The regulated expression of repetitive elements across human cell types and tissues

Dave Tang*[1], the FANTOM5 consortium, and Piero Carninci†[1]

[1]*RIKEN Center for Life Science Technologies (Division of Genomic Technologies)*

November 12, 2014

Abstract

A large fraction of the human genome is composed of repetitive elements (REs), many of which are transposable elements (TEs). Most of these TEs are molecular fossils that have long lost the ability to transpose. However, transcription can initiate from these elements as they contain requisite signals necessary for transcription. However, it is unclear whether these transcriptional products have any functional purpose or are just transcriptional by-products. In order to address this question, we performed the most comprehensive survey of transcriptional events arising from TEs by using the FANTOM5 human atlas of Cap Analysis Gene Expression (CAGE) libraries, which consists of a large panel of human cell and tissues types (totalling 988 samples). We observed that TEs were lowly expressed on average compared to protein-coding genes, have restrictive expression patterns, overlap enhancer sequences, serve as precursors to small RNAs, and drive the transcription of long noncoding RNA. Furthermore, based on the expression patterns of TEs, we could associated families of TEs to specific sample ontologies that have been ascribed to the FANTOM5 samples. This hints at the fact that characteristics of certain families of TEs, such as their sequence composition, may be associated to factors that are specific to particular cell types. These lines of evidence support the hypothesis that various TEs have become exapted and have a functional role in the human genome.

*Email: `dave.tang@riken.jp`

†Email: `carninci@riken.jp`; Corresponding author

1 Introduction

Roughly half of the human genome is made up of repetitive elements (REs), consisting mainly of completely inactive transposable elements (TEs) (Lander et al., 2001). The estimated number of active TEs in the human genome is less than 0.05% (Mills et al., 2007), which is as expected given that TEs are a major factor in causing mutations that lead to human disease (Callinan and Batzer, 2006). However, despite the detrimental effects of TEs, their ability to move around has shaped the evolution of the genomes in which they reside (Cordaux and Batzer, 2009). For example, TEs have been found to contribute transcription factor binding sites (TFBSs) to various mammalian transcription factors (Wang et al., 2007; Bourque et al., 2008; Kunarso et al., 2010) and it has been suggested that TEs have played a major part in wiring and re-wiring regulatory networks (Feschotte, 2008). This phenomenon has been observed in both humans and mice, where transcription factors pervasively bound to TEs (Kunarso et al., 2010). However the TEs that served as "repeatassociated binding sites" in human and mouse were from different families of endogenous retroviruses, suggesting the convergent evolution of TEs as binding sites. In addition to TFBSs, a recent study demonstrated that methylation patterns in TEs were hypomethylated in a tissue-specific manner and displayed enhancer hallmarks (Xie et al., 2013). TEs also promote transcript diversity by providing alternative promoters for extant genes. A genome-wide study examining transcription initiation events arising within TEs observed that TEs were used as alternative promoters in normal tissues and contributed to tissue-specific expression profiles (Faulkner et al., 2009). These studies suggest that TEs are developmentally regulated and may have contributed to driving the large diversity of cell types observed in humans.

With the advent of high-throughput sequencing, numerous RNA sequencing studies have revealed a large set of long non-coding RNAs (lncRNAs) in the human genome (Cabili et al., 2011; Khalil et al., 2009; Guttman et al., 2011). TEs have been suggested to have contributed to the biogenesis of many lncRNAs (Kapusta and Feschotte, 2014) and it was demonstrated that lncRNAs are enriched with TE sequences (Kapusta et al., 2013). The distribution of TEs within lncRNAs was also found to be positioned and orientated in a biased with respect to the transcription start site (TSS), suggesting that TEs provided the initial spark that gave rise to the lncRNA (Kelley and Rinn, 2012). As TEs naturally contain transcriptional regulatory signals, this is not surprising; for example, the long terminal repeats (LTRs) of retrotransposons have been found to drive the expression of nearby genes (Cohen et al., 2009). The TSS of the lncRNAs *BANCR*, *lnc-RoR*, and *lncRNA-ES3* all overlap the LTR sequence of separate endogenous retroviruses (Kapusta et al., 2013). In addition, these lncRNAs only exist in the genomes of organisms that descended from the last common ancestor which contains the retrovirus, clearly supporting the idea that these lncRNAs arose from TEs (Kapusta et al., 2013). Furthermore, a specific TE family was found to drive the expression of stem cell specific lncRNAs (Kelley and Rinn, 2012) and a large number of "non-annotated stem transcripts"

(Fort et al., 2014), suggesting that TEs are able confer tissue-specificity possibly by associating with specific regulatory signals.

The consideration that TEs may be functional products of the genome is generally met with scepticism due to historical views that TEs are simply selfish elements that serve no purpose apart from propagating itself (Doolittle and Sapienza, 1980) and therefore provide little to no selective advantage to organisms (Orgel and Crick, 1980). Furthermore, the elaborate silencing systems that repress the transposition of mobile genetic elements in the human genome suggest that they molecular pests that need to be silenced (Yang and Kazazian, 2006). However, in their original paper, Orgel and Crick suggested at the hypothesis that these selfish DNA sequences may become exapted for control purposes (Orgel and Crick, 1980). Indeed, there have been an increasing number of studies reporting the exaptation of TE insertions, however it has been suggested that many studies are not conclusive and more work in this area is required (de Souza et al., 2013). One reason for the limited number of studies is due to the technical difficulties associated with REs. During the era of microarray technologies, cross-hybridisation issues made it difficult to study the expression of REs. High-throughput sequencing technology has made it possible to quantify the expression of REs, however the short read nature of these technologies has made it difficult to map reads to REs.

It has been previously reported that 75% of reads that are 25 nucleotides long can be mapped uniquely to the human genome and at 60 nucleotides long 95% uniqueness can be achieved (Whiteford et al., 2005). However, this estimate assumes an equal distribution of reads in a given library, which is typically not the case. When dealing with a read that maps to multiple places, there are usually three choices on what to do: a) Discard the read, b) Take the best alignment and if there are multiple best hits, take one randomly, and c) Report all alignments or report up to a certain number (Treangen and Salzberg, 2012). The third choice can also be incorporated with a strategy that uses uniquely mapped reads to probabilistically weight multi-mapping reads (Faulkner et al., 2008). This strategy is based on the basic premise that a region that is transcriptional active is more likely to have given rise to a read than a transcriptional inert region. Other strategies for dealing with multi-mapping reads include mapping to a consensus database of REs, such as RepBase Update (Jurka et al., 2005), or mapping to regions of a genome that has been annotated as being repetitive (Xie et al., 2013), or combining reference genome mapping with consensus sequence mapping (Day et al., 2010). These strategies aim to utilise as many reads as possible to obtain a more representative expression profile.

In this study, we processed over two billion reads from 988 FANTOM5 CAGE libraries and overlaid them to REs annotated using profile hidden Markov models. We found that various REs are expressed in a tissue-specific manner and expression signal from REs can be used to produce biologically meaningful clusters. Furthermore, expression profiles of specific REs can be associated to specific sample ontologies, suggesting that specific families of REs may are associated to tissue-tpes. Lastly, by examining the genomic locality of expressed REs, we observed that they overlapped genomic regions known to produce small RNAs

and lncRNAs, as well as enhancer regions, more often than expected by chance.

2 Methods

2.1 Annotating repeats in the human genome

RepeatMasker (Smit et al., 2004) is a program that screens DNA sequences for repetitive elements (REs); the tool relies on a search algorithm and a database of RE profiles. Traditionally, homology-based tools such as cross_match and variants of BLAST have been used to screen DNA sequence against RE consensus sequences, the most commonly used database being Repbase Update (Jurka et al., 2005). Recently a database of REs based on profile hidden Markov models was developed, called Dfam (Wheeler et al., 2013), which allowed screening of REs using a hidden Markov model search tool, called nhmmer (Wheeler and Eddy, 2013). It has been reported that screening for REs using nhmmer and Dfam is more sensitive and specific than consensus sequence based approaches (Wheeler et al., 2013). For this reason, we annotated REs in the human genome (hg19) using RepeatMasker (4.0.3), nhmmer (hmmer-3.1b1), and Dfam (1.2). Specifically, we ran the command `RepeatMasker -e hmmer -species human -s -xsmall -pa 8 chr.fa`, for each assembled chromosome. REs were classified by class, family, and individual element names.

2.2 Aggregating CAGE reads to repetitive elements

The details describing the preparation of the Cap Analysis Gene Expression (CAGE) libraries for the FANTOM5 project are described elsewhere (Kanamori-Katayama et al., 2011). Briefly, a CAGE protocol optimised for the HeliScope Genetic Analysis System was developed and used to prepare 988 FANTOM5 libraries. A high-throughput short read sequence alignment program called Delve (Djebali et al., 2012) was used to map the CAGE reads to the human genome (hg19). Delve is able to recognise sequencing biases or increased error rates in homopolymer stretches, which makes it suitable for the HeliScope sequencer. Mapping qualities following the Phred scale were provided for each mapped read, where the qualities are probabilities that a mapped read is incorrect (Li et al., 2008).

To aggregate CAGE reads to REs, the coordinates of the mapped reads were intersected with the RE coordinates using intersectBed from the BEDTools suite (Quinlan and Hall, 2010); parallelisation of the computations was achieved using GNU parallel (Tange, 2011). For each repeat class (1087 in total), we tallied the number of reads that intersected that class; thus for each FANTOM5 library, a tally was produced for each repeat class resulting in a 1087×988 matrix. We performed this aggregation step using reads thresholded at various mapping qualities $(0, \ldots, 10)$. Finally, tallies for each library were normalised by tags per million (TPM); library size was the total number of reads that intersected the repeat classes.

2.3 Markov clustering

We calculated the Spearman's rank correlation coefficient between all repeat classes (590,241 pairwise calculations) and all libraries (487,578 pairwise calculations) using the matrix of aggregated CAGE reads. Correlations between REs and libraries were represented as a graph, where the nodes or vertices represented a single RE class and the edges or connections represented a correlation between the two nodes. We used the Markov clustering (MCL) algorithm (Enright et al., 2002) to reveal natural groups within the graphs using only nodes that had a correlation of 0.96 or better to another node. The algorithm simulates flow within a graph and promotes flow in a highly connected region and demotes less connected regions. The MCL algorithm takes one parameter, the inflation parameter, which adjusts the granularity of the clusters. We tested various inflation parameters between two to ten, and used four as this was a good compromise between the number of clusters and cluster sizes. The graphs were visualised using the Cytoscape software (Shannon et al., 2003).

2.4 FANTOM5 sample ontology enrichment analysis

Structured sample ontologies were used to annotate the FANTOM5 libraries, allowing the identification of enriched biological properties based on CAGE expression profiles (Forrest et al., 2014). Specifically, samples were annotated using the The Open Biological and Biomedical Ontologies (`http://www.obofoundry.org/`) and the structured ontologies, were grouped into hierarchical cellular, anatomical, disease and experimental ontologies. Each ontology can be used to separate libraries in a binary manner, where a library either has membership (x) or no membership (y) to a particular ontology. To test for ontology enrichment, the TPM expression of libraries in x were compared to the expression of libraries in y using a Mann-Whitney-Wilcoxon test; this was performed on all 845 sample ontologies and for all all repeat classes (1,087). The p-values were adjusted following the Benjamini & Hochberg method (Benjamini and Hochberg, 1995) and resulted in a 1087×845 matrix of adjusted p-values. Each element of this matrix corresponds to a p-value indicating whether the expression profile of a particular repeat class enriches a particular sample ontology.

2.5 Parametric tag clustering of CAGE reads

The FANTOM5 CAGE libraries were previously clustered using a decomposition-based peak identification (DPI) method that utilised a stringent mapping quality threshold of 20 (Forrest et al., 2014). This criteria removes signal arising from REs and is inappropriate for studying the expression of REs. We used a tag clustering method known as parametric clustering (Frith et al., 2008), which uses maximal scoring segments to clusters reads. Specifically for every maximal scoring segments, the minimum and maximum values of the density parameter, d, are reported and used to assess whether a cluster is robust and not formed due to random fluctuations in the data set. We kept tag clusters

with at least ten raw tags, a maximum density / minimum density ratio of at least two, and limited tag clusters to a length of 200 bps. We performed tag clustering using reads thresholded at various mapping qualities $(0, \ldots, 10)$ and to simplify the large number of tag clusters, we took the largest tag cluster that could encompass all other tag clusters, which we called non-overlapping tag clusters.

2.6 Tag cluster annotation

We used the intersectBed tool from the BEDTools suite to annotate tag clusters to GENCODE (v19) transcripts (Harrow et al., 2012), REs, FANTOM5 permissive enhancers (Andersson et al., 2014), FANTOM5 small RNAs, and long intergenic non-coding RNAs (Cabili et al., 2011). We separated the genome into 5 separate classes based on the GENCODE annotations and annotated each tag cluster hierarchically in the order: promoter, exon, intron, repetitive elements, and intergenic region. Genomic regions +/- 200 bp around the starting site of a GENCODE transcript was considered the promoter region of that transcript. Exonic regions were defined as regions overlapping the exons of GENCODE transcripts. Intronic regions were defined as the region remaining from the subtraction (using substractBed) between a GENCODE gene model and the exonic regions. Intergenic regions were defined as the remaining region from the subtraction between gene models and the genome sequence. FANTOM5 small RNA libraries were clustered in the same manner as the CAGE data (see section 2.5).

We used the GenometriCorr package (Favorov et al., 2012) to calculate potential correlations between two sets of genomic features: a query and a reference set. Relative and absolute distances between query and reference features are tested against a uniform distribution of distances. The significance of overlap between two sets of features is tested using a projection test, using a binomial test, where the probability of overlap is based on the coverage of reference features and by using the Jaccard index defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

which measures the amount of overlap between two features. To test where observed intersections are statistically significant, a null distribution was created by measuring the Jaccard index of 1,000 permutations of the query features. A list of random coordinates was generated by using randomBed from the BEDTools suite running the following command `bedtools random -g hg19.genome -l 300 -n 100000 -seed 31`, where hg19.genome is a file containing the start and end coordinates of the assembled chromosomes (n=25) on hg19.

2.7 Measuring expression specificity

The Shannon entropy has been previously used as a metric to characterise the overall expression specificity of a gene amongst a panel of samples (Schug et al.,

2005). We used the Shannon entropy as a measure for expression specificity and calculated the metric as follows:

$$-\sum_{i=1}^{n} x_i \cdot log_2 x_i$$

where i is the library index and x_i the expression of a feature in a particular library. In addition, we normalised the expression of a feature in a single library by the total expression of the features in all libraries. The Shannon entropy, measured in bits, ranges from zero for transcripts expressed only in a single sample to $log_2 n$ for features that are expressed uniformly across all n samples.

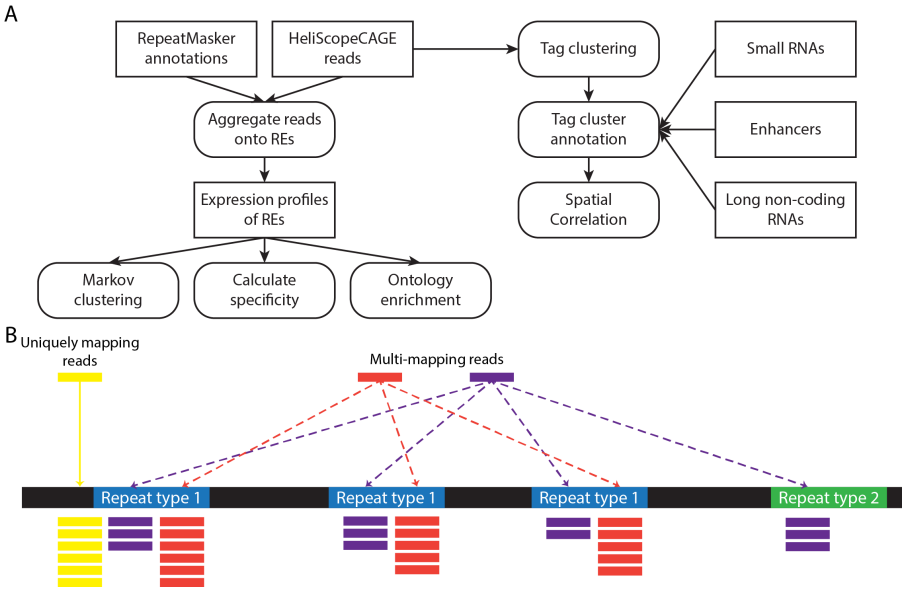

Figure 1: Summary of the methods. A) The pipeline of this study; rectangles show data sources and rounded rectangles show bioinformatic procedures. B) The aggregation strategy used to assign reads to repeat types. In the example above, 3 scenario are shown: the first scenario shows the unambiguous assignment of reads to a repeat type, the second scenario shows how multi-mapping reads are all randomly assigned to the same repeat type, and the third scenario shows how multi-mapping reads can be randomly assigned to different repeat types, which can occur in repeats belonging to a similar class.

3 Results

3.1 Annotating repetitive elements

Repetitive elements (REs) make up a large fraction of the human genome, with estimates ranging from 50% using sequence homology based approaches to up to

two thirds based on *de novo* detection methods (de Koning et al., 2011). However, *de novo* detection methods are annotation agnostic, thus these approaches do not classify the repeats into classes and types. Accurate annotations of REs enables much more precise biological conclusions to be made. For this work, we used a recently developed database of RE called Dfam (Wheeler et al., 2013), which represents each element by a profile hidden Markov model (HMM), built from alignments generated using RepeatMasker (Smit et al., 2004) and Repbase Update (Jurka et al., 2005). Each profile-HMM contains much richer information than the consensus sequence built from multiple alignments of REs, such as those present in Repbase Update. Using the profile-HMMs with a HMM search tool called nhmmer (Wheeler and Eddy, 2013), resulted in an increase in annotations (Wheeler et al., 2013). We compared RE annotations performed using consensus sequences to those based on profile-HMMs and found an 5.4% increase in annotations using profile-HMMs: 46.8% (1,448,043,873 / 3,095,693,983) versus 52.2% (1,614,955,505 / 3,095,693,983) of the human genome was annotated as repetitive using consensus sequences and profile-HMMs, respectively.

3.2 The robustness of aggregating reads

We quantified the expression of REs, in two independent ways: by aggregating reads onto REs and by tag clustering and intersection (Figure 1). In the first approach, CAGE reads were tallied across 1,087 RE classes (Figure 1B) to measure the overall expression strength of REs. This method relies solely on the overlap of a mapped read to a RE, which has its advantages and disadvantages. Positional information of a read with respect to a RE is lost and we cannot distinguish whether a RE is strongly expressed at one locus or weakly expressed at 100 loci. However, this method provides a bird's-eye view of REs across a large panel of samples and is robust to multi-mapping reads. To demonstrate the robustness of this method, we performed the aggregation step using reads filtered at various mapping qualities and compared the expression profiles (Figure 2). The Spearman correlations between the expression profiles were very high, with the lowest correlation (Spearman's rho = 0.83) between the expression profiles prepared using all reads against using reads thresholded with a mapping quality of 10 or better. The results suggest that multi-mapping reads were mapped back to the same RE class.

3.3 The expression specificity of repetitive elements

To measure the specificity of expression, we calculated the Shannon entropy of all REs using their expression profile across the FANTOM5 libraries. The distribution of Shannon entropies show that on an aggregated level, most REs are expressed in roughly the same amount across the libraries (Figure 3). This may be due to summing all genome-wide expression events of a RE class into one value; though it is interesting to note that on average most REs are expressed equally. However, even on this level, there are various classes of REs that are enriched in particular libraries. For example, the expression of LTR7 was

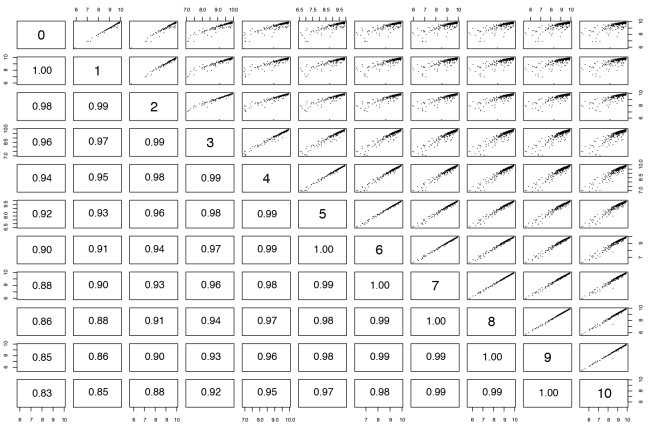

Figure 2: Correlation of repetitive element expression profiles aggregated using reads at different mapping qualities.

enriched in pluripotent and embryonic stem cells, the expression of MER74C was enriched in blood samples, the expression of an ultra-conserved element UCON15 is enriched in brain samples, and the expression of MER41E was enriched in placental samples.

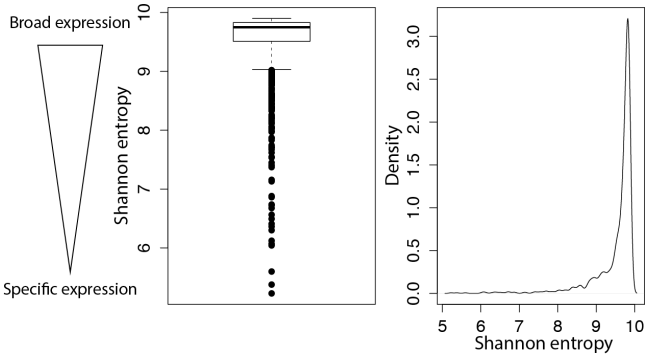

Figure 3: The distribution of Shannon entropies reveals that on an aggregated level, most repetitive elements are expressed at similar levels. However, a small subset shows a much more restrictive expression profile.

Motivated by the observation that some REs were expressed in a tissue-specific manner, we clustered the FANTOM5 libraries using correlations of aggregated RE expression profiles. We represented the correlations as a graph, whereby each node is a single library and nodes were connected if the two libraries had RE expression profiles that were correlated (Figure 4A). The graph showed that most libraries had a very similar RE expression profile. Next, we performed Markov clustering (MCL), an unsupervised clustering algorithm based on the simulation of stochastic flow in graphs, to reveal any natural groups within this graph. Impressively, the clusters formed by the MCL algorithm

grouped together libraries that were technically or biological related together (Figure 4B). For example, induced pluripotent stem cells clustered with human embryonic stem cells and embryoid bodies and whole blood samples taken from different donors clustered together (Table 1).

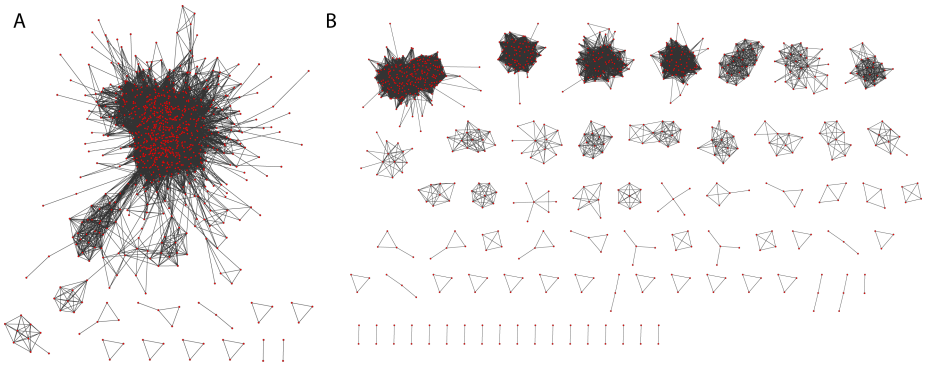

Figure 4: Representing the expression patterns of REs as graphs. A) Each red node represents a FANTOM5 library and each edge represents a Spearman's correlation of 0.96 between two nodes B) Markov clustering revealed natural groups present in the correlation graph.

Markov cluster 1	**Markov cluster 2**
blood_adult_pool1	iPS_to_neuron_control_donor_day00_rep1
Whole_blood_ribopure_donor090612_1	iPS_to_neuron_control_donor_day00_rep2
Whole_blood_ribopure_donor090612_2	iPS_to_neuron_control_donor_day00_rep2
Whole_blood_ribopure_donor090612_3	H9_Embryoid_body_cells_rep1_H9EB_1_d0
Whole_blood_ribopure_donor090325_1	H9_Embryoid_body_cells_rep2_H9EB_2_d0
Whole_blood_ribopure_donor090325_2	H9_Embryoid_body_cells_rep3_H9EB_3_d0
Whole_blood_ribopure_donor090309_1	H9_Embryonic_Stem_cells_rep2_H9ES_2
Whole_blood_ribopure_donor090309_2	H9_Embryonic_Stem_cells_rep3_H9ES_3
Whole_blood_ribopure_donor090309_3	

Table 1: Two examples of the Markov clusters formed (out of ninety) using the correlation values calculated on the aggregated RE expression profiles. The samples names were slightly shorten from the full name used in the FANTOM5 project.

3.4 Sample ontology enrichment analysis

The FANTOM5 libraries have been associated to specific sample ontologies (Forrest et al., 2014), which allowed us to test whether the expression of specific REs enriched specific sample ontologies. Based on a particular ontology, libraries could be separated into two groups: those that are associated with the ontology and those that are not. Based on this separation, we tested whether the RE expression between the two groups was statistically different. We performed 919,602 enrichment tests and found 69,038 associations that were statistically significant (adjusted p-value <0.05) (Figure 5A). Many of the 845 sample ontologies are subtly different but typically associated to the same libraries, thus a RE may enrich several similar ontologies. For example, LTR7 had 80 sample ontologies that were significantly enriched including "embryonic stem cell"

and "H9 embryonic stem cell line" (adjusted p-value for both $\sim$ 0.00693) (Figure 5B). MER74C had 59 ontologies that were significantly enriched including "blood" (adjusted p-value $\sim$ 0.001952) and "whole blood" (adjusted p-value $\sim$ 0.00012) (Figure 5C). UCON15 had 37 sample ontologies significantly enriched, including "regional part of brain" (adjusted p-value $\sim$ 0.000169) and "nucleus of brain" (adjusted p-value $\sim$ 0.016551). We can also lookup specific REs that are enriched to specific sample ontologies (Table 2).

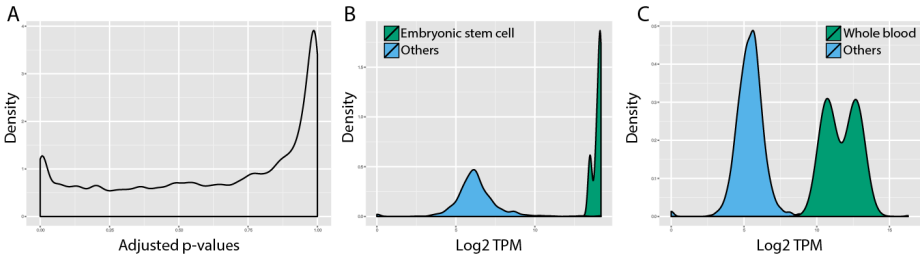

Figure 5: Sample ontology enrichment analysis. A) The distribution of adjusted p-values for the sample ontology enrichment analysis. B) The expression distribution, in tags per million, of LTR7 in samples annotated as "Embryonic stem cell" and those that are not, i.e. others. C) The expression distribution, in tags per million, of MER74C in samples annotated as "Whole blood" and those that are not, i.e. others.

HERVS71-int	LTR28C	HERV4_I-int	MER110A	REP522
0.010	0.010	0.010	0.010	0.010
LTR70	MER4-int	MER67C	HERVH48-int	MSTB1
0.010	0.009	0.008	0.008	0.008
MER61-int	MER4CL34	LTR8A	LTR10D	MER51C
0.008	0.008	0.008	0.007	0.007
LTR1A2	MER4A1_v	MER41A	L1M7	L1M2b
0.007	0.007	0.007	0.007	0.007
LTR10B1	LTR17	LTR33A	LTR22C2	MER48
0.006	0.006	0.006	0.006	0.006
L1M2	MER11C	MER4A1	LTR7B	LTR6A
0.006	0.006	0.006	0.006	0.006
LTR29	LTR5	LTR7Y	LTR6B	Charlie21a
0.006	0.006	0.006	0.006	0.006
LTR27B	LTR48B	LTR7A	LTR49	MER4C
0.006	0.006	0.006	0.006	0.006
MER61D	LTR7	LTR5_Hs	MER61B	LTR9D
0.006	0.006	0.006	0.006	0.006
MER21B	LTR9B	MER61E	LTR26B	LTR7C
0.006	0.006	0.006	0.006	0.006
LTR49-int	MER74A	MER61C	HUERS-P1-int	LTR9C
0.006	0.006	0.006	0.006	0.006
HERVH-int	LTR16D2	MER61A		
0.006	0.006	0.006		

Table 2: Repetitive elements that are enriched in samples with the ontology "embryonic stem cells"; the adjusted p-values are sorted by the level of significance in descending order.

3.5 Repetitive elements overlap small RNAs, enhancers, and long non-coding RNAs

In the second approach of quantifying the expression of REs, we clustered all mapped FANTOM5 CAGE reads at various mapping qualities, using a parametric clustering methods (Frith et al., 2008). This tag clustering approach, as opposed to the aggregation method, allows REs to be put into context of other

genomic features. We first annotated tag clusters to GENCODE transcripts in a hierarchical manner to avoid confounding signal from known transcripts. The expression level of annotated GENCODE promoters is clearly higher much higher than any other genomic feature (Table 3). The median expression of intronic, RE, and intergenic regions are similar; however the mean expression of RE regions is higher, suggesting that a population of REs that are much higher expressed than transcripts arising within intronic and intergenic regions (Table 3).

Genomic feature	**Tag cluster count**	**Mean expression**	**Median expression**
Promoter	390162	9914	52
Exonic	937236	391	33
Intronic	2219891	87	20
Repetitive	410313	210	20
Intergenic	421534	126	21

Table 3: A summary of the number and raw expression levels of tag clusters with respect to their GENCODE annotations.

Next we performed statistical tests to determine whether these REs overlapped different various genomic features more often than by chance. The p-values of the projection and Jaccard index test are highly significant indicating that expressed REs overlap small RNAs, enhancer regions, and lncRNAs more often than by chance (Table 4). A random set of genomic features did not significantly overlap expressed REs.

	sRNAs	**Enhancers**	**lncRNAs**	**Random**
Number of genomic features	610,246	43,011	14,281	100,000
Projection test p-value	0	0	0.00264	0.2246452
Jaccard measure p-value	<0.001	<0.001	<0.001	0.391

Table 4: P-values for the projection and Jaccard index statistical tests for assessing the significance of overlap between RE tag clusters and various genomic features. P-values were calculated using 1,000 permutation tests in each case.

4 Discussion

One of the first genome-wide screenings of transcription events initiating within repetitive elements (REs) was based on roughly 42 million CAGE reads from across 80 human samples, which were clustered into 12 separate tissue types (Faulkner et al., 2009). This initial study revealed that retrotransposon-derived transcription events are generally tissue-specific and function as alternative promoters (Faulkner et al., 2009). This observation was supported by the fact that the human L1 retrotransposon contain internal promoter sequences, including an antisense promoter that can drive the expression of various human genes, even in normal cells (Nigumann et al., 2002; Speek, 2001). Furthermore, these L1 retrotransposon promoters were found to drive tissue-specific transcription of

genes (Matlik et al., 2006). In addition to L1, the long terminal repeats (LTRs) of endogenous retroviruses also have the capacity for driving the expression of human genes, especially in a tissue-specific manner (Cohen et al., 2009). In another large scale study examining the transcriptional landscape in 15 cell lines, it was demonstrated that repetitive elements exhibit cell type specific expression patterns and are enriched in the nucleus (Djebali et al., 2012). Deeply profiled nuclear and cytoplasmic transcriptomes of stem cells also revealed an enrichment of transcripts initiating from REs (Fort et al., 2014).

In this work, we examined the expression pattern of REs in 988 human samples, which included technical and biological replicates, using CAGE technology adopted onto a single molecule sequencing platform (Kanamori-Katayama et al., 2011). This expression atlas has been previously used to systematically study the cohort of genes (Forrest et al., 2014) and enhancers (Andersson et al., 2014) that are used in specific cell types. We leveraged the use of profile hidden Markov models (HMMs) as a replacement over the use of consensus sequences to more accurate RE annotations due to the richer information content contained within a profile (Wheeler et al., 2013). The use of profile-HMMs resulted in a 5.4% increase in annotations, allowing us to ascribe more transcription events to REs. Expression pattern of REs were studied in two independent manners: by aggregating reads onto REs and by tag clustering and annotation. The first approach aimed to quantify the expression of REs on the scale of samples to mitigate multi-mapping issues and to gain an overview of the expression of REs across samples. The second approach puts the expression of REs into the context of other genomic features to identify potential correlations between the expression of REs to nearby genomic features.

To measure the expression specificity of REs, we calculated the Shannon entropy using the expression profiles of REs. On an aggregated level, most REs were broadly expressed with a Shannon entropy near maximum ($log_2(988)$); however, a subset of REs were more specifically expressed. Of the 20 REs with the lowest entropies, 15 corresponded to the long terminal repeats (LTRs), which are the control centres of gene expression for endogenous retroviruses, since they contain all the requisite signals for gene expression. This observation supports the notion that LTRs have become exapted as alternative promoters for driving tissue-specific expression patterns (Cohen et al., 2009). On closer examination of these 15 LTR sequences, we observed that they had enriched expression patterns in a specific cohort of biological samples. For example, the expression of LTR7 is enriched in human embryonic stem cells, induced pluripotent stem cells (iPCSs), and germ line cancers compared to the other samples. LTR7, which is associated with human endogenous retrovirus H (HERVH), has been observed to be enriched in long intergenic non-coding RNAs that are expressed at much higher levels in stem-like cells (Kelley and Rinn, 2012). One potential mechanism by which LTR7 confers specificity to stem cells is the observation that the master transcriptional regulators of pluripotency NANOG, OCT4, and SOX2, were found to bind to LTR7 elements (Loewer et al., 2010). Furthermore, over-expression of LTR7 has been shown to lead to defective human induced pluripotent stem cell clones, suggesting that careful regulation of

LTR7 is required for maintaining pluripotency (Koyanagi-Aoi et al., 2013).

There are other examples of transcription factors (TFs) associating to specific RE sequences. One study discovered the pervasive association of several TFs to sequences in distinctive families of transposable elements (TEs) (Bourque et al., 2008). Since REs harbour sequences that resemble transcription factor binding sites (TFBSs), this may be one mechanism by which REs can confer tissue specificity. Indeed, we demonstrated that the FANTOM5 libraries could cluster biologically and technically similar libraries together, using just the aggregated expression signal from REs. Furthermore, in support of specific RE associating with specific cell types, our sample ontology enrichment analysis, showed that the expression pattern of many REs enriched particular sample ontologies. The enrichment analysis showed that LTR7 is enriched in "embryonic stem cell" samples and while this was observed prior to the enrichment analysis, the ontologies provide a structured manner to associate REs to samples. Furthermore, this analysis can be used to identify REs enriched in specific sample ontologies. For example, many other LTR families are enriched in "embryonic stem cell" samples and one of the most statistically significant LTRs, MER61A, has previously been reported to provide binding sites for p53 (Wang et al., 2007), a tumour suppressor gene that plays an important role in stem cells (Solozobova and Blattner, 2011). This large list of associations between samples and REs is a useful resource for generating hypotheses on the potential role of REs.

In addition to the role of REs in regulatory networks (Feschotte, 2008), REs have been found to be associated with long non-coding RNAs (lncRNAs) (Kelley and Rinn, 2012; Kapusta et al., 2013) and have been proposed to have played a major role in the origin and evolution of lncRNA (Kapusta and Feschotte, 2014). Furthermore, TEs are positioned and orientated in a biased manner at the transcription start site of lncRNAs, suggesting that they may play a role in regulating the expression of the lincRNA (Kelley and Rinn, 2012). In addition to lncRNAs, small regulatory RNAs, such as Piwi-interacting RNAs (piRNAs) and small interfering RNAs (siRNAs) are produced from the sequences of TEs (Cowley and Oakey, 2013). While these small RNAs, primarily function as a host defence mechanism to silence TEs, epigenetically and developmentally regulated bursts in expression of TEs resulted in the production of piRNAs that regulated the expression of messenger RNA (mRNA) by binding to the 3' UTR (McCue and Slotkin, 2012). The enrichment of TE sequences within the 3' UTR sequences of mRNAs (Faulkner et al., 2009) may suggest that this phenomenon may be more widespread. In addition to generating non-coding RNA, a recent study examining the methylation patterns of TEs, demonstrated that methylation patterns of TEs was tissue specific, and may act as enhancers (Xie et al., 2013). Another study found evidence of TEs acting as enhancer regions in stem cells (Fort et al., 2014). By examining the genomic locations of these three genomic features (lncRNAs, enhancers, and small RNAs) to expressed REs, we showed that they overlapped more often than by chance. Additional work needs to be carried out to better characterise the overlap between these features with the expressed REs.

It is becoming evident that TEs have had a major impact in the evolution of regulatory networks. Our work here, takes advantage of the large breath of samples provided by the FANTOM5 project, to illustrate the tissue-specific manner of REs. Given that the technical difficulties with analysing REs are now becoming less of a problem due to longer read lengths in high-throughput sequencers and various computational pipelines have been developed to analyse signal from REs, we may see additional studies in this exciting area of research.

5 Data deposition

All CAGE data has been deposited at DDBJ DRA under accession number DRA000991.

6 Acknowledgements

We would like to thank Dr. Alex Fort for reading an early version of this manuscript and for making suggestions.

7 Funding

FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to YH and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH. DT is supported by the European Union Seventh Framework Programme under grant agreement FP7-People-ITN-2008-238055 ("BrainTrain" project) to PC.

8 Authors' contributions

DT processed the data, designed the analyses, performed the analyses, interpreted the data, and wrote the manuscript. PC oversaw the project. All authors read and approved the final manuscript.

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J. L., Ruan, Y., Wei, C. L., Ng, H. H., and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, 18(11):1752–1762.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, 25(18):1915–1927.

Callinan, P. A. and Batzer, M. A. (2006). Retrotransposable elements and human disease. *Genome Dyn*, 1:104–115.

Cohen, C. J., Lock, W. M., and Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, 448(2):105–114.

Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, 10(10):691–703.

Cowley, M. and Oakey, R. J. (2013). Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet.*, 9(1):e1003234.

Day, D. S., Luquette, L. J., Park, P. J., and Kharchenko, P. V. (2010). Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.*, 11(6):R69.

de Koning, A. P., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.*, 7(12):e1002384.

de Souza, F. S., Franchini, L. F., and Rubinstein, M. (2013). Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol. Biol. Evol.*, 30(6):1239–1251.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., et al. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584.

Faulkner, G. J., Forrest, A. R., Chalk, A. M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D. A., and Grimmond, S. M. (2008). A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 91(3):281–288.

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, 41(5):563–571.

Favorov, A., Mularoni, L., Cope, L. M., Medvedeva, Y., Mironov, A. A., Makeev, V. J., and Wheelan, S. J. (2012). Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput. Biol.*, 8(5):e1002529.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, 9(5):397–405.

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C. A., et al. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.*, 46(6):558–566.

Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. *Genome Res.*, 18(1):1–12.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477(7364):295–300.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467.

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., et al. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.*, 21(7):1150–1159.

Kapusta, A. and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.*, 30(10):439–452.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.*, 9(4):e1003470.

Kelley, D. and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, 13(11):R107.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106(28):11667–11672.

Koyanagi-Aoi, M., Ohnuki, M., Takahashi, K., Okita, K., Noma, H., Sawamura, Y., Teramoto, I., Narita, M., Sato, Y., Ichisaka, T., Amano, N., Watanabe, A., Morizane, A., Yamada, Y., Sato, T., Takahashi, J., and Yamanaka, S. (2013). Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 110(51):20569–20574.

Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., Ng, H. H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, 42(7):631–634.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858.

Loewer, S., Cabili, M. N., Guttman, M., Loh, Y. H., Thomas, K., Park, I. H., Garber, M., Curran, M., Onder, T., Agarwal, S., Manos, P. D., Datta, S., Lander, E. S., Schlaeger, T. M., Daley, G. Q., and Rinn, J. L. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.*, 42(12):1113–1117.

Matlik, K., Redik, K., and Speek, M. (2006). L1 antisense promoter drives tissue-specific transcription of human genes. *J. Biomed. Biotechnol.*, 2006(1):71753.

McCue, A. D. and Slotkin, R. K. (2012). Transposable element small RNAs as regulators of gene expression. *Trends Genet.*, 28(12):616–623.

Mills, R. E., Bennett, E. A., Iskow, R. C., and Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends Genet.*, 23(4):183–191.

Nigumann, P., Redik, K., Matlik, K., and Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*, 79(5):628–634.

Orgel, L. E. and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Schug, J., Schuller, W. P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. J. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, 6(4):R33.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504.

Smit, A. F. A., Hubley, R., and Green, P. (1996-2004). RepeatMasker Open-3.0.

Solozobova, V. and Blattner, C. (2011). p53 in stem cells. *World J Biol Chem*, 2(9):202–214.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.*, 21(6):1973–1985.

Tange, O. (2011). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, 13(1):36–46.

Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K., and Haussler, D. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U.S.A.*, 104(47):18613–18618.

Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., Smit, A. F., and Finn, R. D. (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, 41(Database issue):70–82.

Wheeler, T. J. and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19):2487–2489.

Whiteford, N., Haslam, N., Weber, G., Prugel-Bennett, A., Essex, J. W., Roach, P. L., Bradley, M., and Neylon, C. (2005). An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.*, 33(19):e171.

Xie, M., Hong, C., Zhang, B., Lowdon, R. F., Xing, X., et al. (2013). DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.*, 45(7):836–841.

Yang, N. and Kazazian, H. H. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.*, 13(9):763–771.