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 10, 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 pervasively transcribed, 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 to families of TEs to specific ontological categories 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)[1]. The estimated number of active TEs in the human genome is less than 0.05%[2], which is as expected given that TEs are a major factor in causing mutations that lead to human disease[3]. However, despite the detrimental effects of TEs, their ability to move around has shaped the evolution of the genomes in which they reside[4]. For example, TEs were found to contribute transcription factor binding sites (TFBSs) to various mammalian transcription factors[5] and it has even been suggested that many TFBSs had originated from TEs[6]. Furthermore, in a study that examined the methylation status of TEs, it was demonstrated that methylation patterns in TEs were tissue-specific and displayed enhancer hallmarks[7]. A separate study examining transcription initiation events arising within TEs, observed that TEs were used as alternative promoters and are expressed in a tissue-specific manner[8], corroborating with tissue-specific methylation patterns. These studies imply that TEs are developmentally regulated and may contribute to driving the large diversity of cell types.

Numerous RNA sequencing studies have revealed a large set of long non-coding RNAs (lncRNAs) in the human genome[9, 10, 11]. In a study examining TEs and lncRNAs, it was shown that lncRNAs are enriched with TE sequences, leading to the hypothesis that TEs have contributed to the origin of many lncRNAs[12]. 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[13] and the transcription start site (TSS) of the lncRNAs *BANCR*, *lnc-RoR*, and *lncRNA-ES3* all overlap the LTR sequence of separate endogenous retroviruses[12]. 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[**?**]. Furthermore, a specific TE family was found to drive the expression of stem cell specific lncRNAs[14], suggesting that TEs are able confer tissue-specificity possibly by associating with specific regulatory signals.

However, 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[15] and provides little to no selective advantage to organisms[16]. 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[17]. However, in their original paper, Orgel and Crick suggested at the hypothesis that these selfish DNA sequences may become exapted for control purposes[16]. Furthermore, the technical difficulties associated with REs have limited the number of genome-wide studies investigating their potential roles. In 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[18]. 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[19]. The third choice can also be incorporated with a strategy that uses uniquely mapped reads to probabilistically weight multi-mapping reads[20]. The basic premise is that a region that is transcriptional active is more likely to have given rise to a read than a transcriptional inert region. Other strategies include mapping to a consensus database of REs, such as RepBase Update[21], or mapping to regions of a genome that has been annotated as being repetitive[7], or combining reference genome mapping with consensus sequence mapping[22]. 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 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 ontologies, suggesting that specific families of REs may be involved in specific functions. 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[23] 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[21]. Recently a database of REs based on profile hidden Markov models was developed, called Dfam[24], which allowed screening of REs using a hidden Markov model search tool, called nhmmer[25]. It has been reported that screening for REs using nhmmer and Dfam is more sensitive and specific than consensus sequence based approaches[24]. 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[26]. 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 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[27].

To aggregate CAGE reads to REs, the coordinates of the mapped reads were intersected with the RE coordinates using intersectBed from the BEDTools suite[28]; parallelisation of the computations was achieved using GNU parallel[29]. 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 calcu-

lations) 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[30] 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 [31].

2.4 FANTOM5 sample ontology enrichment analysis

Structured ontologies were developed and used to annotate the FANTOM5 libraries, allowing the identification of enriched biological properties based on CAGE expression profiles[32]. 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 ontological term 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 846 ontologies and for all all repeat classes (1,087). The p-values were adjusted following the Benjamini & Hochberg method[33] and resulted in a 1087×846 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 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[32]. 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[34], 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[35], REs, FANTOM5 permissive enhancers[36], FANTOM5 small RNAs, and long intergenic non-coding RNAs[9]. 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[37] 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.

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[38]. 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.

In addition, we used the h-index[39] as a measure of expression ubiquity, which is defined as:

$$H(x) = max\{i = 1, \ldots, n : x_i \geq i\}$$

where the h-index, $H(x)$, is the maximum value in the set of sorted expression values, i, such that the expression of the x_i library is greater than or equal to i. For example, a tag cluster with a h-index of one is supported by one CAGE read in at least one library. A tag cluster with a h-index of two is supported by at least two CAGE reads in at least two libraries, and so on.

All code used for this work is available at `https://github.com/davetang/fantom5_repeat`.

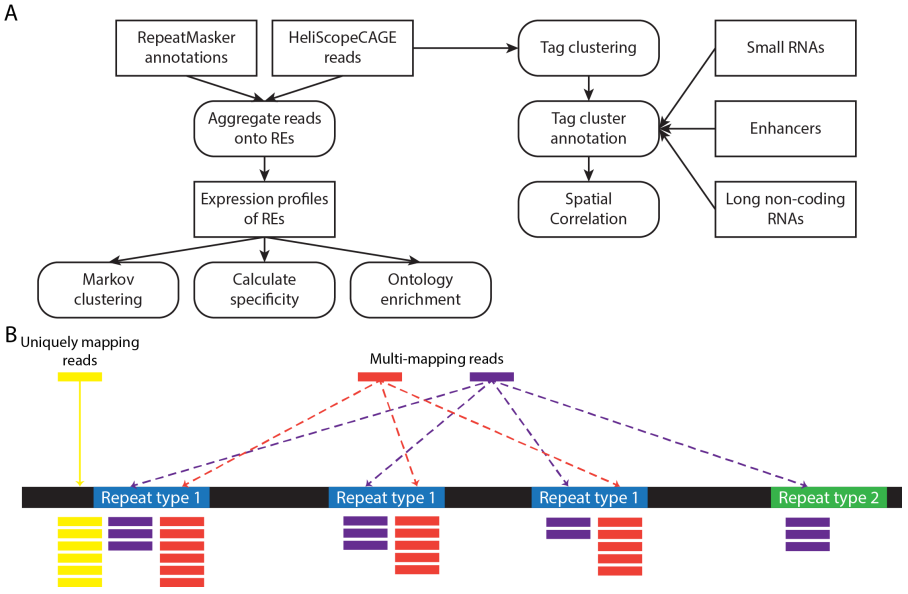

Figure 1: Summary of the methods. A) The pipeline of this study; rectangles show data sources and rounded rectangles show bioinformatic procedures. B) The aggregation strategy used to assign reads to repeat types. In the example above, 3 scenario are shown: the first scenario shows the unambiguous assignment 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 Shannon entropy

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 approach of aggregation is robust against multi-mapping reads, which are likely to multi-map to the same RE class; we compared the expression matrices tallied with reads at different mapping qualities and they showed very high correlations. The lowest correlation (Spearman's rho = 0.83) was between the expression matrices prepared using all reads against using reads thresholded with a mapping quality of 10 or better. Next, we used the Shannon entropy to measure the specificity of RE expression and this revealed several RE classes that had a much more restrictive expression pattern across the FANTOM5 libraries (Figure 2). Expression of LTR7 was restricted to pluripotent and embryonic stem cells, which has been previously reported[14]. Expression of MER74C was found to be restricted to blood samples and MER41E expression was restricted to placental samples. In order to associate the expression of REs to biological functions, we performed a sample ontology enrichment analysis.

Figure 2: The 20 repetitive element classes with the lowest Shannon entropies; the x-axis shows each individual FANTOM5 library and the y-axis shows the tags per million expression.

3.2 Sample ontology enrichment analysis

FANTOM5 libraries have been associated to particular ontologies[32], which allowed libraries to be separated into two groups for each ontology: 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, i.e. sample ontology enrichment. This resulted in 919,602 tests (Figure 3), of which 69,038 associations were statistically significant (adjusted p-value <0.05).

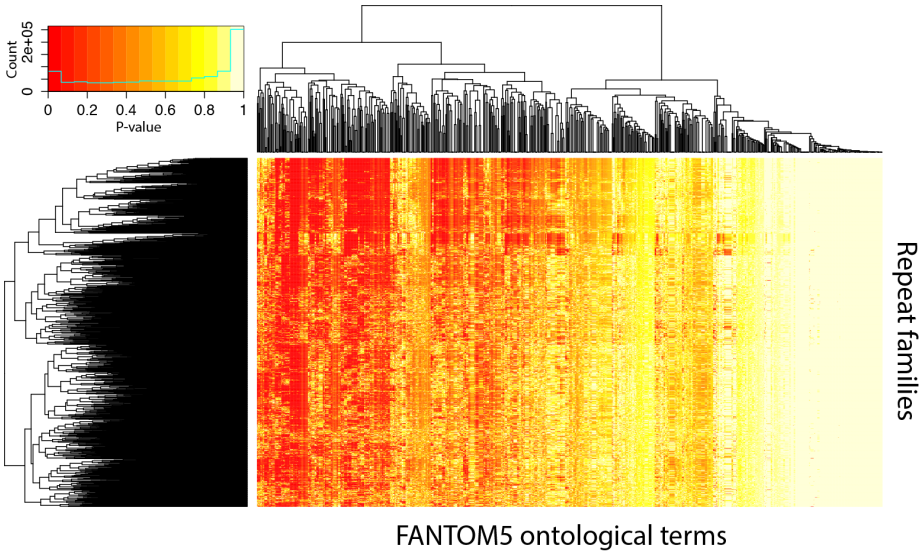

Figure 3: Heatmap of p-values from the sample ontology enrichment analysis; rows represents particular repeat families and columns represent FANTOM5 ontologies.

LTR7 had 80 sample ontologies that were significantly enriched including "embryonic stem cell" and "H9 embryonic stem cell line" (adjusted p-value for both ˜0.00693). MER74C had 59 ontologies that were significantly enriched including "blood" (adjusted p-value ˜0.001952) and "whole blood" (adjusted p-value ˜0.00012). The sample ontology "cancer", contained the most number of enriched REs with 656 out of a total of 1087, which suggests that REs are over-expressed in cancerous samples.

3.3 Markov clustering

We constructed a network of FANTOM5 libraries based on the aggregated expression of REs to determine whether the expression patterns from RE is enough to form biological meaningful clusters. FANTOM5 libraries were connected based on the expression profile correlations of each library against every other

library (Figure 4A). The highly connected network, showed that most repeats were expressed in a similar manner across all libraries. To reveal any natural groups within this graph, we performed Markov clustering (MCL), which is an unsupervised cluster algorithm based on simulation of stochastic flow in graphs. The natural groups revealed by the MCL algorithm (Figure 4B), were FANTOM libraries that were biological related. For example, induced pluripotent stem cells clustered with human embryonic stem cells and embryoid bodies (Supplementary figure 2). This suggests that different cell or tissue types have a pronounced RE expression pattern.

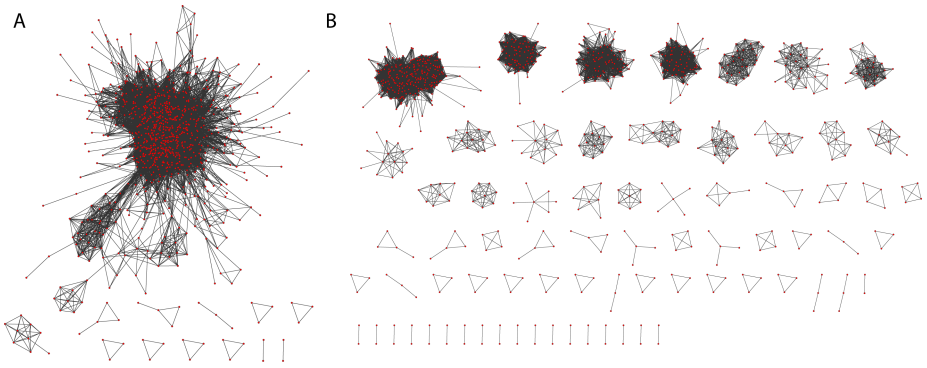

Figure 4: Representing the expression patterns of REs as graphs. A) Each red node represents a FANTOM5 library and each edge represents a Spearman's correlation of 0.96 between two nodes B) Markov clustering revealed the natural groups present in the correlation graph.

3.4 Tag clustering

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[34]. This tag clustering approach, as opposed to the aggregation method, allows REs to be put into context of other genomic features, such as known transcript models. We annotated tag clusters to GENCODE transcripts in a hierarchical manner and to avoid confounding signal; thus tag clusters annotated as RE, do not overlap known transcripts. We performed this annotation step using reads mapped at various mapping qualities to assess the impact of mapping qualities (Figure 5).

We decided to use tag clusters using reads at a mapping quality of ten or better, despite losing a large number of tag clusters.

3.5 Genomic correlations

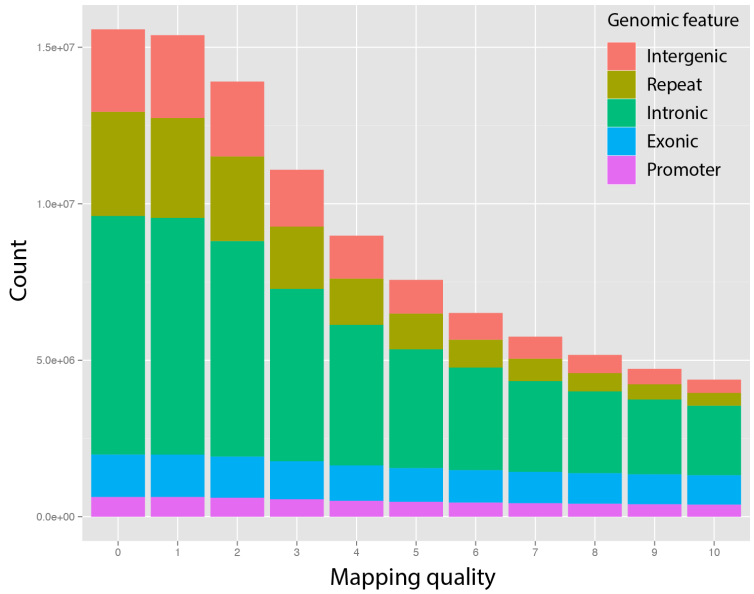

Figure 5: GENCODE annotation of tag clusters using reads at various mapping qualities.

Genomic feature	**Tag clusters**	**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 1: Tag cluster statistics using reads with a mapping quality of 10 or better.

4 Discussion

	sRNAs	**Enhancers**	**lincRNAs**
Number of genomic features	610246	43011	14281
Relative distance p-value	0	0	1.63e-14
Absolute distance p-value	<0.001	<0.001	<0.001
Projection test p-value	0	0	0.00264
Jaccard measure p-value	<0.001	<0.001	<0.001

Table 2: P-values of statistical tests carried out by the GenometriCorr package.

5 Data deposition

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

6 Funding

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

7 Authors' contributions

All authors read and approved the final manuscript.

References

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

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

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

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

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

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

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

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

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

[10] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. 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, Jul 2009.

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

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

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

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

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

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

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

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

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

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

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

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

[23] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0, 1996-2004.

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

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

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

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

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

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

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

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

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

[33] Yoav Benjamini and Yosef Hochberg. 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, 1995.

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

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

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

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

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

[39] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.*, 102(46):16569–16572, Nov 2005.