of TEs within
492 regulatory elements in lincRNA exons and introns was obtained using the same methods as
493 above. The previously described regression models were then used to analyse this dataset.

Association of functional elements with human repetitive elements. To demonstrate the potential functional significance of repetitive elements, the Database for Annotation, Visualization and Integrated Discovery (DAVID) [66] was used to perform the GO classification. We first extracted Gene-IDs from overlapping regions between different gene categories (1000bp proximal promoter, 5'UTR, 3'UTR, and the combination of these 3 regions) and TEs. These gene-lists were then submitted to the DAVID Functional Classification Tool. We chose the third level of GO terms to describe the over-represented functional terms for the three datasets and visualized the functional over-representation of overlapped genes using the R package *heatmap.2*. The p-value was applied in the GO analysis as the standard index to determine the degree of enrichment. The threshold for over-represented GO terms was set to an FDR (Benjamini-Hochberg method) less than 0.05. Protein-coding genes with Alus were also visualised with the UCSC genome browser (http://genome.ucsc.edu/) to compare their mRNA with various gene datasets and annotations.

Association of alternative splicing and protein coding regions containing TEs. In order to assess the relationship between transposable elements and exonization, an alternative splicing annotation dataset (SIB Alt-splicing) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/). These data were generated from RefSeq genes, Genbank RNAs and ESTs that aligned to the human genome. A total of 46,973 alternatively spliced transcripts were intersected with gene models containing transposable elements.

ACKNOWLEDGEMENT

The authors wish to thank Dan Kortschak, Atma Ivancevic, Joy Raison, Reuben Buckley and Sim Lin Lim for valuable discussions and critical reading of drafts.

AUTHOR CONTRIBUTION

DLA and CCW conceived, designed and managed the study. LZ collected the datasets, implemented the analysis pipeline, and analyzed the data. SMP analyzed the data. ZPQ, DFC, ZQH prepared datasets. LZ, DLA and CCW wrote and revised the manuscript. All authors reviewed and approved the final manuscript.

CONFLICT OF INTEREST

The author(s) declare that they have no competing interests.

REFERENCES

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current opinion in genetics & development 9: 657-663.

Cordaux R and Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nature reviews Genetics 10: 691-703.

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature genetics 42: 631-634.

Lynch VJ, Leclerc RD, May G and Wagner GP (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nature genetics 43: 1154-1159.

542 Cowley M and Oakey RJ (2013) Transposable elements re-wire and fine-tune
543 the transcriptome. PLoS genetics 9: e1003234.
544 Pereira V, Enard D and Eyre-Walker A (2009) The effect of transposable
545 element insertions on gene expression evolution in rodents. PloS one 4: e4321.
546 Britten RJ (1996) DNA sequence insertion and evolutionary variation in gene
547 regulation. Proceedings of the National Academy of Sciences of the United
548 States of America 93: 9374-9377.
549 van de Lagemaat LN, Landry JR, Mager DL and Medstrand P (2003)
550 Transposable elements in mammals promote regulatory variation and
551 diversification of genes with specialized functions. Trends in genetics : TIG 19:
552 530-536.
553 Jjingo D, Huda A, Gundapuneni M, Marino-Ramirez L and Jordan IK (2011)
554 Effect of the transposable element environment of human genes on gene length
555 and expression. Genome biology and evolution 3: 259-271.
556 Han JS, Szak ST and Boeke JD (2004) Transcriptional disruption by the L1
557 retrotransposon and implications for mammalian transcriptomes. Nature 429:
558 268-274.
559 Rebollo R, Romanish MT and Mager DL (2012) Transposable elements: an
560 abundant and natural source of regulatory sequences for host genes. Annual
561 review of genetics 46: 21-42.
562 Kelley D and Rinn J (2012) Transposable elements reveal a stem cell-specific
563 class of long noncoding RNAs. Genome biology 13: R107.
564 Jacques PE, Jeyakani J and Bourque G (2013) The majority of primate-specific
565 regulatory sequences are derived from transposable elements. PLoS genetics 9:
566 e1003504.
567 Conley AB, Piriyapongsa J and Jordan IK (2008) Retroviral promoters in the
568 human genome. Bioinformatics 24: 1563-1567.
569 Medstrand P, Landry JR and Mager DL (2001) Long terminal repeats are used
570 as alternative promoters for the endothelin B receptor and apolipoprotein C-I
571 genes in humans. The Journal of biological chemistry 276: 1896-1903.
572 Franchini LF, Lopez-Leal R, Nasif S, Beati P, Gelman DM, et al. (2011)
573 Convergent evolution of two mammalian neuronal enhancers by sequential
574 exaptation of unrelated retroposons. Proceedings of the National Academy of
575 Sciences of the United States of America 108: 15270-15275.
576 Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, et al. (2007)
577 Comparative analysis of transposed element insertion within human and
578 mouse genomes reveals Alu's unique role in shaping the human transcriptome.
579 Genome biology 8: R127.
580 Piriyapongsa J, Polavarapu N, Borodovsky M and McDonald J (2007)
581 Exonization of the LTR transposable elements in human genome. BMC
582 genomics 8: 291.

583 Hadjiargyrou M and Delihas N (2013) The Intertwining of Transposable
584 Elements and Non-Coding RNAs. International journal of molecular sciences 14:
585 13307-13328.
586 de Souza FS, Franchini LF and Rubinstein M (2013) Exaptation of transposable
587 elements into novel cis-regulatory elements: is the evidence always strong?
588 Molecular biology and evolution 30: 1239-1251.
589 Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, et al. (2014)
590 Mammalian-wide interspersed repeat (MIR)-derived enhancers and the
591 regulation of human gene expression. Mobile DNA 5: 14.
592 Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping
593 and analysis of chromatin state dynamics in nine human cell types. Nature 473:
594 43-49.
595 McCue AD and Slotkin RK (2012) Transposable element small RNAs as
596 regulators of gene expression. Trends in genetics : TIG 28: 616-623.
597 Gong C and Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA
598 decay by duplexing with 3' UTRs via Alu elements. Nature 470: 284-288.
599 Krull M, Petrusma M, Makalowski W, Brosius J and Schmitz J (2007) Functional
600 persistence of exonized mammalian-wide interspersed repeat elements (MIRs).
601 Genome research 17: 1139-1145.
602 Belancio VP, Hedges DJ and Deininger P (2006) LINE-1 RNA splicing and
603 influences on mammalian gene expression. Nucleic acids research 34: 1512-
604 1521.
605 Belancio VP, Roy-Engel AM and Deininger P (2008) The impact of multiple
606 splice sites in human L1 elements. Gene 411: 38-45.
607 Cenik C, Chua HN, Zhang H, Tarnawsky SP, Akef A, et al. (2011) Genome
608 analysis reveals interplay between 5'UTR introns and nuclear mRNA export for
609 secretory and mitochondrial genes. PLoS genetics 7: e1001366.
610 Cenik C, Derti A, Mellor JC, Berriz GF and Roth FP (2010) Genome-wide
611 functional analysis of human 5' untranslated region introns. Genome biology
612 11: R29.
613 Slotkin RK and Martienssen R (2007) Transposable elements and the
614 epigenetic regulation of the genome. Nature reviews Genetics 8: 272-285.
615 Barrett LW, Fletcher S and Wilton SD (2012) Regulation of eukaryotic gene
616 expression by the untranslated gene regions and other non-coding elements.
617 Cellular and molecular life sciences : CMLS 69: 3613-3634.
618 Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, et al. (2009) The regulated
619 retrotransposon transcriptome of mammalian cells. Nature genetics 41: 563-
620 571.
621 Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, et al. (2013)
622 The Xist lncRNA exploits three-dimensional genome architecture to spread
623 across the X chromosome. Science 341: 1237973.

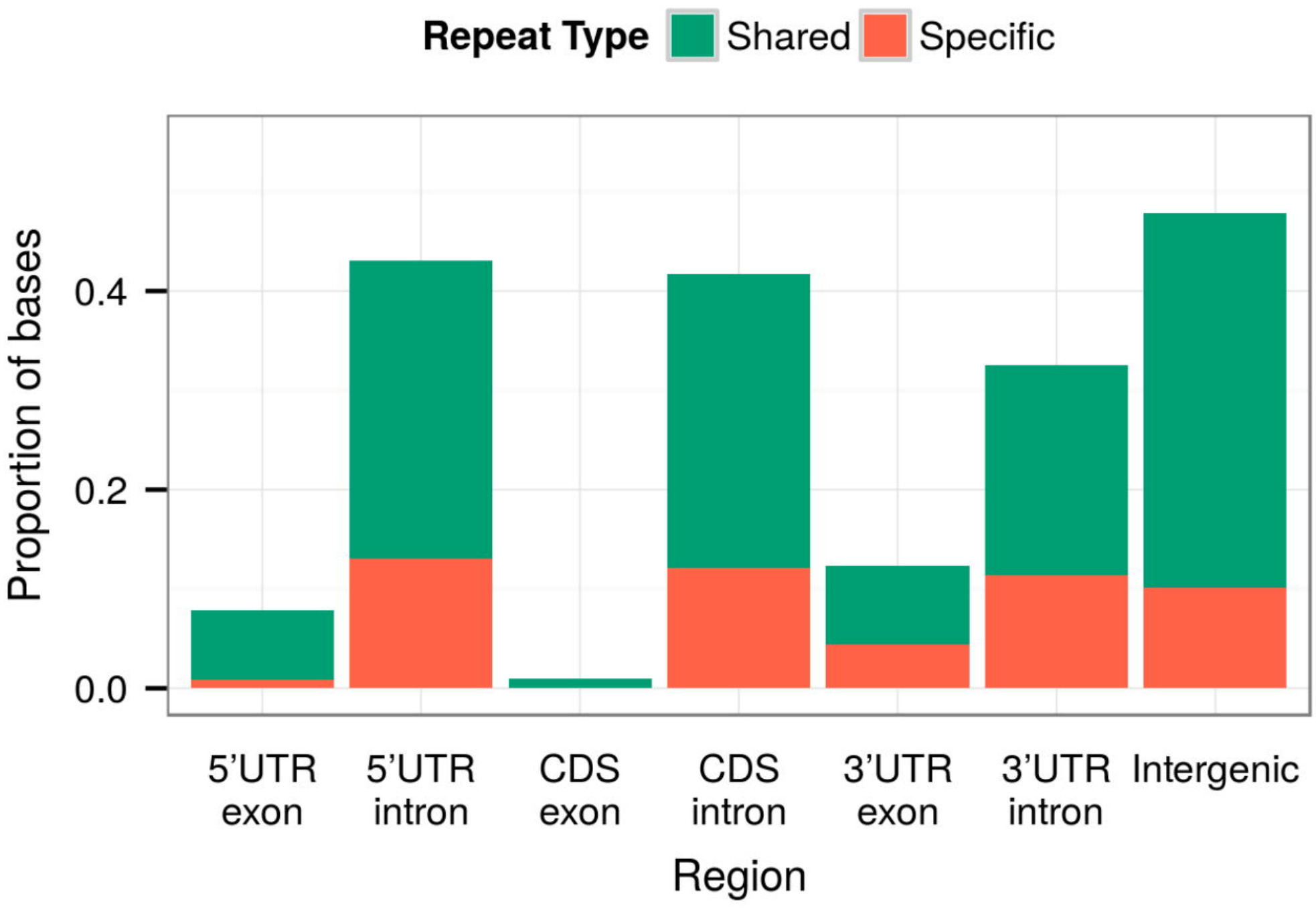
624 Kim DD, Kim TT, Walsh T, Kobayashi Y, Matise TC, et al. (2004) Widespread
625 RNA editing of embedded alu elements in the human transcriptome. Genome
626 research 14: 1719-1725.
627 Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, et al. (2004)
628 Systematic identification of abundant A-to-I editing sites in the human
629 transcriptome. Nature biotechnology 22: 1001-1005.
630 Lin L, Jiang P, Shen S, Sato S, Davidson BL, et al. (2009) Large-scale analysis of
631 exonized mammalian-wide interspersed repeats in primate genomes. Human
632 molecular genetics 18: 2204-2214.
633 Lavie L, Maldener E, Brouha B, Meese EU and Mayer J (2004) The human L1
634 promoter: variable transcription initiation sites and a major impact of
635 upstream flanking sequence on promoter activity. Genome research 14: 2253-
636 2260.
637 Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, et al. (2007) Repetitive
638 sequence environment distinguishes housekeeping genes. Gene 390: 153-165.
639 Cohen CJ, Lock WM and Mager DL (2009) Endogenous retroviral LTRs as
640 promoters for human genes: a critical assessment. Gene 448: 105-114.
641 Bellone RR, Holl H, Setaluri V, Devi S, Maddodi N, et al. (2013) Evidence for a
642 retroviral insertion in TRPM1 as the cause of congenital stationary night
643 blindness and leopard complex spotting in the horse. PloS one 8: e78280.
644 Guntaka RV (1993) Transcription termination and polyadenylation in
645 retroviruses. Microbiological reviews 57: 511-521.
646 Deininger PL and Batzer MA (2002) Mammalian retroelements. Genome
647 research 12: 1455-1465.
648 Lee JY, Ji Z and Tian B (2008) Phylogenetic analysis of mRNA polyadenylation
649 sites reveals a role of transposable elements in evolution of the 3'-end of genes.
650 Nucleic acids research 36: 5581-5590.
651 Lewis BP, Green RE and Brenner SE (2003) Evidence for the widespread
652 coupling of alternative splicing and nonsense-mediated mRNA decay in
653 humans. Proceedings of the National Academy of Sciences of the United States
654 of America 100: 189-192.
655 Callicott JH, Straub RE, Pezawas L, Egan MF, Mattay VS, et al. (2005) Variation in
656 DISC1 affects hippocampal structure and function and increases risk for
657 schizophrenia. Proceedings of the National Academy of Sciences of the United
658 States of America 102: 8627-8632.
659 Rapoport JL, Addington AM, Frangou S and Psych MR (2005) The
660 neurodevelopmental model of schizophrenia: update 2005. Molecular
661 psychiatry 10: 434-449.
662 Pacanowski MA, Zineh I, Cooper-Dehoff RM, Pepine CJ and Johnson JA (2009)
663 Genetic and pharmacogenetic associations between NOS3 polymorphisms,
664 blood pressure, and cardiovascular events in hypertension. American journal of
665 hypertension 22: 748-753.

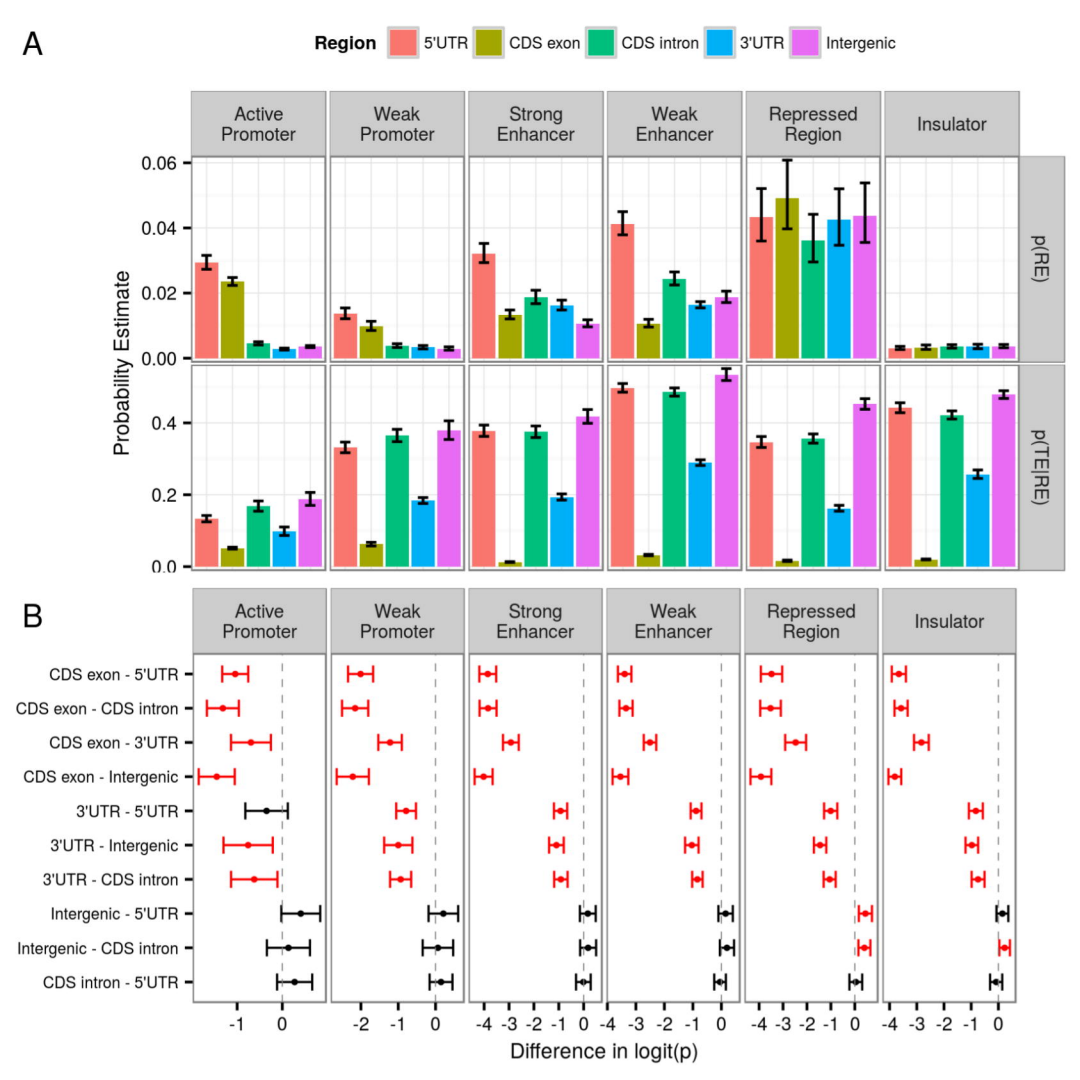
666 Hingorani AD, Liang CF, Fatibene J, Lyon A, Monteith S, et al. (1999) A common
667 variant of the endothelial nitric oxide synthase (Glu298-->Asp) is a major risk
668 factor for coronary artery disease in the UK. Circulation 100: 1515-1520.
669 Nekrutenko A and Li WH (2001) Transposable elements are found in a large
670 number of human protein-coding genes. Trends in genetics : TIG 17: 619-621.
671 R Core Team (2014) R : A language and environment for statistical computing.
672 Vienna, Austria: R Foundation for Statitical Computing.
673 Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for
674 comparing genomic features. Bioinformatics 26: 841-842.
675 Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, et al. (2012) The
676 UCSC Genome Browser database: extensions and updates 2011. Nucleic acids
677 research 40: D918-923.
678 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human
679 genome browser at UCSC. Genome research 12: 996-1006.
680 Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and
681 transcript expression analysis of RNA-seq experiments with TopHat and
682 Cufflinks. Nature protocols 7: 562-578.
683 Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer New
684 York.
685 Pruitt KD, Tatusova T and Maglott DR (2007) NCBI reference sequences
686 (RefSeq): a curated non-redundant sequence database of genomes, transcripts
687 and proteins. Nucleic acids research 35: D61-65.
688 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase
689 Update, a database of eukaryotic repetitive elements. Cytogenetic and genome
690 research 110: 462-467.
691 Kohany O, Gentles AJ, Hankus L and Jurka J (2006) Annotation, submission and
692 screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.
693 BMC bioinformatics 7: 474.
694 Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, et al. (2013) The UCSC
695 Genome Browser database: extensions and updates 2013. Nucleic acids
696 research 41: D64-69.
697 Zeileis A (2004) Econometric Computing with HC and HAC Covariance Matrix
698 Estimators. Journal of Statistical Software 11(10): 1-17.
699 Westfall THaFBaP (2008) Simultaneous Inference in General Parametric
700 Models. Biometrical Journal 50: 346--363.
701 Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: a flexible trimmer for
702 Illumina sequence data. Bioinformatics 30: 2114-2120.
703 Li B and Dewey CN (2011) RSEM: accurate transcript quantification from RNA-
704 Seq data with or without a reference genome. BMC bioinformatics 12: 323.
705 Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative
706 annotation of human large intergenic noncoding RNAs reveals global properties
707 and specific subclasses. Genes & development 25: 1915-1927.

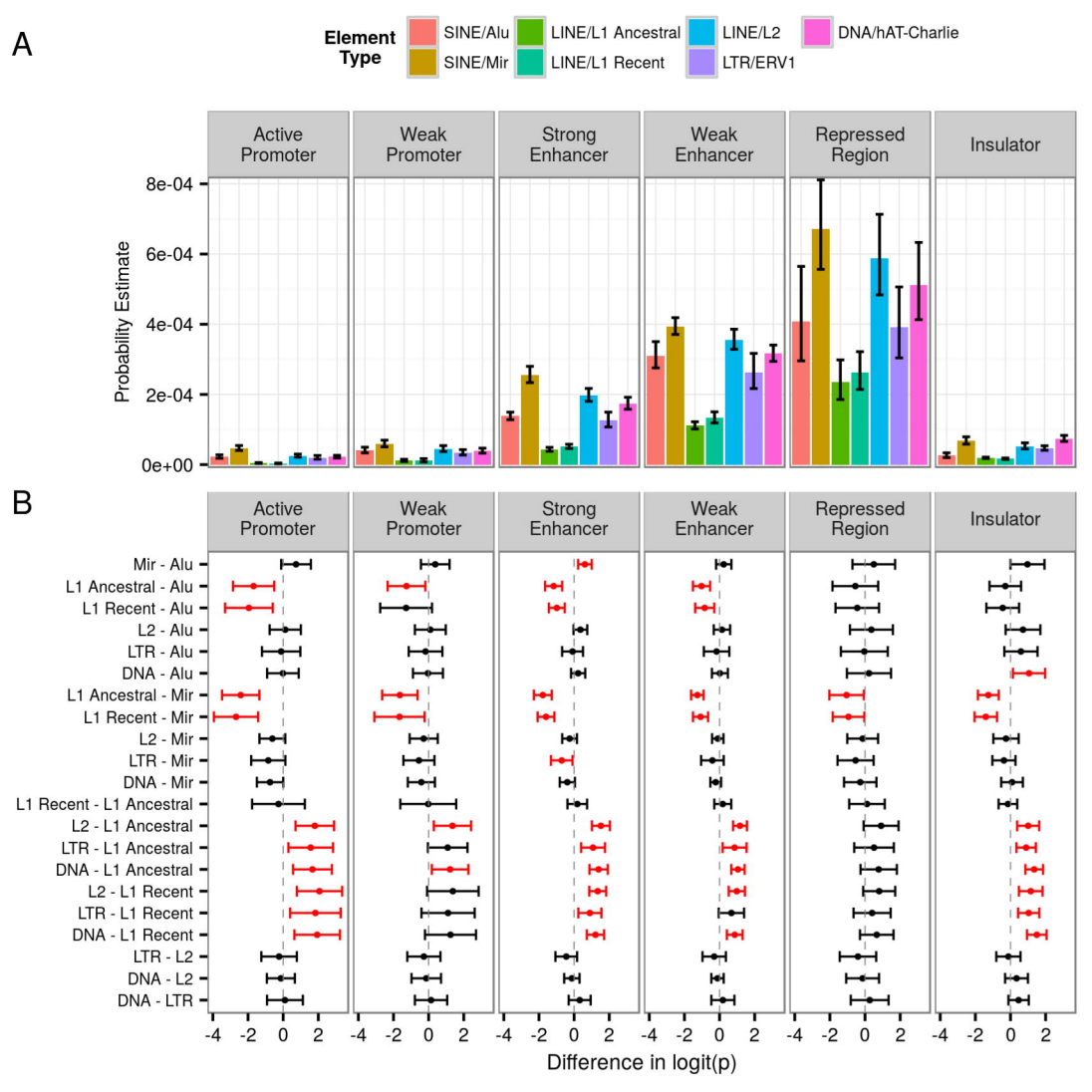
708 Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID:
709 Database for Annotation, Visualization, and Integrated Discovery. Genome
710 biology 4: P3.

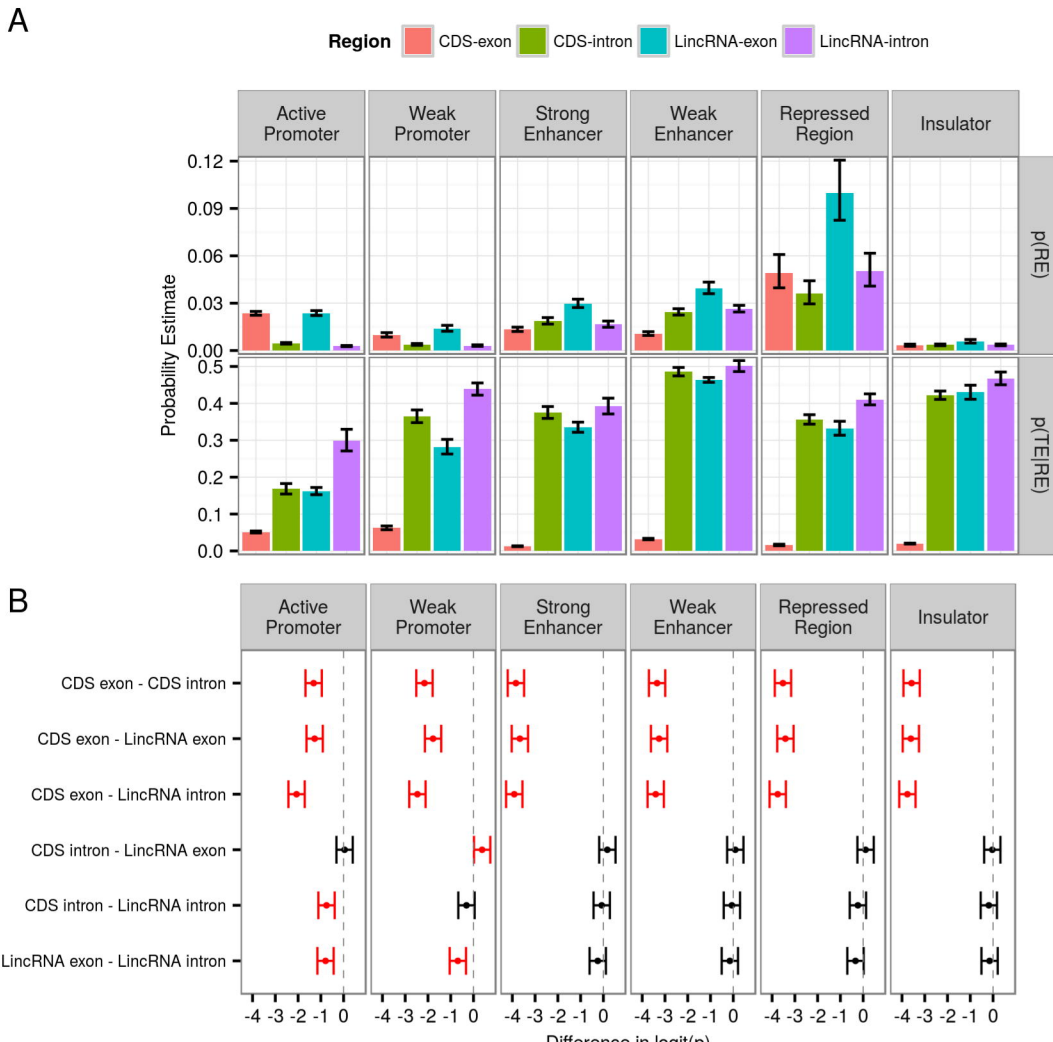
711

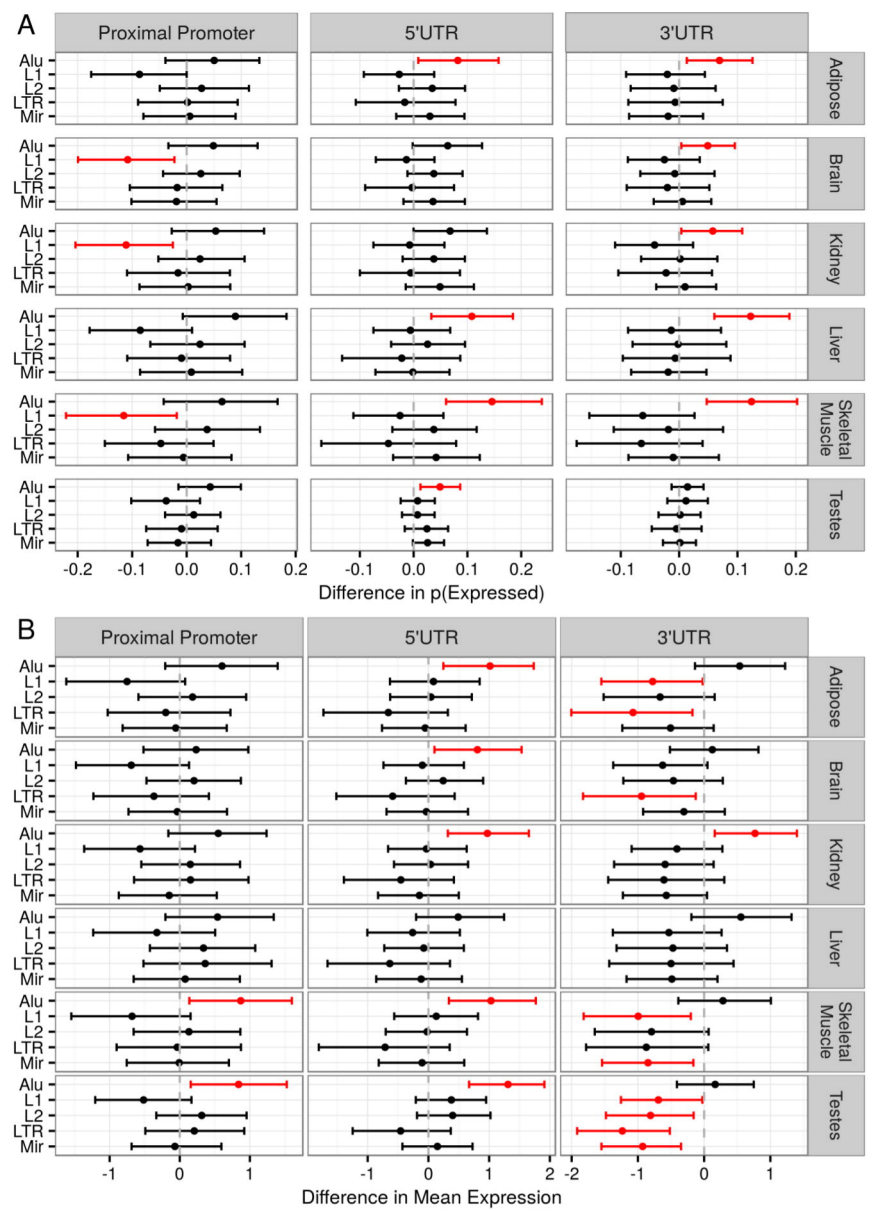
712

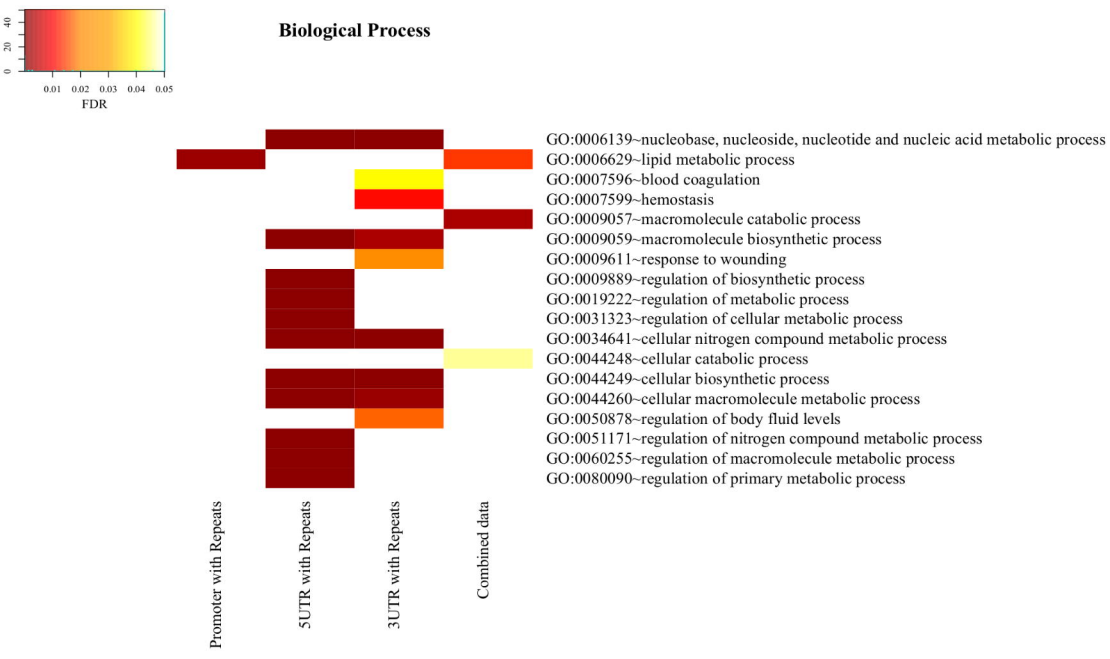

Repeat Type
Shared
Specific
Proportion of bases
0.4
0.2
0.0
5'UTR exon
5'UTR intron
CDS exon
CDS intron
3'UTR exon
3'UTR intron
Intergenic
Region


A
Region
5'UTR
CDS exon
CDS intron
3'UTR
Intergenic
Active Promoter
Weak Promoter
Strong Enhancer
Weak Enhancer
Repressed Region
Insulator
p(RE)
p(TE|RE)
Probability Estimate
0.06
0.04
0.02
0.00
0.4
0.2
0.0
B
CDS exon - 5'UTR
CDS exon - CDS intron
CDS exon - 3'UTR
CDS exon - Intergenic
3'UTR - 5'UTR
3'UTR - Intergenic
3'UTR - CDS intron
Intergenic - 5'UTR
Intergenic - CDS intron
CDS intron - 5'UTR
-1 0
-2 -1 0
-4 -3 -2 -1 0
Difference in logit(p)


A
Element Type
SINE/Alu
SINE/Mir
LINE/L1 Ancestral
LINE/L1 Recent
LINE/L2
LTR/ERV1
DNA/hAT-Charlie
Active Promoter
Weak Promoter
Strong Enhancer
Weak Enhancer
Repressed Region
Insulator
Probability Estimate
8e-04
6e-04
4e-04
2e-04
0e+00
B
Mir - Alu
L1 Ancestral - Alu
L1 Recent - Alu
L2 - Alu
LTR - Alu
DNA - Alu
L1 Ancestral - Mir
L1 Recent - Mir
L2 - Mir
LTR - Mir
DNA - Mir
L1 Recent - L1 Ancestral
L2 - L1 Ancestral
LTR - L1 Ancestral
DNA - L1 Ancestral
L2 - L1 Recent
LTR - L1 Recent
DNA - L1 Recent
LTR - L2
DNA - L2
DNA - LTR
-4 -2 0 2
Difference in logit(p)


A
Region
CDS-exon
CDS-intron
LincRNA-exon
LincRNA-intron
Active Promoter
Weak Promoter
Strong Enhancer
Weak Enhancer
Repressed Region
Insulator
p(RE)
p(TE|RE)
Probability Estimate
0.12
0.09
0.06
0.03
0.00
0.5
0.4
0.3
0.2
0.1
0.0
B
CDS exon - CDS intron
CDS exon - LincRNA exon
CDS exon - LincRNA intron
CDS intron - LincRNA exon
CDS intron - LincRNA intron
LincRNA exon - LincRNA intron
-4 -3 -2 -1 0
Difference in logit(p)


A
Proximal Promoter
5'UTR
3'UTR
Alu
L1
L2
LTR
Mir
Adipose
Brain
Kidney
Liver
Skeletal Muscle
Testes
-0.2
-0.1
0.0
0.1
0.2
Difference in p(Expressed)
B
Proximal Promoter
5'UTR
3'UTR
-1
0
1
2
-2
Difference in Mean Expression


Biological Process
0
20
40
0.01
0.02
0.03
0.04
0.05
FDR
GO:0006139~nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0006629~lipid metabolic process
GO:0007596~blood coagulation
GO:0007599~hemostasis
GO:0009057~macromolecule catabolic process
GO:0009059~macromolecule biosynthetic process
GO:0009611~response to wounding
GO:0009889~regulation of biosynthetic process
GO:0019222~regulation of metabolic process
GO:0031323~regulation of cellular metabolic process
GO:0034641~cellular nitrogen compound metabolic process
GO:0044248~cellular catabolic process
GO:0044249~cellular biosynthetic process
GO:0044260~cellular macromolecule metabolic process
GO:0050878~regulation of body fluid levels
GO:0051171~regulation of nitrogen compound metabolic process
GO:0060255~regulation of macromolecule metabolic process
GO:0080090~regulation of primary metabolic process
Promoter with Repeats
5UTR with Repeats
3UTR with Repeats
Combined data