Functional analysis of repeat regions in the eukaryotic genomes

Functional analysis of repetitive DNA derived from transposable elements in the human genome

Lu Zeng

A thesis submitted for the degree of Master of Philosophy

Discipline of Genetics

School of Molecular and Biomedical Science

The University of Adelaide

July 2013

Table of Contents

Abstract

Nearly half of the human genome is made up of transposable elements (TEs). With the rapid progress of sequencing technologies, we are now much better able to systematically analyze these TEs. We have used multiple types of omics data, including the genomic sequences, epigenetic data and transcriptomic data, to investigate the potential functions of TEs across the entire human genome. Comparative analysis revealed that a large proportion of potentially functional transposable elements were located in introns, and they were mainly associated with gene repression. Functional classification from GO enrichment showed that different functions were enriched in protein coding regions containing TEs compared to non-protein coding regions. For example, protein coding genes with Alus in non-coding regions are enriched with respect to intracellular membrane-bounded organelles, while protein coding genes with Alus in coding regions are more enriched at intracellular non-membrane-bounded organelles. Significantly, transcriptome data showed that the genes with TEs had lower expression levels compared with genes without TEs, revealing a novel aspect of the impact of TEs on the human genome. In addition, genome wide analysis of repeats with regulatory elements showed that MIR and L2 repeats were more probable to be active regulators while L1 repeats were less probable to be regulators. In conclusion, the role of TEs is significant across the genome. Repeats reduce or repress the expression of related gene, either through the proximal promoter, 5'UTR or 3'UTR or perhaps as components of lincRNA exons.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university of other tertiary institution to Lu Zeng, and to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright holder(s) of these works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed…… Date…….

Acknowledgements

I would like to express my sincere gratitude to the following people: Dr. Dave Adelson, gave me the opportunity and support to join the Master by Research program in the University of Adelaide. I am so lucky to have such a great supervisor. Dave, the knowledge that I learned from you is not just valuable for my Master, but will support me for my whole academic career.

Dr. Chaochun Wei, my supervisor in SHANGHAI JIAOTONG University, introduced me to a completely new field of bioinformatics. Over the year I stayed in Adelaide, he was always in contact with Dave and me to co-advise my research. Furthermore, he gave me this opportunity to do the Master in Adelaide.

Dan Kortschak, my co-supervisor in the University of Adelaide, helped me do research and always provide valuable suggestions.

Joy Raison, helped me generate nice figures. Sim Lim Lin, helped me sort out the tough problems I encountered in my research, and kept encouraging me. Zhipeng Qu, made both my life and research easier and fluent. Reuben Buckley, helped me practice English and providing me valuable advices. All other members of the Adelson lab, past and present, made it such a supportive and enjoyable place to work.

I would like to also specially thank SHANGHAI JIAOTONG University and The University of Adelaide, who provided me this opportunity to come here for my research.

STATEMENT OF AUTHORSHIP

Functional analysis of repeat regions in the human genome

Submitted, July 2013

Lu Zeng (Candidate)

Designed and performed experiments, analyzed results and wrote the manuscript.

I hereby certify that the statement of contribution is accurate

Signed…… Date…….

David L. Adelson & Chaochun Wei

Supervised development of work and assisted in analyzing results and writing the manuscript.

I hereby certify that the statement of contribution is accurate and I give permission for inclusion of the paper in the thesis.

Signed…… Date……

Functional analysis of repeat regions in the human genome

Lu Zeng

School of Molecular and Biomedical Science

The University of Adelaide

Adelaide, SA

Australia

School of Life Sciences and Biotechnology

SHANGHAI JIAOTONG University

Shanghai

P.R.China

Chapter 1

Functional analysis of repeat regions in the human genome

Lu Zeng and David L. Adelson and Chaochun Wei

School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, SA, Australia

School of Life Sciences and Biotechnology, SHANGHAI JIAOTONG University，
P.R.China

Introduction

1.1 Background

Eukaryotic genomes contain vast amounts of repetitive DNAs derived from TEs that contribute significantly to biological activity and genome evolution. Furthermore, TEs are mutagens; they may damage their host cells through various mechanisms [1]. For example, a transposon or a retrotransposon that inserts itself into a functional gene may disrupt or alter it, disrupting gene function. Similarly, a DNA transposon that excises from a genome may result in a deletion that cannot be repaired. Due to the presence of multiple copies of repetitive elements, such as Alu sequences, precise chromosomal pairing during meiosis may be deleted, causing unequal crossovers and deletion or insertion of genetic materials. Through these mutagenic mechanisms, repeats are known to cause a variety of human genetic disorders [2].

A number of recent studies have shown that TEs can influence host genes by providing novel promoters, splice sites or post-transcriptional modification to re-wire different developmental regulatory and transcriptional networks [3-5]. TEs tend to regulate gene expression through several mechanisms [5-7]. For example, the expression levels of protein coding genes containing repeats are significantly associated to the number of repeats in rodent genomes [6]. Moreover, TEs have been shown to influence gene expression through non-coding RNAs, resulting in the reduction or silencing of gene expression [8]. Past studies have also found that TEs have contributed to nearly half of the active regulatory elements to the human genome [9], such as altering gene promoters, creating alternative promoters and enhancers to regulate gene activity [10-12]. According to previous research, 60% of TEs in both human and mouse were located in intronic regions and all TE families in human and mouse can exonize [13], supporting the view that TEs may create new genes and exons by promoting the formation of

novel or alternative transcripts [14, 15]. The association between repeats and RNAs has also been investigated, some findings showed that tRNA can use TinT events to drive the formation of novel SINE [16]. Telomeric repeats may be transcribed as telomeric RNAs or telomeric repeat-containing RNAs [14, 15] and the insertion of TEs may also drive the evolution of lincRNAs and alter their biological functions [17].

1.2 Research questions

In order to uncover the hidden information of TEs, my research will focus on these following questions:

1) What is the distribution of TEs in functional regions in the human genome? Functional regions here include protein coding genes, ncRNAs and regulatory elements like TFBS, promoters and enhancers.

2) What is the association between repetitive elements and functional elements in the human genome?

3) Do repetitive elements impact on gene expression?

1.3 Aims and objectives

1) To analyze the distribution of repeat-associated functional elements in human genes.

2) To classify and analyze repeats in different cell lines

3) To build the consensus sequences for some important classes of repeats and use these sequences to identify full-length repeats and their function and distribution in the human genome.

4) To analyze the relationship between repeats and lincRNAs.

5) To conduct analysis of the expression level of specific repeats in the human genome.

1.4 Significance

1.4.1 Definition and classification of TEs

TEs are DNA sequences that can change their position within the genome, potentially giving rise to mutations or altering genome size and structure [18]. These characteristics of TEs can affect biological activities and thus may contribute to genome evolution. Moreover, TEs are able to insert at new locations without having a sequence relationship with the target locus. TEs make up about 50 percent of the human genome.

Transposons fall two major classes: RNA (retrotransposons) and DNA (DNA transposons), according to whether their replication is via RNA or DNA [19] intermediates. DNA transposons use a cut-and-paste transposition mechanism instead of involving the RNA intermediate.

Retrotransposons include two classes of elements, autonomous and non-autonomous. Autonomous transposons contain open reading frames (ORFs), which encode proteins essential for transposition and are thus able to autonomously transpose. Non-autonomous transposons do not encode these functions and so rely on replication machinery provided by autonomous transposons.

Retrotransposons can also be separated into two groups with respect to different characteristics: Long terminal repeat (LTR), and Non-LTR. LTR retrotransposons have transcription control sequences and open reading frames encoding retrotranspositional activities [19].They range in size from ~100bp to over 5kb. About 8% of the human genome and approximately 10% of the mouse genome are composed of LTR transposons [20].

Non-LTR retrotransposons include two sub-types; long interspersed elements (LINEs) and short interspersed elements (SINEs) respectively, both of which are widespread in eukaryotic genomes. Furthermore, LINEs are autonomous retrotransposons, while SINEs are non-autonomous retrotransposons.

LINEs [21] are genetic elements that contribute significantly to eukaryotic genomes, they are transcribed into RNA using an RNA polymerase II promoter. LINEs account for 17% of the human genome.

SINEs [21] are short DNA sequences, usually less than 500 bases long [22] originally transcribed by RNA polymerase III into tRNA, 5s ribosomal RNA and other small nuclear RNAs. The most common SINEs are Alu sequences, which account for 10.6% of the human genome.

According to previous research, SINEs and LINEs have similar nucleotide sequences at the 3'end, and SINEs usually dependent on LINE RT/EN function for transposition [21]. This finding was the starting point for the concept of LINE machinery involved in the retrotransposition of SINEs [23, 24]. Moreover, from the most recent human genome sequence, I found that there are 1500000 SINEs and 850,000 LINEs that account for 34% of the human genome in total. 70% of SINEs are Alu elements.

1.4.2 Functions of TEs

Retrotransposons can impact on human genome structure, which can dramatically affect genome evolution.

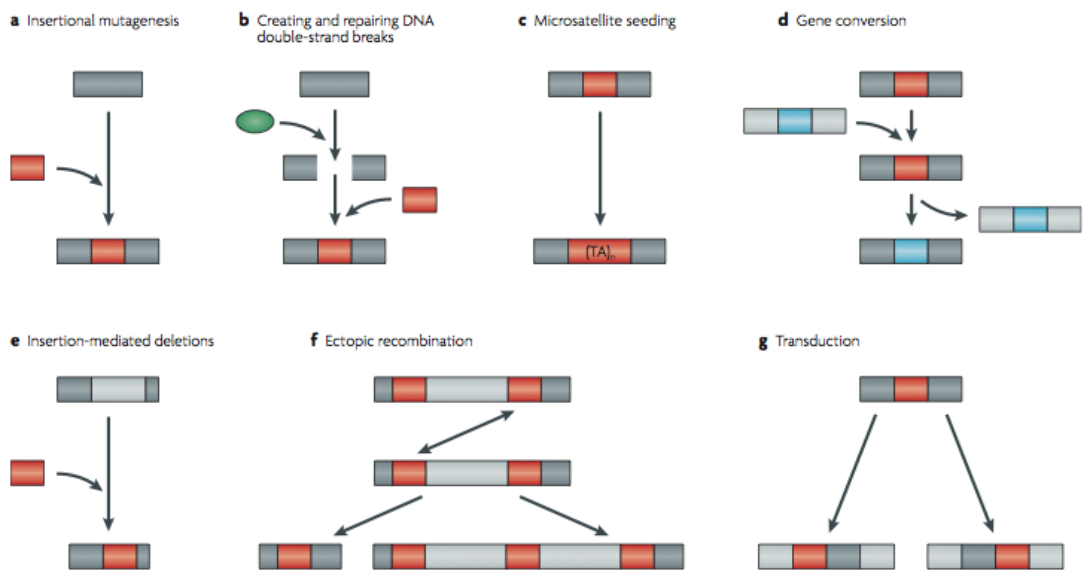


Figure1: Richard et al [25] showed that how retrotransposons have an impact on human genome structure through 7 mechanisms.

Retrotransposons can affect the human gene expression.

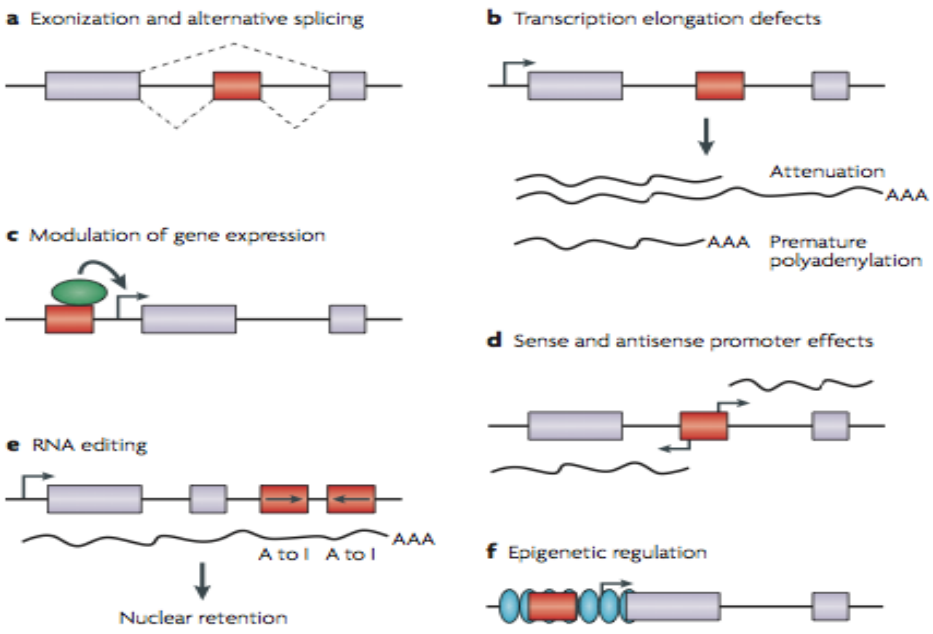


Figure2: Retrotransposons impact on human gene expression from Richard et al [25]

1.4.2.1 SINEs/Alu elements are primate-specific repeats and influence gene expression

Alu insertion is ongoing in modern human genomes, including somatic insertion events, generating genetic diversity and causing disease through insertional mutagenesis as well as causing copy number variation. Many Alu elements affect polyadenylation [26, 27], splicing [28-30], and double-stranded RNA-specific adenosine deaminase (ADAR) editing [31, 32].

1.4.2.2 LINE/L1 insertions have a high frequency of retrotransposition

L1 elements can cause human disease by inserting into human genes. After transcription to RNA, they can be reverse-transcribed into cDNA and integrated into other genomic locations. LINE/L1 has two open reading frames, ORF1 encodes a nucleic acid binding protein [33, 34], and ORF2 encodes a protein with endonuclease activity [35], reverse transcriptase activity [36] and a C-terminal cysteine-rich motif [37]. The 5'UTR of LINEs contains an internal promoter sequence, while the 3'UTR has a polyadenylation signal and a poly-A tail.

1.4.3 Association between RNAs and TEs

SINEs are derived from RNA [38], for example, Alu elements come from the ubiquitous 7SL RNA [39]. A functional sequence within Alu RNA transcripts has revealed a modular structure analogous to the organization of domains in protein transcription factors. According to recent studies, telomeric repeats may be transcribed as telomeric RNAs or telomeric repeat-containing RNAs [14, 15]. Furthermore, ncRNAs can modulate the function of transcription factors, and as far as I know, retrotransposons also contain transcription factor binding sites, which can combine with transcription factors to alter gene expression.

1.4.4 Conclusions

The recent explosion of retrotransposon studies has brought about a great improvement in understanding of TEs. It is clear that gene-regulatory networks are complicated, but this is not just the realm of genes and proteins, but also repeats. It appears that genome structure, especially for complex organisms, is very complicated as well. Genomes possess a high proportion of repeat regions, which may represent a hidden level of gene regulation. TEs impact the transcriptome through both transcriptional and post-transcriptional [5], as well as some disease-related mechanisms. Although evidence suggests that TEs are highly expressed from different regions of genomes, and possess a wide range of functionality in gene regulation, these discoveries still constitute just a glimpse of the hidden repeats.

In general, most studies of TEs are constrained to several model organisms, such as human, mouse and cow. There are few, and sometimes no studies focusing on other well-known organisms, such as chicken, pigs and so on.

2 Methods

2.1 Theoretical framework and methods

The main framework of this project was to build a pipeline to analyze the distribution, function and expression of repeats from both human genes and bovine genes.

2.1.1 The pipeline for the identification and distribution of functional repetitive elements from human genome

The identification and classification of TEs from the human and bovine genome was conducted by developing a pipeline based on free software, Perl, R and The UCSC Genome

Browser (University of California, Santa Cruz) database. Perl is a programming language that can be used for a large variety of tasks. One of the most powerful functions of Perl is for extracting information from a text file and printing out a report or for converting a text file into another form. This feature makes Perl popular in bioinformatics. In this project, Perl was used as a glue language to conduct result parsing and program linking. R is a free software programming language and a software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software [40, 41]. In this project, R was used to build graphs in order to illustrate the distribution, classification and function of TEs. The UCSC Genome Browser is an on-line genome browser [42, 43] that offers access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms. In this project, I have used this service to retrieve the repeat data and RefGene annotation data for my experiment.

2.1.1.1 Identification and distribution of functional TEs

In order to study the distribution of repetitive elements in human genes as well as the classes of various classes of TEs that exists in the human genome, I collected the datasets that applied to my experiment. First, NCBI's human Reference Gene Collection (RefSeq hg19) [44] and the associated annotation table were downloaded from the UCSC genome browser [42, 45]. In order to analyze the function of repeat regions, I have downloaded the regulatory elements data of nine human cell lines from UCSC. These regulatory element annotations, including active promoters, weak promoters, strong enhancers, weak enhancers, insulators and poly-comb repressed regions, which were derived from different chromatin states that have been marked by histone methylation, acetylation as well as histone variant H2AZ, PolIII, and CTCF [45]. I also chosen six human cell types from those nine cell lines that are useful for studying human disease, they are GM12878, HepG2, HMEC, HUVEC, K562, NHLF, Table2

shows the resource and information of these human tissues. I also have downloaded these datasets from group regulation track Broad ChromHMM and divided them into six parts according to their functional roles: active-promoter, weak-promoter, strong-enhancer, weak-enhancer, insulator and polycomb-repressed regions. Then, I retrieved the most recent human RefGene (hg19) from UCSC [46], separating it into different sections according to the human genome regions, which include 5'UTR, start codon, CDS exon, CDS intron, and stop codon and 3' UTR. Next, BED intersection was applied to get the overlap between RepeatMasker and Human regulatory elements, and then rerun overlap between the union data and Human RefGene respectively. From this operation I determined the distribution of repeat-associated regulatory elements with respect to human gene sections. I normalized my results with respect to the number of base pairs in each gene region.

2.1.1.2 Analyze the classes of regulatory repetitive elements

In this part, I have used BED intersection to get the overlap between regulatory elements in human cell lines and human RepeatMasker annotation. As I have already described, I have acquired the various classes of regulatory repetitive elements in different cell lines with respect to various regulators. According to the results I obtained in this part, in the next step, I built six consensus sequences in order to study specific TEs.

2.1.1.3 Building consensus sequences

I have identified thousands of short fragments of repetitive elements. Using consensus sequences from RepeatMasker database I can identify the 5'/3'UTRs only, often annotating repeats as having 5'/3'ends from different repeats. Thus, in order to study the impact of full-length specific retrotransposons on human gene structure and function; I have built complete consensus sequences of specific TEs.

Multiple sequence alignments of full-length sequences were performed using MUSCLE software with default parameters [47], these alignments were then used to run fasttree [48] to generate full-length repetitive elements respectively and used the tool archaeopteryx to generates the classification (See Supplementary Material S1) [49]. Next, I reran the multiple alignment tool MUSCLE to get the alignment sequences between different classes of each TEs, in the last step, I used Gblocks [50] to pile up these alignment results to acquire complete consensus sequences (See Supplementary S2).

Next, I used BED intersection to obtain the distribution of these consensus sequences in different human genes' sections; I only kept the first intron when I encountered alternative splicing. I normalized my data with respect to the length of relevant repetitive elements.

2.2 Functional analysis of human/bovine repeats

To demonstrate the functional significance of repetitive elements, I used DAVID (The Database for Annotation, Visualization and Integrated Discovery) to perform the GO (Gene Ontology) classification, which represents gene product properties. First, I extracted the gene-IDs from the results that were overlapped with repeat consensus sequences in the human genome; I then submitted these gene-IDs to the DAVID Gene Functional Classification Tool [51]. From the results I chose the third level of GO terms to acquire the over-represented function terms of genes that contained repetitive elements. According to the GO term hierarchy; the third level of GO terms contains annotation categories for my analysis. Then I visualized the functional over-representation of genes overlapped with those six specific repeat consensus sequences in the human genome. The thresholds for over-represented GO terms were set as gene count >5 and p-value (EASE score) <0.05. The web

server REVIGO was used to reduce redundancy and visualize the overrepresented GO terms based on semantic similarity.

2.2.1 Repetitive element expression in different human tissues

In order to analyze the expression level of TEs, I have taken advantage of the Human RNA-seqdata Illumina bodyMap2 transcriptome (http://www.ebi.ac.uk/ena/data/view/ERP000546) datasets. These RNA-seq data represent 16 different human tissues. Then, the normalized expression for each transcript dataset is based on my colleague (Z. Qu, unpublished) previous work in our lab. In order to conduct my work, I obtained the proximal promoters that belong to the gene upstream 1000bp. First, I obtained the overlap between repeat and human genes; I then intersected this data with proximal promoter dataset, 3'UTR and 5'UTR of the human genome. Next, I divided the data into repeats in proximal promoter and repeats not in proximal promoter. In the last step, I applied the intersection between the data I have already obtained and the transcriptome datasets, and obtained the expressed genes' interacting repeats in 16 human tissues. I found liver, kidney, testes, brain, skeletal muscle and adipose had relatively higher expression levels in my datasets. Therefore, I extracted these six tissues to analyze the expression level with respect to those three specific repetitive elements. To make the results more clear and significant in my graphs, I used log transformation to better separate the visualize differences in my expression data.

2.3 Relationship between lincRNAs and TEs

Considering the previous study, we have known that TEs are a source of endogenous small RNAs in animals and plants, and they are considered as functionally significant from gene-regulating small RNAs [52]. Furthermore, according to past research, it was found that there are many Alu elements inserted into RNAs. Thus, I planned to study the association between lincRNAs and different classes of TEs.

I have acquired a dataset that contains 8196 putative human lincRNAs (Long intergenic non-coding RNA) [53], then I extracted the sequences of these lincRNAs from whole human sequences, and used BLASTN [54] to align these lincRNAs sequences against the previously constructed repeat consensus sequences I had built. In the next step, I extracted the aligned sequences from the BLAT result in order to analyse their coverage characteristic within lincRNAs.

3 Results

3.1 The distribution of chromatin state associated transposable elements (CSTEs) from six different cell lines

I used the pipeline I have built to study the distribution of transposable elements (TEs) in the human and cow genomes (Figure 3). The large numbers of TEs overlapping with genome intervals from the human and cow RefGene datasets showed that TEs was enriched in different gene sections.

I found that TEs have a similar distribution between the human and cow genome; in total, they are mainly located on 5' UTR intron, CDS intron, 3'UTR intron and intergenic regions. Comparing the TE locations between these two species, I found there were more TEs in each section of the human genome. Intergenic regions in particular have a higher number of TEs in the human compared to cow. The exception is for 5'UTR introns, where there are more TEs in the cow genome (Figure 3).

When I examined to chromatin state functional transposable elements (CSTEs), which represent the TEs that overlap with chromatin state segmentation [45] for each of six human

cell types, I chose six states that have predicted functional elements. Considering the CSTE distribution in the human genome, I divided my results into two parts, the first part is based on CSTE association with regulators in these functional cell lines, the second part is the relationship between CSTEs and different types of those chromatin human cell lines (Figure 4A, B), the human cell lines based on the tissues karyotype (Table 2). My aim was to look at how these two datasets are distributed in the human genome. When we are looked at the CSTEs that overlap with regulatory region distributions in the human genome, I found that the TEs that overlap with active promoters are the most highly represented in the gene start region. However, CSTEs that overlap with weak enhancers and polycomb-repressed regions are enriched in other regions of the human genome (Figure 4A). Among them, I discovered that most of CSTEs' that overlap with regulatory elements are located within the intergenic regions, followed by 5' UTR introns and CDS introns, which may indicate that TEs affect gene transcription through alteration of intron length. Moreover, the CSTEs that may down regulate the gene transcription were always located within the intergenic region, while regulators that can up regulate gene expression were usually located in the 5'UTRs and CDS introns.

In the analysis of the distribution of histone modification that overlapped with TEs in the six human cell lines (Figure 4B), I found that blood cancer cells had CSTEs that were highly enriched in every gene section, while in other cell types, CSTEs were less enriched in various gene features, except for intergenic regions, all of the six cell lines that contained CSTEs were highly enriched in this part, especially blood cancer, lung normal and blood vessel normal. To the contrary, breast normal had CSTEs were always located on the lowest enrichment in those gene sections, which may indicates that the chromatin state breast normal is less active when it contains TEs.

3.2 The proportions of different repeat classes in active chromatin from six distinct cell lines

In this section I set out to study the distribution of specific classes of TEs that overlap with active chromatin from functional regions of the human genome (Figure 5A, B).

Active chromatin from blood cancer cells contained the highest percentage of repeats from virtually all repeat classes, but active chromatin from liver cancer cells ranked either in the mid-range or towards the bottom of repeat percentage (Figure 5B). Over-represented repeats in active chromatin include SINE Mir, LINE L2 and DNA transposons, which are normally represent as much lower percentages of the genome as a whole. LINE L1 elements were present at a surprisingly low level in active chromatin from all cell lines.

When I looked at the overlap of functional regulators (extracted from active chromatin regions) with TEs, I found a similar pattern in each repeat type, with repressed region containing the highest proportion of repeat sequence and active promoters the smallest proportion of repeat sequence (Figure 5A). I also found that MIR and L2 elements overlapped more with regulators compared to other TEs, while L1 had the lowest coverage percentage for all regulators. Because MIR and L2 are molecular fossils, while L1 is a currently active repeat type, this may indicate that ancient TEs may have been exapted during the course of evolution. From this graph I also can conclude that while the single biggest association of repeat coverage is with repressor activity, there is also a strong association with up-regulators of gene expression.

3.3 Repeat sequence distribution in the human genome

I have generated six repeat consensus sequences that represent the principal classes of TEs in the human genome; they are MIR, Alu, L1, LTR, SVA and ERV. These classes of retrotransposons cannot only affect the human genome structure, but also gene transcription activity. I have determined the distribution of these six specific TE classes with respect to different gene features (Figure 6). I found that a large proportion of these TEs are located in 5'UTR, cds-intron and intergenic regions, with L1 and Alu the most common repeats represented.

3.4 Functional representation of repeat consensus sequence in the human genome

I have mapped six primary repeat consensus sequences to the human genome, including gene models. I have used Gene Ontology (GO) annotation of the genes that overlapped with repeats to assess the functional contribution of repeats. By comparing the P-value as my standard index to acquire the definite functions of these repeat consensus sequences.

GO category results fall into three categories: cellular component, molecular function and biological process, with respect to their gene product properties (Table 1). I discovered that most of TEs' are associated with specific functional products that execute biological processes. Figure.7A illustrates that L1 and MIR are significant associated with genes that have many functions in the human genome. L1 in particular is associated with genes that have many functions in biological processes, such as establishment of localization in cell, cell-cell adhesion and cellular component morphogenesis. Moreover, gene associated with MIR participates in cell development processes, protein transport and neuron projection development. ERV associated genes are significantly over-represented with respect to cell-cell adhesion. Furthermore, Mir and L1 have similar clustering with respect to their functions if the genes they are associated with in terms of biological processes.

The second most annotated gene product property is cellular components (Figure 7B). Genes containing L1 and MIR have a number of associated functions. For example, MIR containing genes act as a part of plasma membrane. L1 containing genes are apparently enriched in the functions extrinsic to membrane, dendritic spine, dendritic shaft, extracellular matrix and protein serine/threonine. ERV containing genes have several functions as well, for example, endomembrane system and constitution of the membrane.

The third gene product property is molecular functions (Figure 7C). L1 containing genes functions include: purine nucleotide binding, substrate-specific transmembrane transporter activity and calmodulin binding. Genes containing ERV have functions including cell adhesion molecular binding and GTPase activator activity and enzyme binding. MIR containing genes have functions such as passive transmembrane transporter activity, while genes containing LTR participate in cell adhesion molecular binding.

3.5 The effect of Alu, L1 and LTR on gene expression in 6 human tissues

I analyzed the effect of different TEs on human gene expression by comparing the expression of genes with TEs in either the proximal promoter region, the 5'UTR or the 3'UTR to genes without TEs in those regions. I selected six human tissue transcriptome sets (kidney, liver, brain, testes, skeletal muscle and adipose tissue) from the Illumina BodyMap2 dataset for this analysis.

In order to carry out the analysis, I compared the genomic intervals for TEs to the genomic intervals for gene models to identify genes with and without TEs in the proximal promoter, 5'UTR and 3'UTR. I then plotted log normalized gene expression levels (determined by

TopHat/Cufflinks) of genes from the different TE categories (Figure 8, 9 and 10) to see if overall gene expression levels were correlated with the presence or absence of TEs.

I found that Alus, L1s and LTRs appeared to be associated with lower levels of gene expression in all tissues if they were present in the proximal promoter regions. The exception to this was for testes expressed (Figure 8), where the presence of L1 in the proximal promoter made no different to gene expression and where the effects of other TEs were also less obvious. I did not see changes in gene expression associated with repeats in the 5'UTR, but in the 3'UTR the presence of LTR was associated with decreased gene expression in all six tissues. These 2 suggest that TEs in proximal promoters and 3'UTR could possibly act as repressors of gene expression.

3.6 Are specific repeat sequences present in lincRNAs?

In order to determine if TEs contribute specific sequences to Long Intergenic Non-Coding RNAs (lincRNAs), I aligned lincRNAs [53] and my TE sequences with BLASTN [54]. I found that many lincRNA sequences contained almost full length Alu, but that the alignments showed a greater contribution of left and right monomer sequences from Alu elements (Figure 11A). In particular, the left monomer sequence peak (Figure 11A) might correspond with the internal Alu promoter region. This result suggests a potential regulatory role for Alu sequences in lincRNAs expression.

Mir elements contributed the sequences between position 50 and 200 of their consensus sequence to lincRNAs (Figure 11B). When I extracted this subsequence and scanned it for regulatory motifs (http://asp. ii.uib.no: 8090/cgi-bin/CONSITE/consite), I found this 150bp

subsequence contained low complexity sequence with motifs for Myf and Snail binding snail, CREB and bZIP910.

L1 elements are generally about 7kbp long and L1 alignments to lincRNAs showed a preponderance of sequences from the 3'end of L1, consistent with known 5'truncation of inserted L1 sequences (Figure 11C).

LTR sequences were generally found at low levels in lincRNAs except for a sharp 17bp peak in the middle of the LTR sequence (Figure 11D). When this sequence was extracted from the LTR it was found to be a low complexity tetranucleotide repeat sequence. This 17bp sequence has been annotated as containing regulatory motifs (See Supplementary Material S3).

ERV sequences were also found a generally low level in lincRNAs, but also showed some sharp spikes (Figure 11E). I extracted the subsequences for the two tallest spikes, 21bp at about 3kpb and 17bp at about 7kbp within the ERV consensus. The 21bp subsequence contains a binding site for the Pbx transcription factor and the 17bp subsequence contains low complexity sequence with motifs for Myf and Snail binding (See Supplementary Material S3).

Initial alignment of SVA sequences to lincRNAs showed that the Alu domain of SVA was probably aligning to the Alu regions of lincRNAs (Figure 11F). The VNTR domain of SVA showed large numbers of hits in lincRNAs and the 100bp subsequence from this region contained binding motifs for 8 transcription factors (See Supplementary Material S3).

The fact that alignment peaks in lincRNAs contained transcription factor binding motifs contributed from 4 TEs in suggestive of a role for these motifs in lincRNAs regulation or function.

Discussion

In this work, I have analyzed the function, distribution and expression of various classes of TEs in the human genome. I have found that TE distribution is similar in human and cow that TEs appear to be able to alter gene expression, based on their distribution in functional regions and their correlation with gene expression. Furthermore, I found that genes containing TEs were over-represented in terms of their GO annotations in all three categories. Finally, I found that TEs are present in ncRNAs, specifically lincRNAs, and that some repeat classes appeared to contribute specific subsequences that contain functions to the lincRNAs.

TEs had similar distributions in human and cow (Figure 3), with a greater proportion of sequence originating from TEs in the 5' and 3'UTRs compared to coding exons or start and stop codons. 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 regulatory functions within the UTRs [55, 56]. The repeat content for 5'UTR introns was comparable to remaining introns, but this may be significant in the context of transcriptional repression, where genes with shorter 5' UTR introns were expressed at higher levels [57, 58]. It would appear that in some genes the exaptation of repeats into UTRs is neutral or functional (see below).

The distribution of TEs in genomic regions epigenetically modified to regulate transcription (active chromatin) was consistent with a potential role as regulators of gene expression. From my results (Figure 4A, B), I found that some functional regions of active chromatin contained

higher levels of TE, specifically polycomb-repressed regions and weak enhancers. This was most striking for the polycomb-repressed regions, which was consistent with epigenetic silencing of TEs. This was also consistent with the higher level of repeat content seen in 5'UTR, which are also known to regulate gene [59]. Furthermore, it has been shown that TEs in 3'UTR are associated with lower transcript abundance [60]. This suggests that exaptation of repeat into regulatory regions is most often associated with repression of gene expression. In fact, older repeat, such as fossil L2 and MIR, are more prevalent in active chromatin than recently inserted repeats, such as L1 or Alu, supporting the argument for exaptation.

Different repeat classes were clearly present at different levels in active chromatin or specific regulatory regions, such as polycomb-repressed (Figure 5), implying an association with gene expression. My analysis of gene expression as a function of regulatory region repeats content indicated that there was a general association between repeat content and lower gene expression (Figures 8, 9 and 10). These results are consistent with previous reports showing TEs such as Alus can be exapted as regulators of alternative splicing, as transcription factor binding sites and as sense and antisense promoters [32, 61]. Furthermore, the presence of Alu pairs in opposite orientations in gene transcripts can lead to adenosine to inosine editing, resulting in suppression of expression through nuclear retention of edited RNA transcripts [25]. My results are consistent with TEs mainly down regulating gene expression if they were present in the proximal promoter or 3'UTR region. This association was weakest or non-existent for L1 elements (Figures 8, 9 and 10). Furthermore, L1 elements were less prevalent in regulatory regions or active chromatin, compared to other repeat classes (Figure 5). This makes sense because most L1 elements in the human genome are 5' truncated [60], lacking promoters and other regulatory sequences. LTR repeat were similarly distributed, compared to non-LTR repeats, and associated with repression of gene expression. This is in contrast to

previous work that implicates LTRs as alternative promoters [62]. However, an LTR in the first intron of the equine TRPM gene suppresses gene expression by acting as an alternative poly-A site (Bellone et al, unpublished), and the insertion of LTRs in introns has been associated with premature termination of transcription [63].

In addition to regulating gene expression, TEs may also be associated with specific functional characteristics of expressed protein coding genes. When I examined the functional annotation of repeat containing genes, I found that some functions were over-represented (Figure 7). This might occur because exaptation of TEs into coding sequences is associated with particular functions or that particular functional classes of genes are co-regulated and that repeats are involved in this co-regulation.

Non-coding genes also included exapted TEs or functionally important TE subsequences. Specifically I found that lincRNAs, which are known to regulate gene expression through epigenetic mechanisms and competition for transcription factors [64-66], contained TE sequences. Not all TEs contributed sequences equally, most of the Alu consensus was found in lincRNAs, but for L1s, 3' truncated sequences were dominant and for MIRs, the central 150bp was most often found in lincRNAs. This was in contrast to LTR, SVA and ERV TEs, which contributed much more defined, shorter sequences that were shown to contain transcription factor binding motifs. This result is consistent with a role for lincRNAs as competitors for transcription factors that bind to promoter regions.

In conclusion I can say that TEs are most strongly associated with repression of gene expression, either through the proximal promoter or 3'UTR or by contributing sequences to lincRNAs. These results extend previous published work in a novel and more comprehensive

fashion. Finally, I speculate that TEs may influence gene expression by co-regulating functionally similar genes, and I believe this to be a novel hypothesis.

Future Directions

I have analyzed the repeats distribution, association with different gene sections and functional elements as well as expression of various classes of repeats in the human genome. However, my research target was limited to Homo sapiens. Thus, applying these methods and pipelines to other species is my next work, which may include chimpanzee, cow and elephant. Furthermore, in my analysis of transposable elements influences on human gene expression, I only focused on Alu, L1 and LTR in six human tissues. In next step, I may explore other classes of TEs impact on human gene transcript level, and also study the association between different types of TEs and tissues, to find out whether the influence of TEs in genes has any tissue specific. In this paper, I have studied the association between repeats and various regulatory elements, which were only based on active chromatin states. Therefore, I may transfer to the association between repeats and heterochromatin in our next work, giving us better understanding of the association between TEs and epigenetic regulation.

According to recent studies, piRNAs from RNA-protein complexes through interactions with piwi proteins, and these piRNA complexes have been linked to both epigenetic and post-transcriptional gene silencing of retrotransposons and other genetic elements in germ line [67]. This finding can help me to further explore the mechanism about how retrotranspsons interact with piwi interacting RNA to impact the gene expression.

Abbreviations List:

TEs: transposable elements

ORFs: open reading frames

LTR: long terminal repeats

LINEs: long interspersed elements

SINEs: short interspersed elements

ADAR: double-stranded RNA-specific adenosine deaminase

GO: Gene ontology

lincRNAs: long intergenic non-coding RNA

CSTE: chromatin states transposable elements

Figures and table legends:

Figure 3. Gene feature distributions of repetitive elements in human and cow

The y-axis has been normalized by the equation

$$percentage = \frac{repeats\ in\ specific\ genome\ section\ (bp)}{genome\ specific\ section\ (bp)} \times 100$$

Figure 4. Gene feature distributions of biologically activity of repetitive elements

The biological activity of repetitive elements was measured as the percentage of bases that overlapped repetitive elements and genomic areas of a particular chromatin state. Areas of genome that have been named biologically active are those that do not overlap heterochromatic areas. Shown are the specific biological activities of repetitive elements, averaged over six different cell lines (A), and the total biological activity of repetitive elements for each cell line (B). Percentages in A and B were normalized by the proportions of

genome occupying a particular gene feature. The method of normalization is same with the figure 3.

Figure 5. Biological activity of different repeat classes

The biological activity of a repeat class was measured as the percentage of bases that overlapped with particular repeat class and genomic areas of a particular chromatin state. Areas of genome that have been named biologically active are those that do not overlap heterochromatic areas. Shown are the specific activities of repeat classes, averaged over six different cell lines (A), and the total activity of repeat classes for each cell line (B). Percentages in A and B have been normalized by the proportions of genome occupied by a particular repeat class. Normalization equation:

$$Percentage = \frac{Specific\ chromatin\ state\ TEs\ (bp)}{Specific\ TEs\ (bp)} \times 100$$

Figure 6. Gene feature distributions of repeat classes

Percentages of genome that overlap a certain repeat class and gene feature. Percentages are normalized by the proportion of genome occupied by a particular gene feature. The method of normalization is same with the figure 3.

Figure 7. Enrichment of GO terms of various repeat classes in the human genome

Enrichment is based on a log (p-value) transformation, which is shown over three GO domains, biological process (A), cellular compartment (B), and molecular function (C).

Figure 8. Expression level of three repeats in kidney (A), liver (B), brain (C), testes (D), skeletal muscle (E) and adipose (F) in the part of proximal promoter of human genes. I have used log transformation to acquire these violin plots. The white dot in the middle

represents the medium value of each sub graphs; the width of each subplot shows the gene number that enriched with respect to the y-axis.

Figure 9. Expression level of three repeats in kidney (A), liver (B), brain (C), testes (D), skeletal muscle (E) and adipose (F) in 5'UTR region of human genes. I have used log transformation to obtain these violin plots. The white dot in the middle represents the medium value of each sub graphs; the width of each subplot shows the gene number that enriched with respect to the y-axis.

Figure 10. Expression level of three repeats in kidney (A), liver (B), brain (C), testes (D), skeletal muscle (E) and adipose (F) in 3'UTR section of human genes. I have used log transformation to get these violin plots. The white dot in the middle represents the medium value of each sub graphs; the width of each subplot shows the gene number that enriched with respect to the y-axis.

Figure 11. Coverage plot showing the length distribution of aligned sequence between lincRNA and different types of repetitive elements. The plot shows Alu, Mir, LTR, SVA, L1 and ERV these six repeat consensus sequences association with lincRNAs. X-axis shows the whole length of each relevant repeat consensus sequence, y-axis signifies the number of the alignment length that covered in that region.

Figure 3:

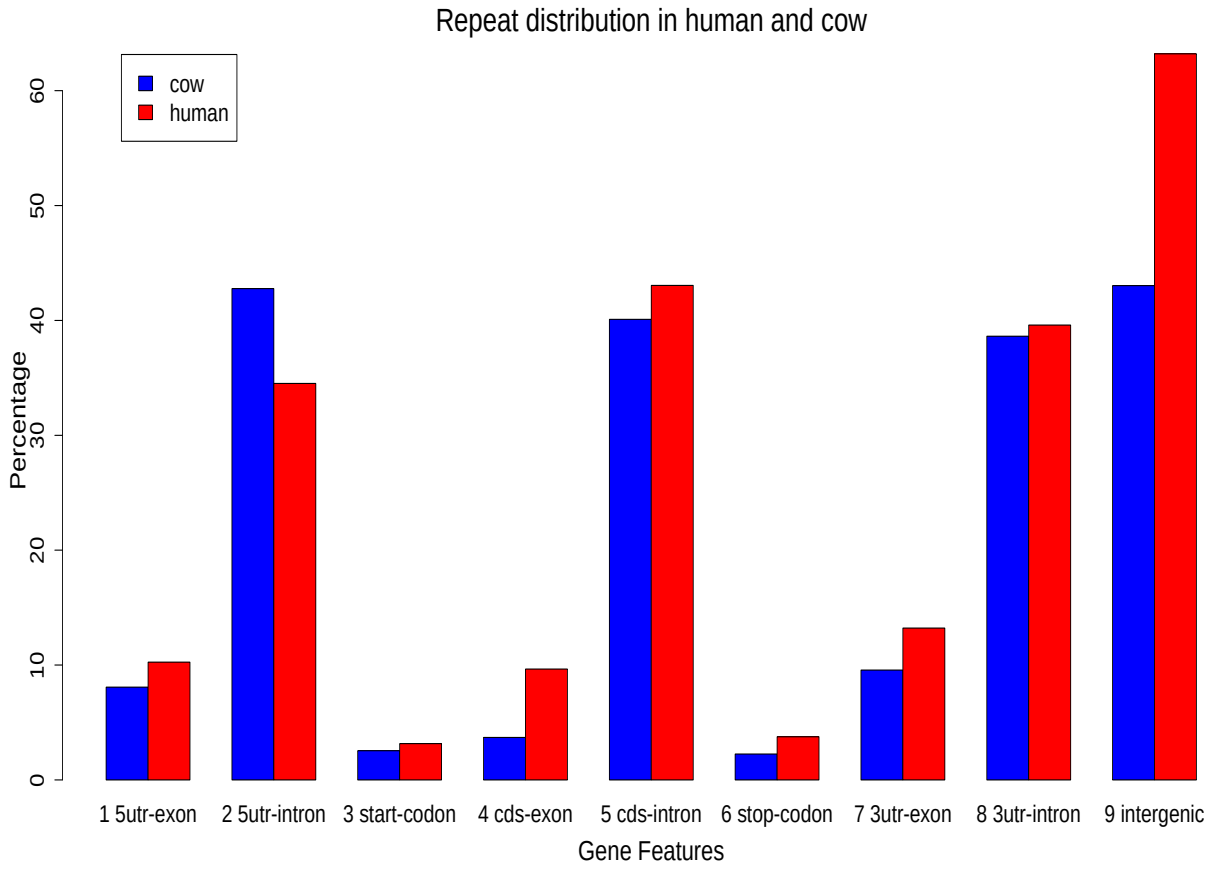

Repeat distribution in human and cow
cow
human
Percentage
0
10
20
30
40
50
60
1 5utr-exon
2 5utr-intron
3 start-codon
4 cds-exon
5 cds-intron
6 stop-codon
7 3utr-exon
8 3utr-intron
9 intergenic
Gene Features

Figure 4:

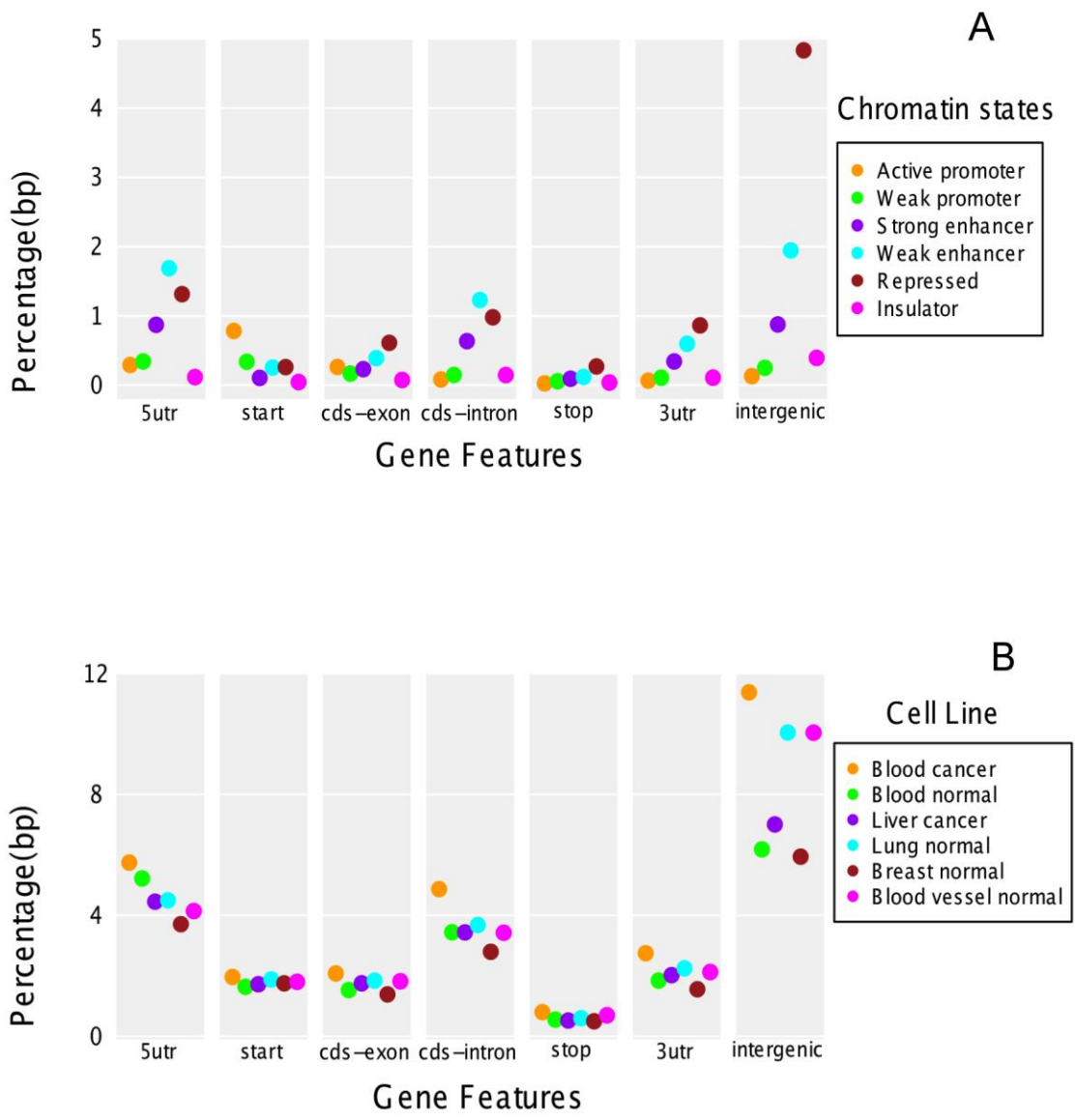

A
Chromatin states
Active promoter
Weak promoter
Strong enhancer
Weak enhancer
Repressed
Insulator
Percentage(bp)
0
1
2
3
4
5
5utr
start
cds–exon
cds–intron
stop
3utr
intergenic
Gene Features
B
Cell Line
Blood cancer
Blood normal
Liver cancer
Lung normal
Breast normal
Blood vessel normal
Percentage(bp)
0
4
8
12
5utr
start
cds–exon
cds–intron
stop
3utr
intergenic
Gene Features

Figure 5:

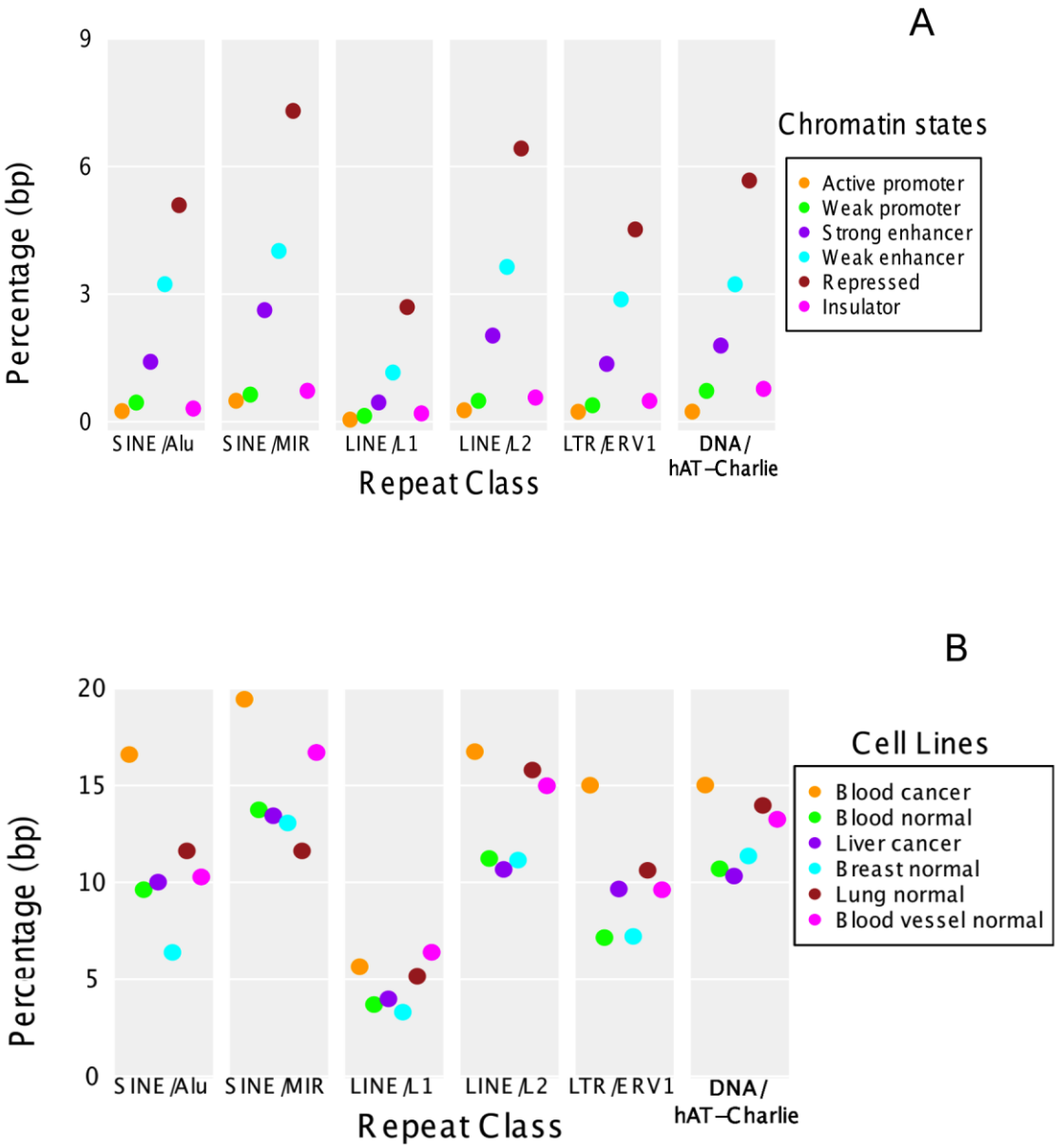

A
Percentage (bp)
9
6
3
0
SINE /Alu
SINE /MIR
LINE /L1
LINE /L2
LTR /ERV1
DNA/
hAT–Charlie
Repeat Class
Chromatin states
Active promoter
Weak promoter
Strong enhancer
Weak enhancer
Repressed
Insulator
B
Percentage (bp)
20
15
10
5
0
SINE /Alu
SINE /MIR
LINE /L1
LINE /L2
LTR /ERV1
DNA/
hAT–Charlie
Repeat Class
Cell Lines
Blood cancer
Blood normal
Liver cancer
Breast normal
Lung normal
Blood vessel normal

Figure 6:

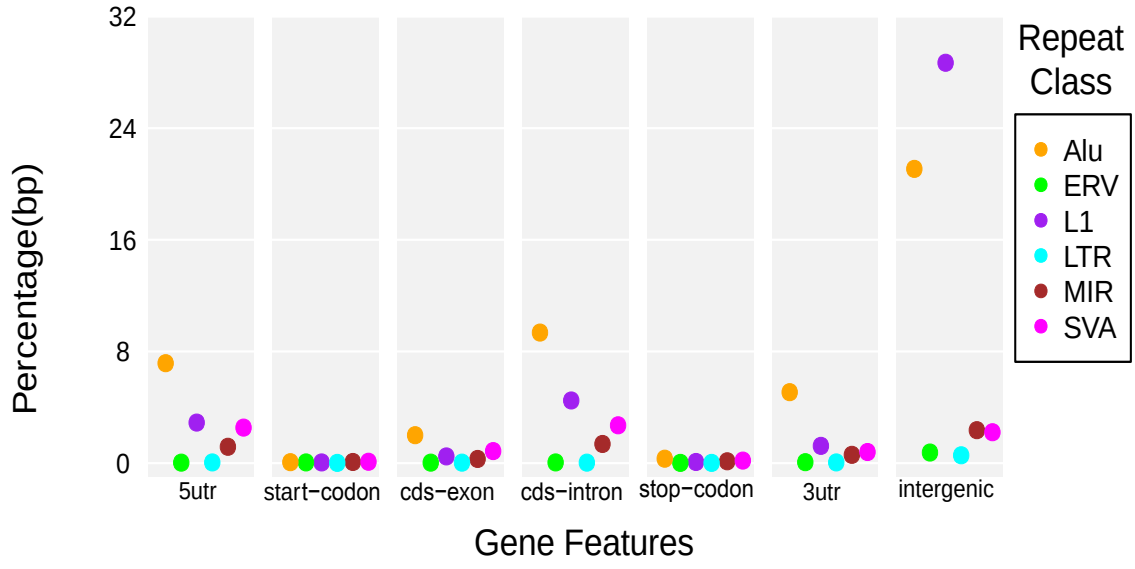
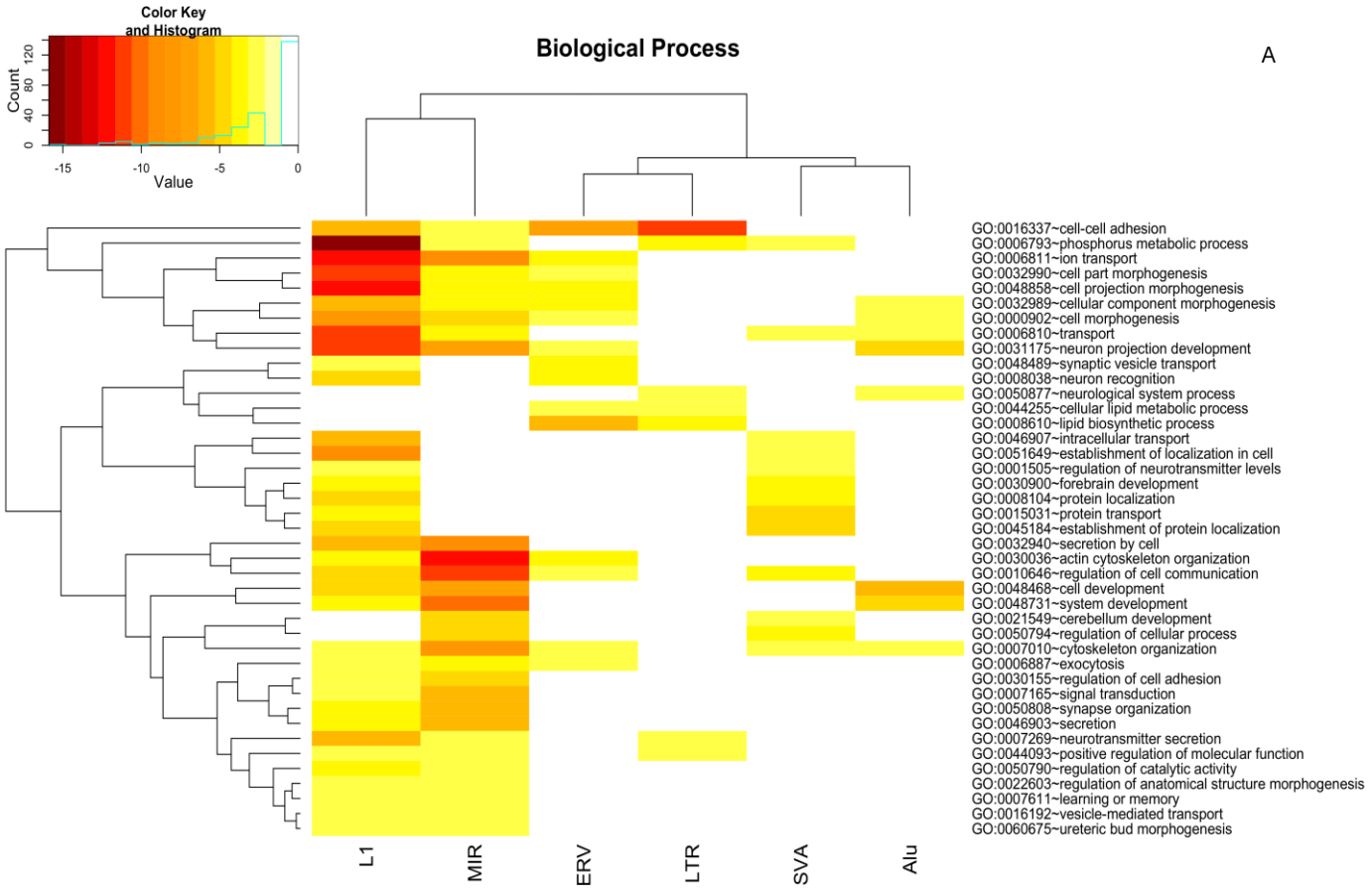
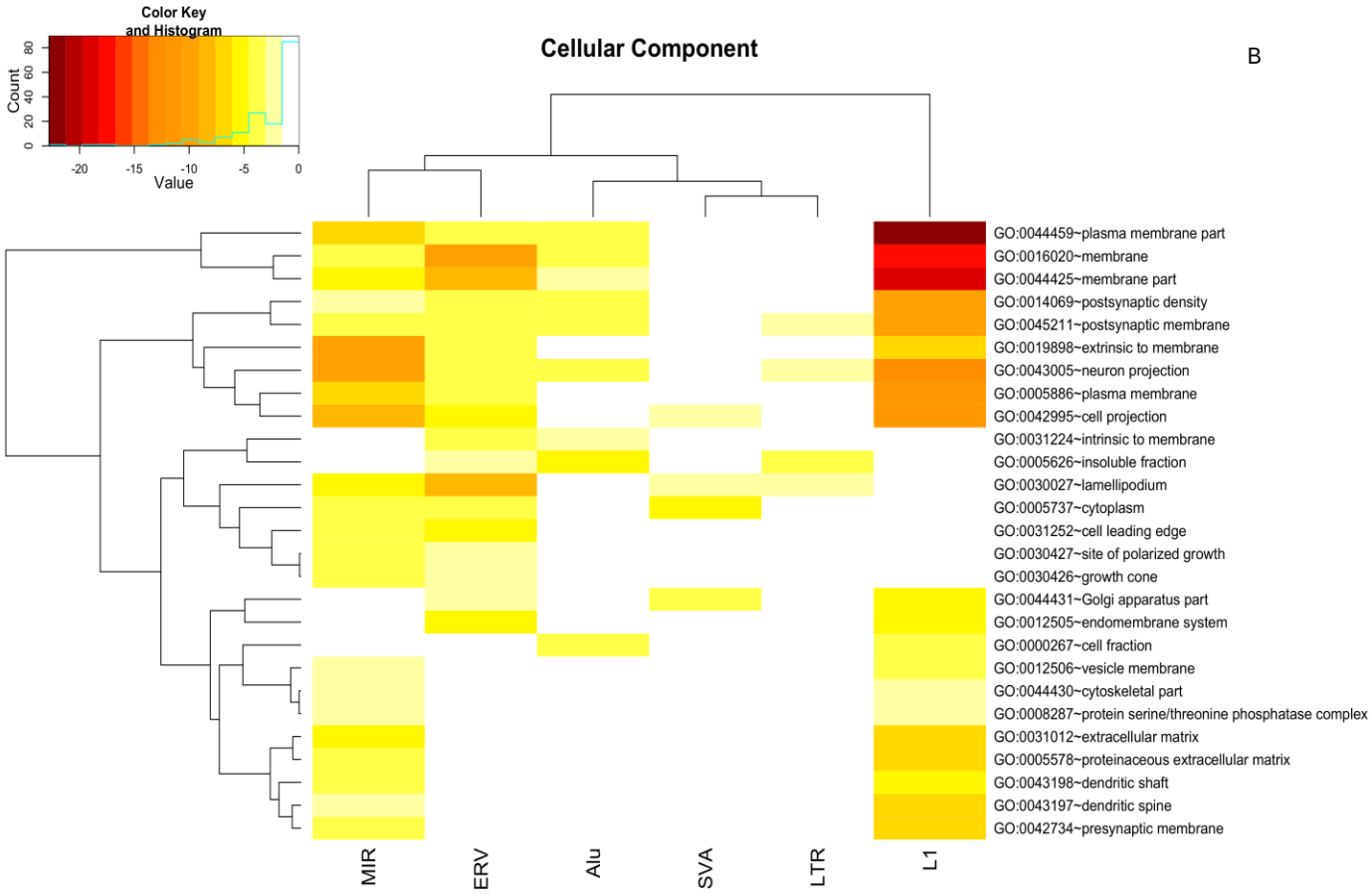

32
24
16
8
0
Percentage(bp)
5utr
start−codon
cds−exon
cds−intron
stop−codon
3utr
intergenic
Gene Features
Repeat Class
Alu
ERV
L1
LTR
MIR
SVA

Figure 7:


Color Key and Histogram
Count
0
40
80
120
-15
-10
-5
0
Value
Biological Process
A
GO:0016337~cell-cell adhesion
GO:0006793~phosphorus metabolic process
GO:0006811~ion transport
GO:0032990~cell part morphogenesis
GO:0048858~cell projection morphogenesis
GO:0032989~cellular component morphogenesis
GO:0000902~cell morphogenesis
GO:0006810~transport
GO:0031175~neuron projection development
GO:0048489~synaptic vesicle transport
GO:0008038~neuron recognition
GO:0050877~neurological system process
GO:0044255~cellular lipid metabolic process
GO:0008610~lipid biosynthetic process
GO:0046907~intracellular transport
GO:0051649~establishment of localization in cell
GO:0001505~regulation of neurotransmitter levels
GO:0030900~forebrain development
GO:0008104~protein localization
GO:0015031~protein transport
GO:0045184~establishment of protein localization
GO:0032940~secretion by cell
GO:0030036~actin cytoskeleton organization
GO:0010646~regulation of cell communication
GO:0048468~cell development
GO:0048731~system development
GO:0021549~cerebellum development
GO:0050794~regulation of cellular process
GO:0007010~cytoskeleton organization
GO:0006887~exocytosis
GO:0030155~regulation of cell adhesion
GO:0007165~signal transduction
GO:0050808~synapse organization
GO:0046903~secretion
GO:0007269~neurotransmitter secretion
GO:0044093~positive regulation of molecular function
GO:0050790~regulation of catalytic activity
GO:0022603~regulation of anatomical structure morphogenesis
GO:0007611~learning or memory
GO:0016192~vesicle-mediated transport
GO:0060675~ureteric bud morphogenesis
L1
MIR
ERV
LTR
SVA
Alu

B


Color Key and Histogram
Count
0
20
40
60
80
-20
-15
-10
-5
0
Value
Cellular Component
GO:0044459~plasma membrane part
GO:0016020~membrane
GO:0044425~membrane part
GO:0014069~postsynaptic density
GO:0045211~postsynaptic membrane
GO:0019898~extrinsic to membrane
GO:0043005~neuron projection
GO:0005886~plasma membrane
GO:0042995~cell projection
GO:0031224~intrinsic to membrane
GO:0005626~insoluble fraction
GO:0030027~lamellipodium
GO:0005737~cytoplasm
GO:0031252~cell leading edge
GO:0030427~site of polarized growth
GO:0030426~growth cone
GO:0044431~Golgi apparatus part
GO:0012505~endomembrane system
GO:0000267~cell fraction
GO:0012506~vesicle membrane
GO:0044430~cytoskeletal part
GO:0008287~protein serine/threonine phosphatase complex
GO:0031012~extracellular matrix
GO:0005578~proteinaceous extracellular matrix
GO:0043198~dendritic shaft
GO:0043197~dendritic spine
GO:0042734~presynaptic membrane
MIR
ERV
Alu
SVA
LTR
L1

C

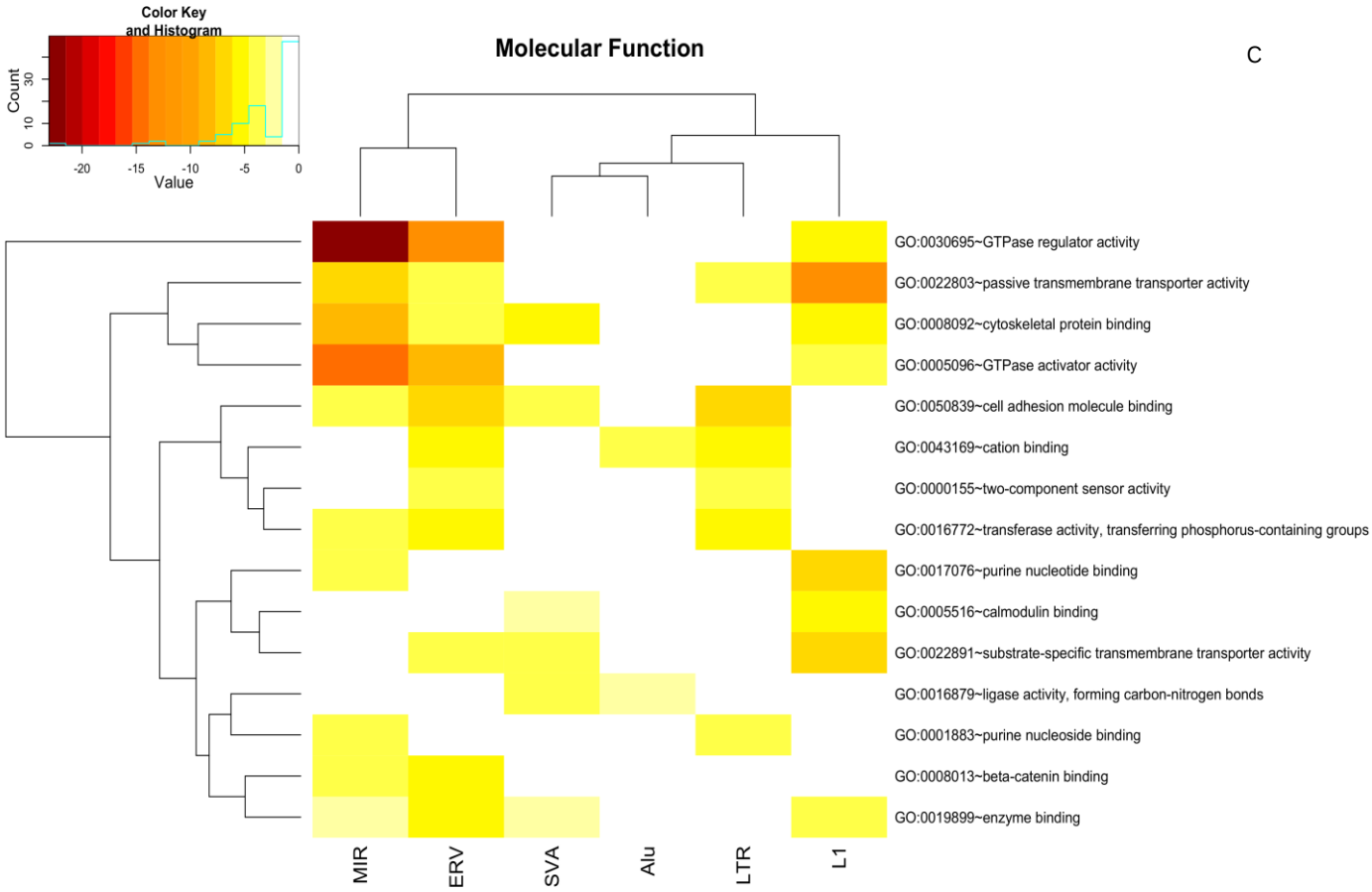

Color Key
and Histogram
Count
0
10
30
-20
-15
-10
-5
0
Value
Molecular Function
GO:0030695~GTPase regulator activity
GO:0022803~passive transmembrane transporter activity
GO:0008092~cytoskeletal protein binding
GO:0005096~GTPase activator activity
GO:0050839~cell adhesion molecule binding
GO:0043169~cation binding
GO:0000155~two-component sensor activity
GO:0016772~transferase activity, transferring phosphorus-containing groups
GO:0017076~purine nucleotide binding
GO:0005516~calmodulin binding
GO:0022891~substrate-specific transmembrane transporter activity
GO:0016879~ligase activity, forming carbon-nitrogen bonds
GO:0001883~purine nucleoside binding
GO:0008013~beta-catenin binding
GO:0019899~enzyme binding
MIR
ERV
SVA
Alu
LTR
L1

Figure 8:

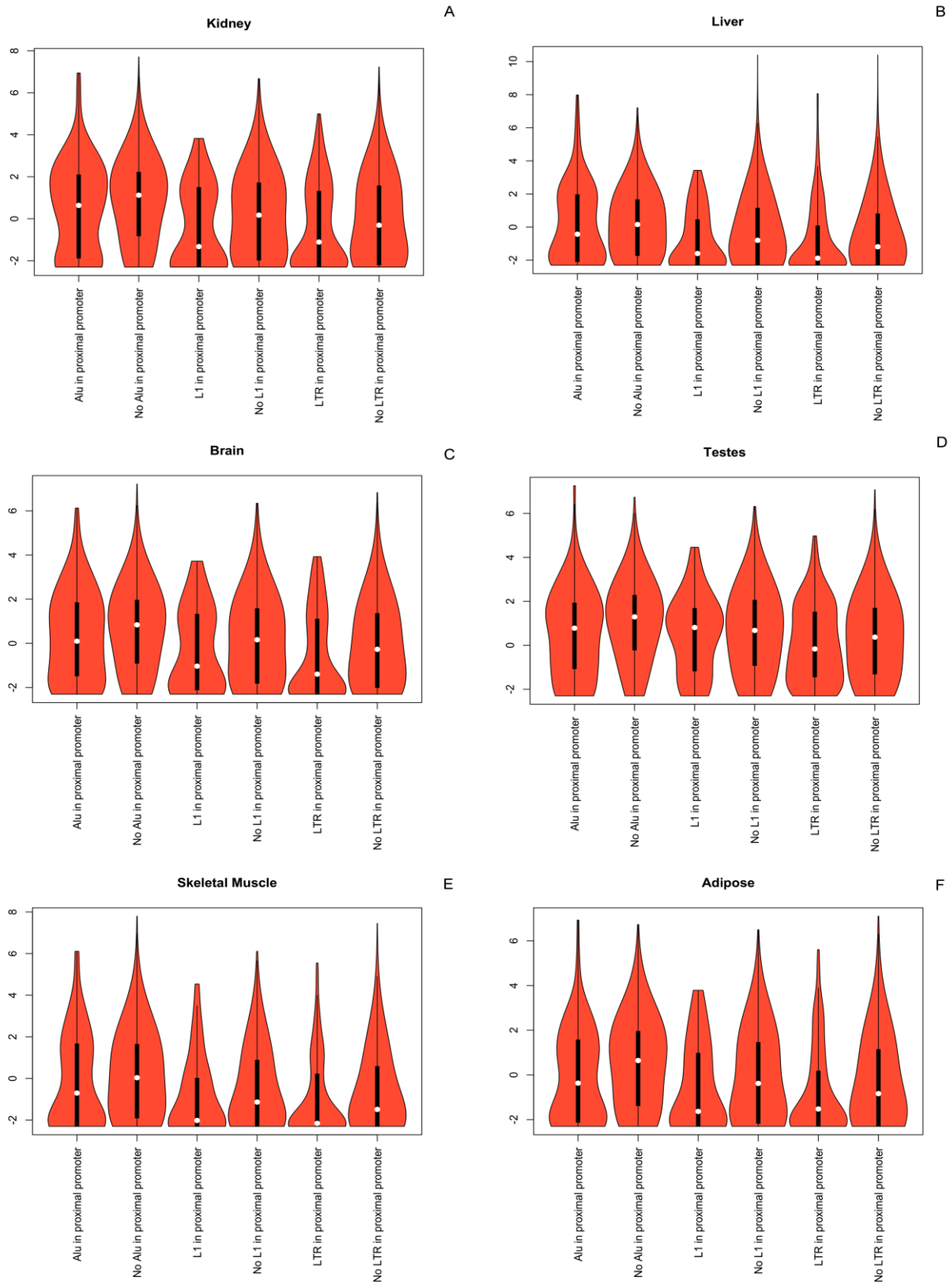

A
Kidney
B
Liver
C
Brain
D
Testes
E
Skeletal Muscle
F
Adipose
Alu in proximal promoter
No Alu in proximal promoter
L1 in proximal promoter
No L1 in proximal promoter
LTR in proximal promoter
No LTR in proximal promoter

Figure 9:

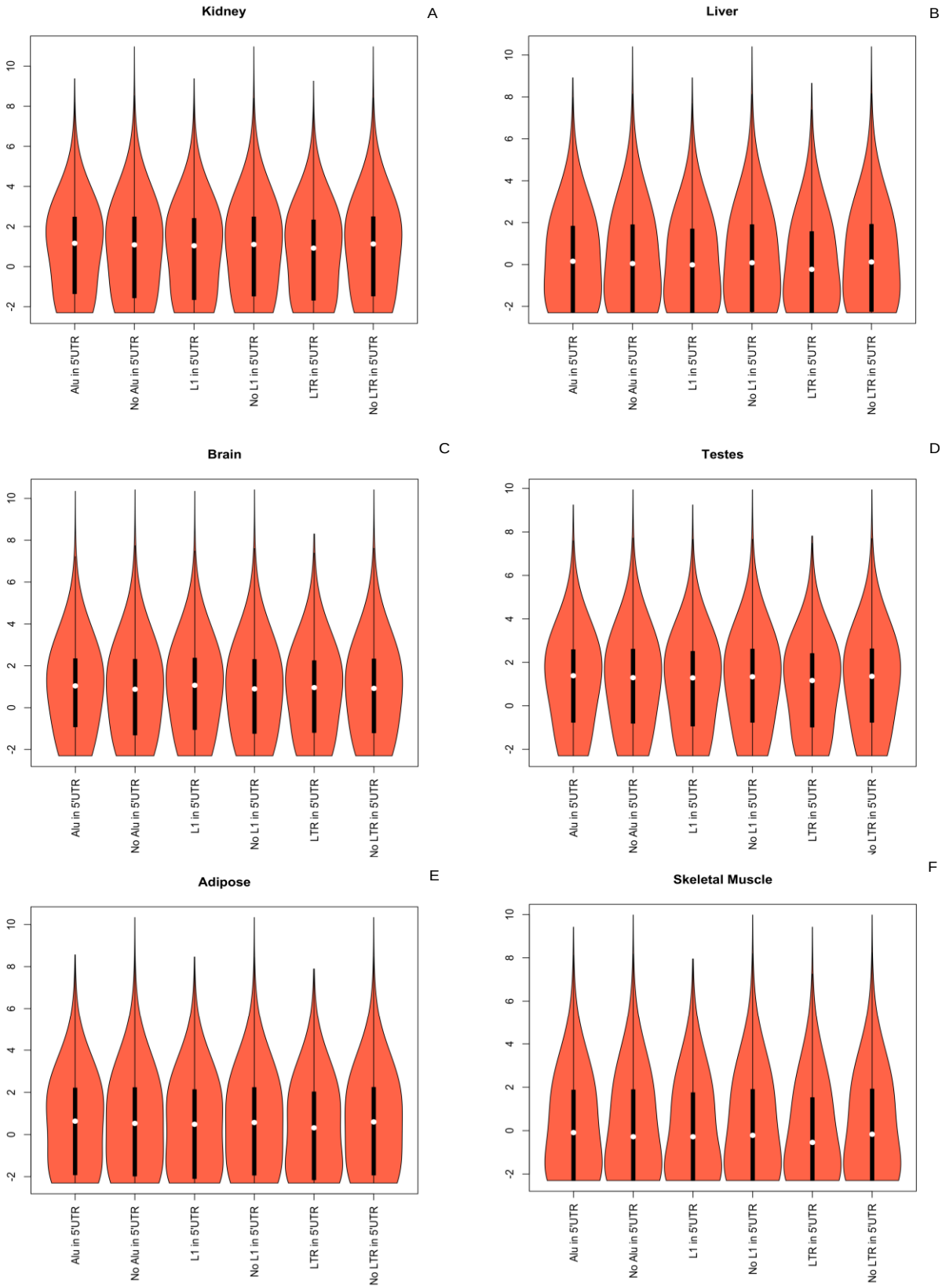

A
Kidney
B
Liver
C
Brain
D
Testes
E
Adipose
F
Skeletal Muscle
-2
0
2
4
6
8
10
Alu in 5'UTR
No Alu in 5'UTR
L1 in 5'UTR
No L1 in 5'UTR
LTR in 5'UTR
No LTR in 5'UTR

Figure 10:

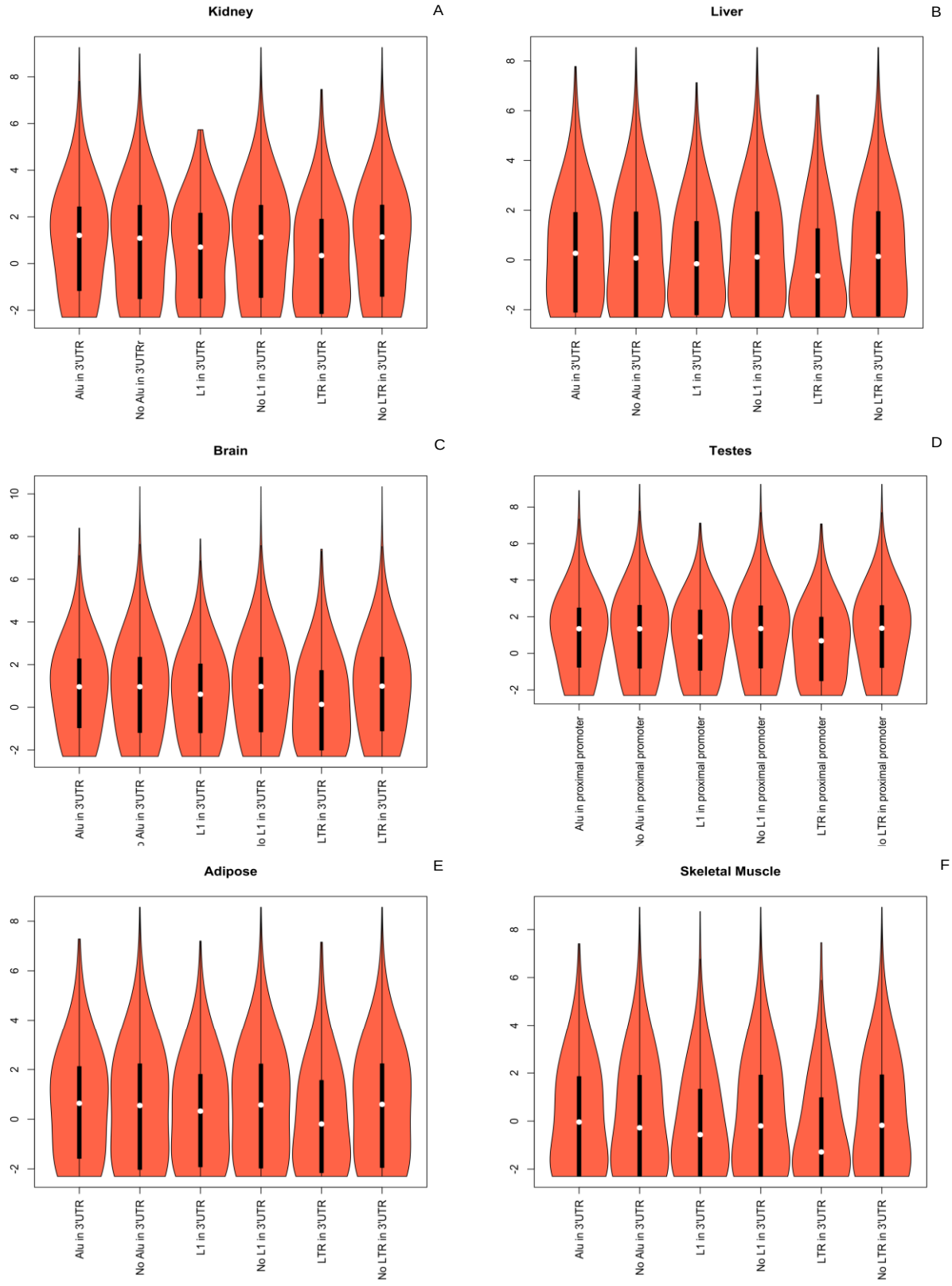
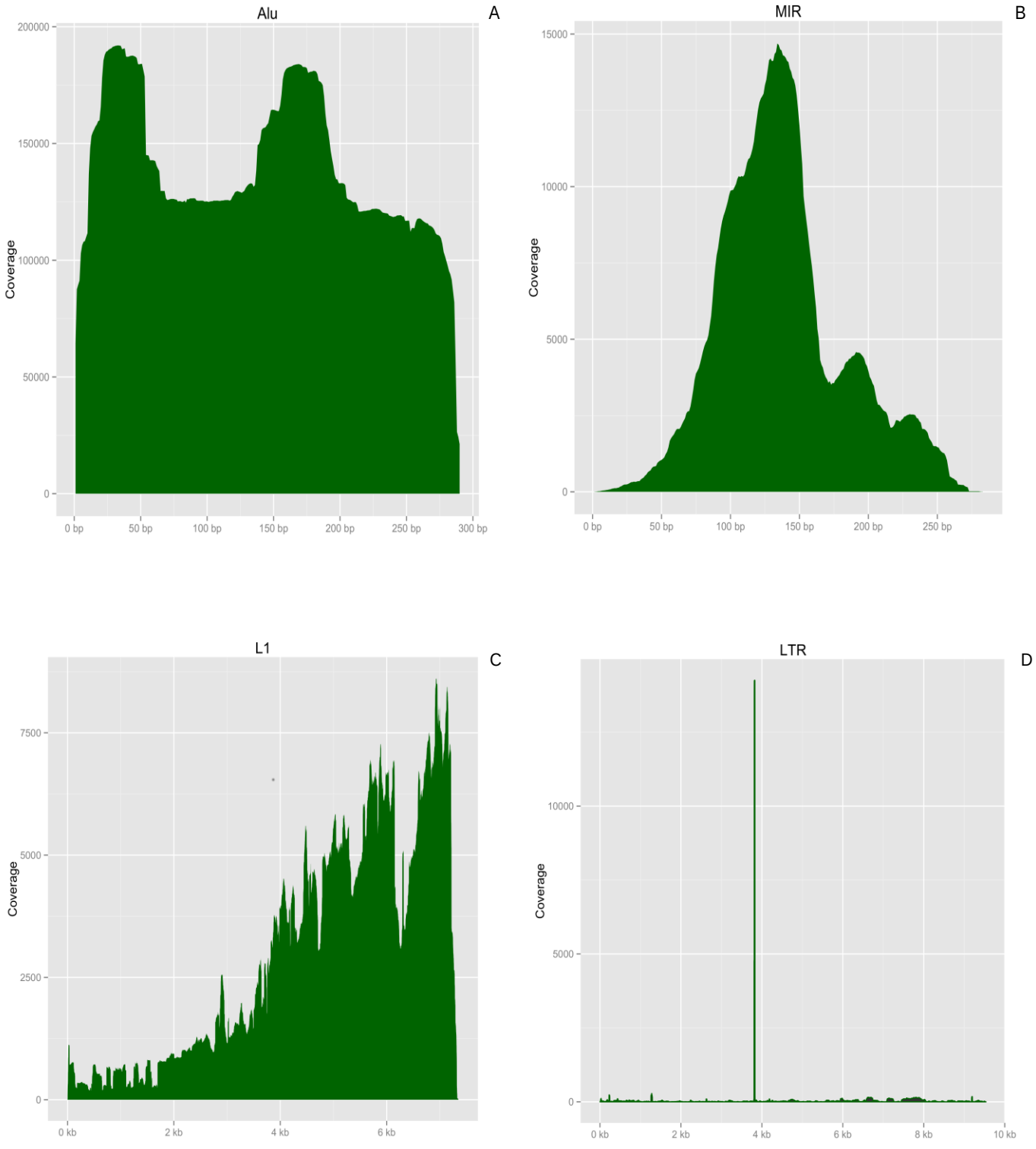
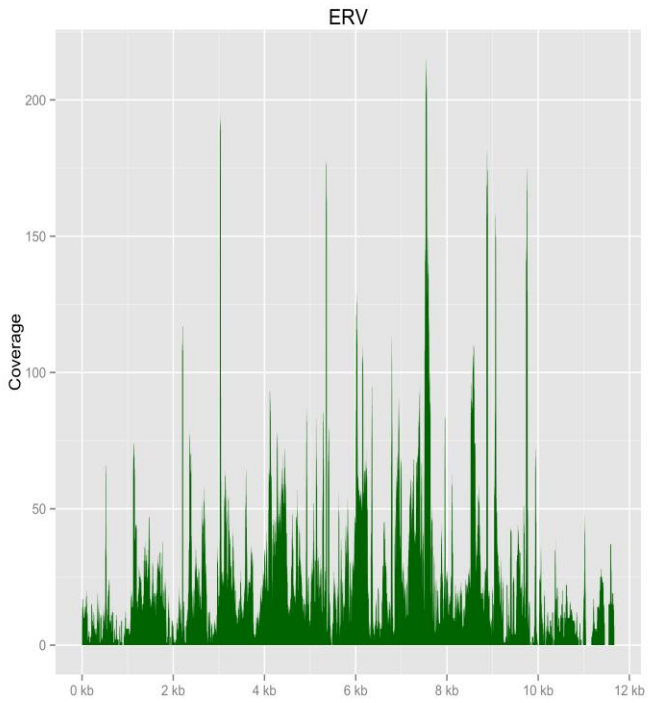
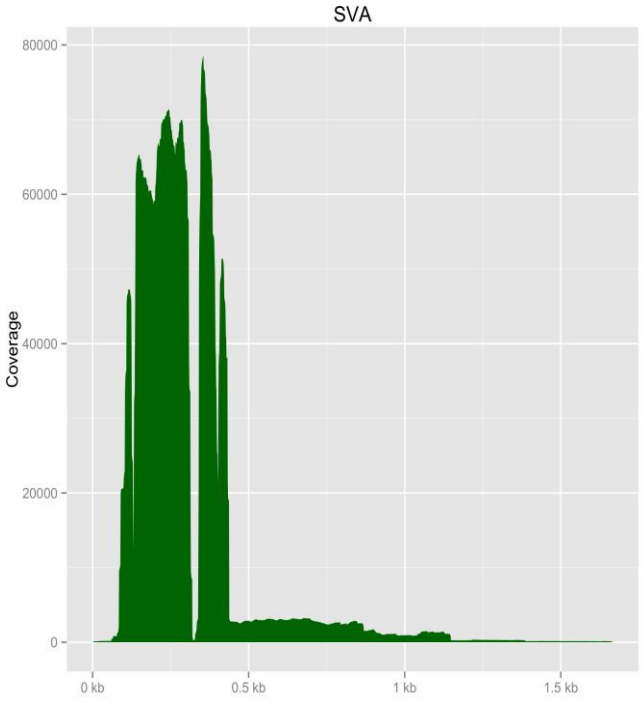

Kidney
A
Alu in 3'UTR
No Alu in 3'UTR
L1 in 3'UTR
No L1 in 3'UTR
LTR in 3'UTR
No LTR in 3'UTR
Liver
B
Brain
C
Testes
D
Alu in proximal promoter
No Alu in proximal promoter
L1 in proximal promoter
No L1 in proximal promoter
LTR in proximal promoter
Adipose
E
Skeletal Muscle
F

Figure 11:


Alu
A
Coverage
200000
150000
100000
50000
0
0 bp
50 bp
100 bp
150 bp
200 bp
250 bp
300 bp
MIR
B
Coverage
15000
10000
5000
0
0 bp
50 bp
100 bp
150 bp
200 bp
250 bp
L1
C
Coverage
7500
5000
2500
0
0 kb
2 kb
4 kb
6 kb
LTR
D
Coverage
10000
5000
0
0 kb
2 kb
4 kb
6 kb
8 kb
10 kb


ERV
Coverage
200
150
100
50
0
0 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb

E


SVA
Coverage
80000
60000
40000
20000
0
0 kb
0.5 kb
1 kb
1.5 kb

F

Table 1. GO terms with high enrichment in the human genome within six different classes of TEs

Category	Term	P-value	Repeats
Biological Process	GO:0048468~cell development	0.002433068	Alu
Biological Process	GO:0031175~neuron projection development	0.010950207	Alu
Biological Process	GO:0048731~system development	0.014218732	Alu
Biological Process	GO:0000902~cell morphogenesis	0.045430207	Alu
Biological Process	GO:0016337~cell-cell adhesion	6.76E-04	ERV
Biological Process	GO:0008610~lipid biosynthetic process	0.004836686	ERV
Biological Process	GO:0007565~female pregnancy	0.007958811	ERV
Biological Process	GO:0010942~positive regulation of cell death	0.021215202	ERV
Biological Process	GO:0030036~actin cytoskeleton organization	0.022252018	ERV
Biological Process	GO:0006629~lipid metabolic process	0.029784268	ERV
Biological Process	GO:0006811~ion transport	0.030293183	ERV
Biological Process	GO:0048489~synaptic vesicle transport	0.033240211	ERV
Biological Process	GO:0048858~cell projection morphogenesis	0.041187851	ERV
Biological Process	GO:0032989~cellular component morphogenesis	0.04121194	ERV
Biological Process	GO:0008038~neuron recognition	0.0416136	ERV
Biological Process	GO:0000902~cell morphogenesis	0.0448662	ERV
Biological Process	GO:0006793~phosphorus metabolic process	1.26E-07	L1
Biological Process	GO:0006811~ion transport	5.44E-06	L1
Biological Process	GO:0048858~cell projection morphogenesis	6.02E-06	L1
Biological Process	GO:0032990~cell part morphogenesis	1.54E-05	L1
Biological Process	GO:0031175~neuron projection development	1.54E-05	L1
Biological Process	GO:0006810~transport	1.83E-05	L1
Biological Process	GO:0051649~establishment of localization in cell	1.13E-04	L1

Biological Process	GO:0000902~cell morphogenesis	2.82E-04	L1
Biological Process	GO:0007411~axon guidance	7.57E-04	L1
Biological Process	GO:0046907~intracellular transport	0.001803868	L1
Biological Process	GO:0032835~glomerulus development	0.002073621	L1
Biological Process	GO:0032989~cellular component morphogenesis	0.002253662	L1
Biological Process	GO:0007267~cell-cell signaling	0.002650669	L1
Biological Process	GO:0016337~cell-cell adhesion	0.003030857	L1
Biological Process	GO:0055085~transmembrane transport	0.00371976	L1
Biological Process	GO:0007269~neurotransmitter secretion	0.003912961	L1
Biological Process	GO:0032940~secretion by cell	0.003993018	L1
Biological Process	GO:0048468~cell development	0.007052635	L1
Biological Process	GO:0008038~neuron recognition	0.007550272	L1
Biological Process	GO:0008104~protein localization	0.00963923	L1
Biological Process	GO:0045184~establishment of protein localization	0.010047077	L1
Biological Process	GO:0030031~cell projection assembly	0.011954353	L1
Biological Process	GO:0010646~regulation of cell communication	0.013264599	L1
Biological Process	GO:0048731~system development	0.015254729	L1
Biological Process	GO:0046903~secretion	0.016725195	L1
Molucular Function	GO:0043169~cation binding	0.030378079	Alu
Molucular Function	GO:0005102~receptor binding	0.035545206	Alu
Molecular Function	GO:0030695~GTPase regulator activity	3.76E-06	ERV
Molecular Function	GO:0005096~GTPase activator activity	4.18E-04	ERV
Molecular Function	GO:0050839~cell adhesion molecule binding	9.41E-04	ERV
Molecular Function	GO:0019899~enzyme binding	0.002702865	ERV
Molecular Function	GO:0043169~cation binding	0.005550116	ERV
Molecular Function	GO:0008013~beta-catenin binding	0.00576055	ERV
Molecular Function	GO:0016772~transferase activity, transferring phosphorus-containing	0.007591362	ERV

	groups		
Molecular Function	GO:0000155~two-component sensor activity	0.012793986	ERV
Molecular Function	GO:0008092~cytoskeletal protein binding	0.015298604	ERV
Molecular Function	GO:0022803~passive transmembrane transporter activity	0.021237808	ERV
Molecular Function	GO:0008656~caspase activator activity	0.022962103	ERV
Molecular Function	GO:0022891~substrate-specific transmembrane transporter activity	0.035814672	ERV
Molecular Function	GO:0022803~passive transmembrane transporter activity	2.97E-06	L1
Molecular Function	GO:0043169~cation binding	2.99E-06	L1
Molecular Function	GO:0016772~transferase activity, transferring phosphorus-containing groups	1.20E-05	L1
Molecular Function	GO:0001883~purine nucleoside binding	1.37E-04	L1
Molecular Function	GO:0032553~ribonucleotide binding	7.33E-04	L1
Molecular Function	GO:0017076~purine nucleotide binding	0.001280726	L1
Molecular Function	GO:0022891~substrate-specific transmembrane transporter activity	0.002093544	L1
Molecular Function	GO:0005516~calmodulin binding	0.003082402	L1
Molecular Function	GO:0030695~GTPase regulator activity	0.005323101	L1
Molecular Function	GO:0008092~cytoskeletal protein binding	0.007427365	L1
Molecular Function	GO:0005096~GTPase activator activity	0.011907	L1
Molecular Function	GO:0019899~enzyme binding	0.013789565	L1
Molecular Function	GO:0005543~phospholipid binding	0.016279339	L1
Molecular Function	GO:0019887~protein kinase regulator activity	0.030812232	L1
Molecular Function	GO:0050839~cell adhesion molecule binding	0.037913538	L1
Molecular Function	GO:0008093~cytoskeletal adaptor activity	0.044567764	L1
Molecular Function	GO:0050839~cell adhesion molecule binding	6.11E-04	LTR
Molecular Function	GO:0016772~transferase activity, transferring phosphorus-containing groups	0.00807089	LTR
Molecular Function	GO:0043169~cation binding	0.009748826	LTR
Molecular Function	GO:0005496~steroid binding	0.010019163	LTR

Molecular Function	GO:0000155~two-component sensor activity	0.012243649	LTR
Molecular Function	GO:0022803~passive transmembrane transporter activity	0.020931795	LTR
Molecular Function	GO:0001883~purine nucleoside binding	0.046141345	LTR
Molecular Function	GO:0030695~GTPase regulator activity	9.81E-11	MIR
Molecular Function	GO:0005096~GTPase activator activity	5.53E-07	MIR
Molecular Function	GO:0008092~cytoskeletal protein binding	3.32E-04	MIR
Molecular Function	GO:0022803~passive transmembrane transporter activity	6.24E-04	MIR
Molecular Function	GO:0043169~cation binding	0.00144037	MIR
Molecular Function	GO:0022891~substrate-specific transmembrane transporter activity	0.008004068	MIR
Molecular Function	GO:0005516~calmodulin binding	0.008041464	MIR
Molecular Function	GO:0046983~protein dimerization activity	0.015808567	MIR
Molecular Function	GO:0001883~purine nucleoside binding	0.023661931	MIR
Molecular Function	GO:0019992~diacylglycerol binding	0.026389336	MIR
Molecular Function	GO:0032553~ribonucleotide binding	0.028150117	MIR
Molecular Function	GO:0016772~transferase activity, transferring phosphorus-containing groups	0.035853993	MIR
Molecular Function	GO:0008013~beta-catenin binding	0.037247081	MIR
Molecular Function	GO:0050839~cell adhesion molecule binding	0.037247081	MIR
Molecular Function	GO:0017076~purine nucleotide binding	0.042483365	MIR
Molecular Function	GO:0008092~cytoskeletal protein binding	0.002789855	SVA
Molecular Function	GO:0050839~cell adhesion molecule binding	0.011650127	SVA
Molecular Function	GO:0030695~GTPase regulator activity	0.012213165	SVA
Molecular Function	GO:0005543~phospholipid binding	0.022869129	SVA
Molecular Function	GO:0016879~ligase activity, forming carbon-nitrogen bonds	0.031038513	SVA
Cellular Component	GO:0005626~insoluble fraction	0.002754261	Alu
Cellular Component	GO:0014069~postsynaptic density	0.01455023	Alu
Cellular Component	GO:0045211~postsynaptic membrane	0.016366668	Alu

Cellular Component	GO:0000267~cell fraction	0.023283704	Alu
Cellular Component	GO:0043005~neuron projection	0.037089783	Alu
Cellular Component	GO:0016020~membrane	0.038961333	Alu
Cellular Component	GO:0044459~plasma membrane part	0.039979154	Alu
Cellular Component	GO:0016020~membrane	1.03E-04	ERV
Cellular Component	GO:0030027~lamellipodium	2.77E-04	ERV
Cellular Component	GO:0044425~membrane part	4.90E-04	ERV
Cellular Component	GO:0012505~endomembrane system	0.003112417	ERV
Cellular Component	GO:0031252~cell leading edge	0.005978395	ERV
Cellular Component	GO:0042995~cell projection	0.008750212	ERV
Cellular Component	GO:0043005~neuron projection	0.010567625	ERV
Cellular Component	GO:0005886~plasma membrane	0.015943166	ERV
Cellular Component	GO:0014069~postsynaptic density	0.017268836	ERV
Cellular Component	GO:0031224~intrinsic to membrane	0.018157998	ERV
Cellular Component	GO:0044459~plasma membrane part	0.019160607	ERV
Cellular Component	GO:0031965~nuclear membrane	0.019851813	ERV
Cellular Component	GO:0031300~intrinsic to organelle membrane	0.022330112	ERV
Cellular Component	GO:0045211~postsynaptic membrane	0.031613444	ERV
Cellular Component	GO:0019898~extrinsic to membrane	0.034741239	ERV
Cellular Component	GO:0005737~cytoplasm	0.039586133	ERV
Cellular Component	GO:0044459~plasma membrane part	1.28E-10	L1
Cellular Component	GO:0044425~membrane part	6.73E-09	L1
Cellular Component	GO:0016020~membrane	3.08E-08	L1
Cellular Component	GO:0043005~neuron projection	1.73E-06	L1
Cellular Component	GO:0005886~plasma membrane	6.19E-06	L1
Cellular Component	GO:0042995~cell projection	1.89E-05	L1
Cellular Component	GO:0045211~postsynaptic membrane	5.98E-05	L1

Cellular Component	GO:0014069~postsynaptic density	7.58E-05	L1
Cellular Component	GO:0044463~cell projection part	8.04E-05	L1
Cellular Component	GO:0005578~proteinaceous extracellular matrix	0.00100319	L1
Cellular Component	GO:0019898~extrinsic to membrane	0.001395374	L1
Cellular Component	GO:0042734~presynaptic membrane	0.001400923	L1
Cellular Component	GO:0043197~dendritic spine	0.001680735	L1
Cellular Component	GO:0031012~extracellular matrix	0.001904228	L1
Cellular Component	GO:0005875~microtubule associated complex	0.003288719	L1
Cellular Component	GO:0044431~Golgi apparatus part	0.003519217	L1
Cellular Component	GO:0012505~endomembrane system	0.003851358	L1
Cellular Component	GO:0043198~dendritic shaft	0.008184011	L1
Cellular Component	GO:0000267~cell fraction	0.0254723	L1
Cellular Component	GO:0009986~cell surface	0.032262217	L1
Cellular Component	GO:0012506~vesicle membrane	0.040162796	L1
Cellular Component	GO:0044441~cilium part	0.04790736	L1
Cellular Component	GO:0044420~extracellular matrix part	0.050716684	L1
Cellular Component	GO:0009897~external side of plasma membrane	0.052729331	L1
Cellular Component	GO:0005581~collagen	0.057803073	L1
Cellular Component	GO:0035085~cilium axoneme	0.061178965	L1
Cellular Component	GO:0044430~cytoskeletal part	0.086692602	L1
Cellular Component	GO:0008287~protein serine/threonine phosphatase complex	0.091900824	L1
Cellular Component	GO:0005604~basement membrane	0.09873182	L1
Cellular Component	GO:0005626~insoluble fraction	0.047082603	LTR
Cellular Component	GO:0030027~lamellipodium	0.063143676	LTR
Cellular Component	GO:0043005~neuron projection	0.079788863	LTR
Cellular Component	GO:0005929~cilium	0.088069784	LTR
Cellular Component	GO:0045211~postsynaptic membrane	0.099886093	LTR

Cellular Component	GO:0019898~extrinsic to membrane	6.70E-05	MIR
Cellular Component	GO:0043005~neuron projection	8.70E-05	MIR
Cellular Component	GO:0042995~cell projection	1.68E-04	MIR
Cellular Component	GO:0044459~plasma membrane part	6.25E-04	MIR
Cellular Component	GO:0005886~plasma membrane	0.001138769	MIR
Cellular Component	GO:0034702~ion channel complex	0.001420407	MIR
Cellular Component	GO:0030027~lamellipodium	0.00671882	MIR
Cellular Component	GO:0044425~membrane part	0.009441945	MIR
Cellular Component	GO:0031012~extracellular matrix	0.00961077	MIR
Cellular Component	GO:0005578~proteinaceous extracellular matrix	0.010895064	MIR
Cellular Component	GO:0016020~membrane	0.015034447	MIR
Cellular Component	GO:0031252~cell leading edge	0.018580787	MIR
Cellular Component	GO:0016459~myosin complex	0.020015626	MIR
Cellular Component	GO:0005737~cytoplasm	0.02689117	MIR
Cellular Component	GO:0042734~presynaptic membrane	0.031856375	MIR
Cellular Component	GO:0030426~growth cone	0.040758303	MIR
Cellular Component	GO:0030427~site of polarized growth	0.043152743	MIR
Cellular Component	GO:0043198~dendritic shaft	0.043472116	MIR
Cellular Component	GO:0045211~postsynaptic membrane	0.044889397	MIR
Cellular Component	GO:0002142~stereocilia ankle link complex	0.047579822	MIR
Cellular Component	GO:0002139~stereocilia coupling link	0.047579822	MIR
Cellular Component	GO:0002141~stereocilia ankle link	0.047579822	MIR
Cellular Component	GO:0032420~stereocilium	0.049378586	MIR
Cellular Component	GO:0005622~intracellular	2.59E-04	SVA
Cellular Component	GO:0044424~intracellular part	0.001178181	SVA
Cellular Component	GO:0043229~intracellular organelle	0.001696152	SVA
Cellular Component	GO:0043231~intracellular membrane-bounded organelle	0.003537599	SVA

Cellular Component	GO:0005737~cytoplasm	0.009380962	SVA
Cellular Component	GO:0000776~kinetochore	0.033366997	SVA
Cellular Component	GO:0044431~Golgi apparatus part	0.043688535	SVA
Cellular Component	GO:0044444~cytoplasmic part	0.044648231	SVA

Table 2. Cell lines source and kinds that used to analyze the function of retrotransposons in the human genome

Cell	Tier	Description	Lineage	Tissue	Karyotype	Sex	Documents	Vendor ID	Term ID	Label
GM12878	1	B-lymphocyte,lymphoblastoid,international HapMap Project-CEPH/Utah-European Caucasion,Epstein-Barr VIrus	mesoderm	blood	normal	F	ENCODE	Coriell GM12878	BTO: 0002062	GM 12878
HepG2	2	Hepatocellular carcinoma	endoderm	liver	cancer	M	ENCODE	ATCC HB-8065	BTO: 0000599	HepG2
HMEC	3	Mammary epithelial cells	ectoderm	breast	normal	U	Bernstein Crawford stam	Lonza CC-2551	BTO: 0002178	HMEC
HUVEC	2	Umbilical vein endothelial cells	mesoderm	blood vessel	normal	U	Encode	Lonza CC-2517	BTO: 0001949	HUVEC
NHLF	3	Lung fibroblasts	endoderm	lung	normal	U	Bernstein stam	Lonza CC-2512	BTO: 0000161	NHLF
K562	1	leukemia, "the continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC	mesoderm	blood	cancer	F	ENCODE	ATCC CCL-243	BTO: 0000664	K562

Supplementary Materials

Supplementary Material S3:

Figure 12: The tree we used to build the Alu consensus sequences

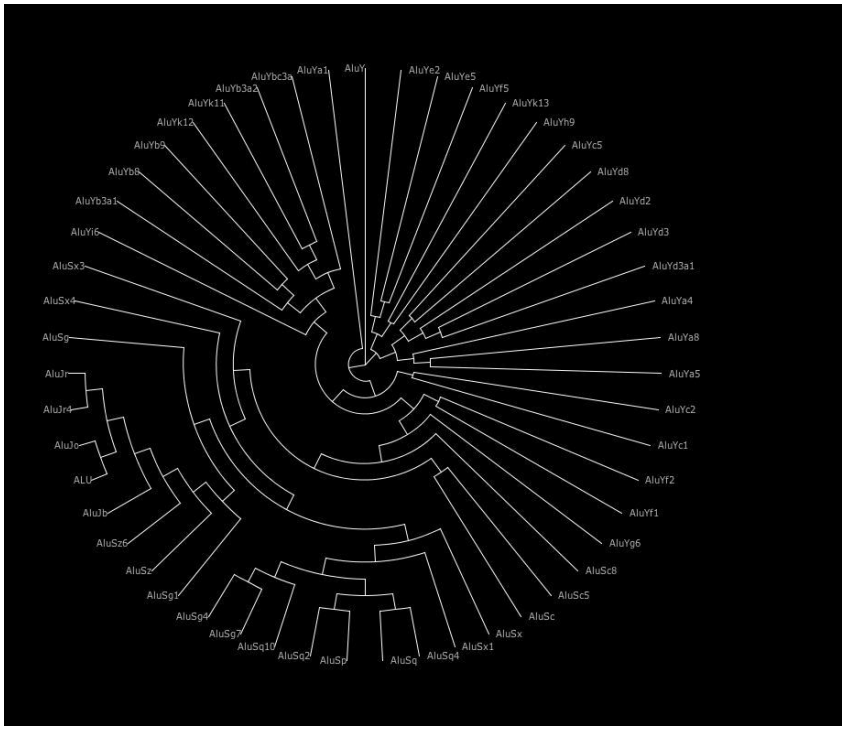


Figure 13: The tree we used to built the Mir consensus sequence

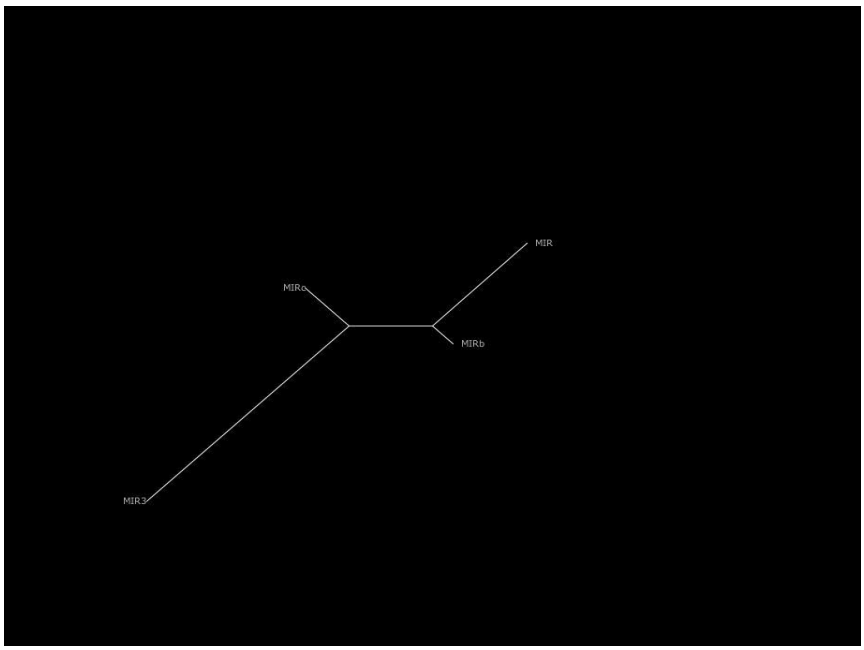


Figure 14: The tree we used to built the L1 consensus sequence

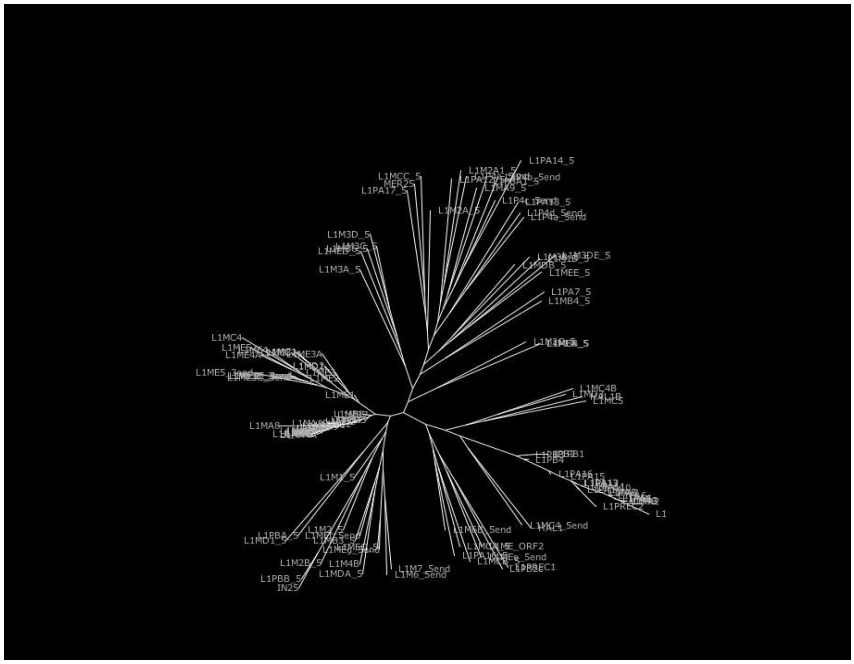


Figure 15: The tree we used to built the LTR consensus sequence

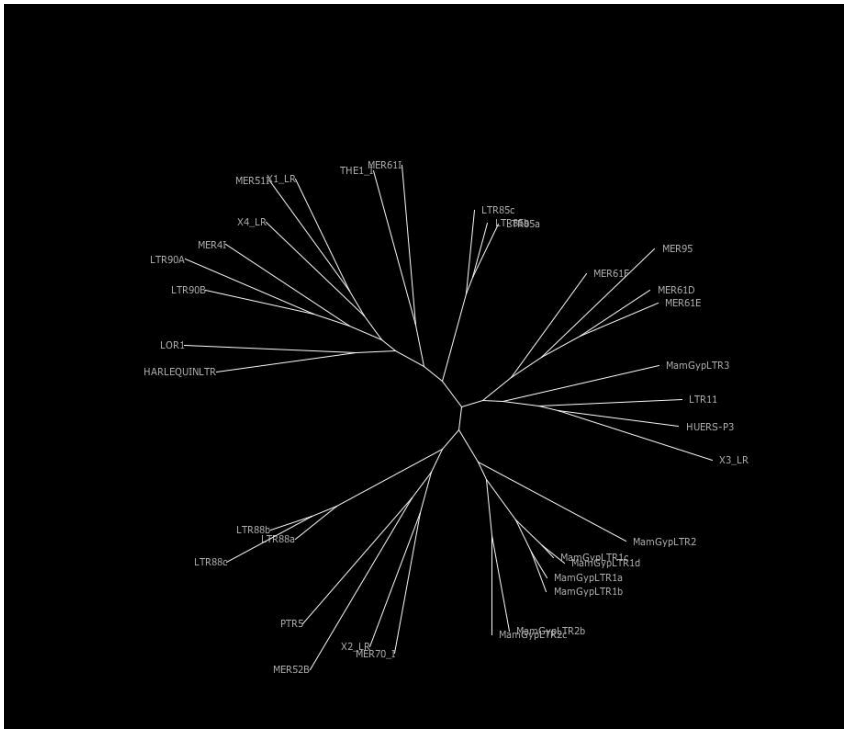


Figure 16: The tree we used to built the ERV consensus sequence

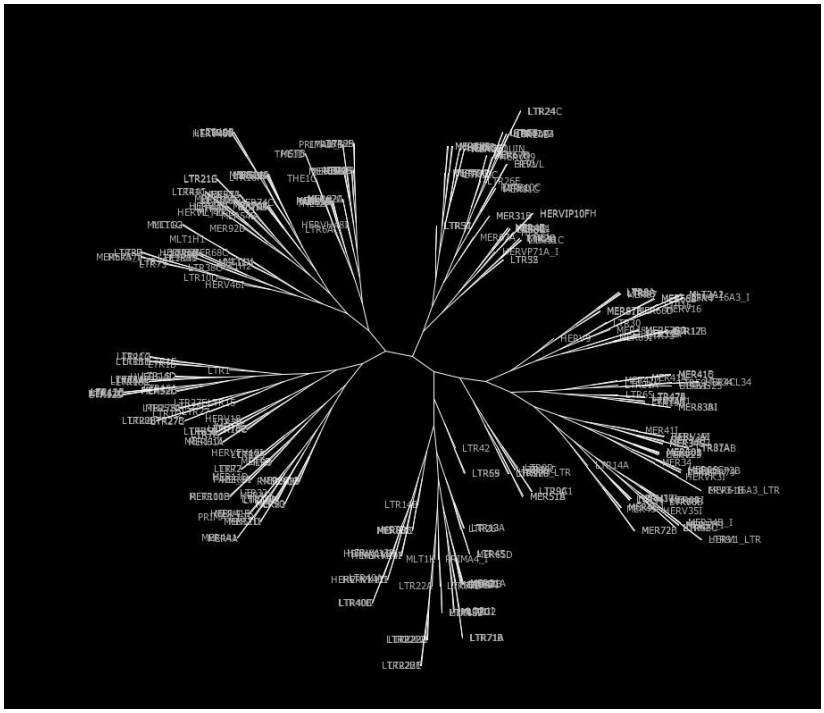
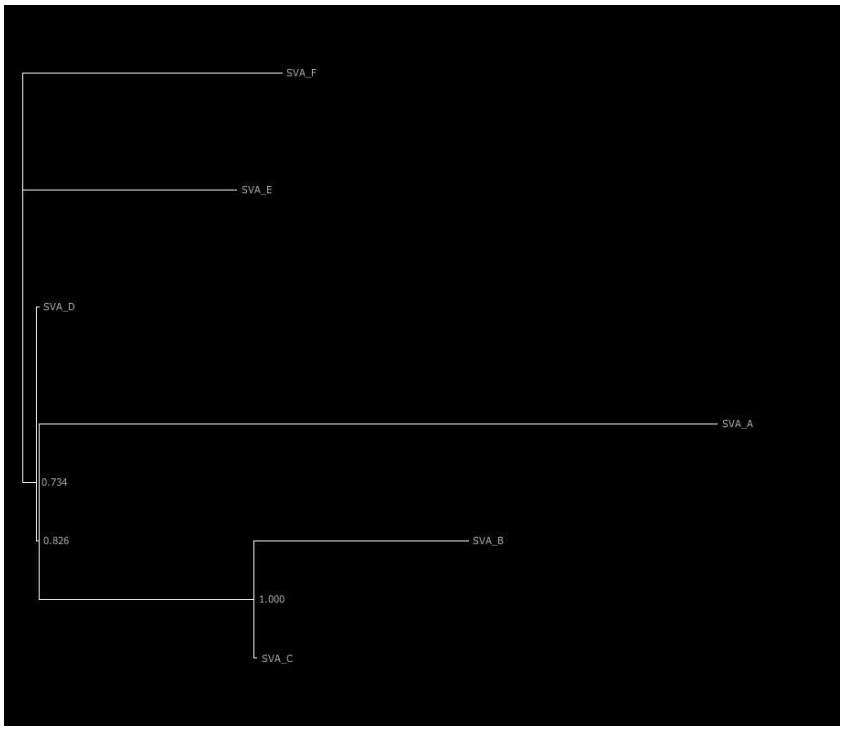


Figure 17: The tree we used to built the SVA consensus sequence



Supplementary Material S2:

Alu consensus sequences:

>1_cons
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA
TCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACT
AAAAATACAAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGA
GGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGC
GCCACTGCACTCCAGCCTGGGCGACAAGAGCGAGACTCCGTCTCAAA

>2_cons
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA
TCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAA
AAATACAAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAG
GCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATTGCG
CCACTGCAGTCCGCAGTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCA

>3_cons
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGA
TCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAA
AAAATACAAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGAG
AGGCTGAGGCAGGAGAATGGCGTGAACCCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATC
GCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCA

MIR consensus sequences:

>1_cons
CAGAGGGACAGCATAGCACAGTGGTAAAGAGCACGGACTCTGGAGCCAGACAGACCTGGG
TTCGAATCCCGGCTCTGCCACTTACTAGCTGTGTGACCTTGGGCAAGTCACTTAACCTCT
CTGAGCCTCAGTTTCCTCATCTGTAAAATGGGGATAATAATAGTACCTACCTCACAGGGT
TGTTGTGAGGATTAAATGAGATAATACATGTAAAGCGCTTAGAACAGTGCCTGGCACACA
GTAAGCGCTCAATAAATGGTAGCTCTATTATT

>2_cons
TTCTCGAAGCAGTATGGTACAGTGGAAAGAACAACTGGACTAGGAGTCAGGAAGACCTGG
GTTCGAGTCCTAGCTCTGCCACTAACTAGCTGTGTGACCTTGGGCAAGTCACTTAACCTC
TCTGAGCCTCAGTTTTCCTCATCTGTAAAATGAGGATAATAATACCTGCCCTGCCTACCT
CACAGGGTTGTTGTGAGGATCAAACGAGATAATCTATGTGAAAGCGCCCTGCAAACTCTA
AAATGCTATACAAATGTAAGGGGATACTATGATTCTAAAAAAA

L1 consensus sequences:

>1_cons
GGGGGGAGGAGCCAAGATGGCCGAATAGAAACAGCACCGGTCTACAGCTCCCAGCGTGAG
CGACCACAGAAGACGGGTGATTCCTGCATCGCCAACTGAGGTACCAGGTTCATCTCACTA
GGAAGTGCCAGACAGCGGGCGCACCCACAGACCCTCTGAAGGAAGCGGACTGCTCCTGCA
GGACCCGGGAGACACCCCAAATACTGTGAGTGCCCAAACTGCGGAAGTGGGAAAGGGAGA
TCCTCCGCTCCCGAACACACACCCCCACTGGGGAAACTGAAGGTCTAGTTTGCGGGAGAA
GTTTCCGACCTTACCTGGAGCTGAGTCAATTTAGAGAGCCGAGCGAAATACAGGGGTAGA
GGAAGCAGCGAGGAAAGGCCCTGGGAGCTCGCTGGGTCCCCAAGCAGGCCATTCCTGCCT
GGCACCACAGGGATCCTTCGGGAGGGCGGACAGAGGAGCGAGCGCACCGAGCGCAAGCCG
AAGCAGGGCGAGGCAGCGCCTCACCTGAGAAGCGCAAGGGGTCAGGGAATCCCCTTTCCC
AGTCAAAGAAAGGGGGGACGGACGCCACCTGCAAAATCGGGTCACTCCCGCCCTAACAAT

GCGCTTTTCCGACCCACTTAAGAAACGGCGCACCACGAGAATATACCCCACAGGGGCCTA
GGGTCCCAACCCTGGAGCCGCGCAGATTCTCAACAGCCTCTCAGCTGGAATCTGCTTAAG
CCTGCCGAGCTCCTGGCTCGGAGGGTCCTACGCACACGGACTCTCGCTGATTGCTAGCAC
AGCAGTCTGAGATCAAACTGCAAGGGGCAGCAGCCAGCACTGGGACTCATAACTGCCTAA
CACACTAAGCTCCCGGCAACGAGGCTGGGCGAGGGGCGCCCGCCAATGCCCAGGCTTCCC
TAGGTAAACAAAGCAGCCGGGAAGCTCGAACGGGGTGGACCCCACCACAGCTCAAGCACA
CCTCCATGCACCTGTAGGCTCCACCTCTGGCCGCAGCGCACAGACAAACAAAAACACAGC
AGTAACCTCCGCAGACTTAAGTGTCCCTGTCCGACAGCTTCGAAAAGAGCAGTCGTTCTC
CCAGCACGCAGCTGCAGATCTGACTGGGCCTGAGCCCCTAGAGGGAGGGGTGGCCGCAGT
CTCTGCGGACCAGCAGACTTAGCCTTTCCTCCTGGTAGTTCTGAGGAATCCGGGCAGCCC
AGATGAGTGGGTTTCCCCCCAGCGAAGCACACCCCCTGAACGAGCAGACTGCCACCTCAA
GTGGGTAAATGACCCCTGACCCCCGAGCAGCCTAACTGGGAGGCACCCCCCAGCAGGGGC
GGACTGACACCCCACACGGCAGCGATCCTACTGGCATCAGGTTGGTACACCTCGAAGACA
AAAATACCAGAAGAACGACCAGGCAACAAACTCTGCCGTTCTCCAATATCCACCACTGAC
ACCACCCAACGCGGAAGAGAACCAGAAAAACAGGGTCTGAAGTGAACCCCCAGCAAACTC
CAACAGACCTGCAGCAGAGGGTCCTGACTATTAGAAGGAAAACTAACAAACAGAAAGGAC
ATCCACACCCAAAACCCATATAAACATCACTGCAGCTCGGCTCACAGGAAGCCACATCCA
TAGGAAAAGGGGGAGAGTACTACATCAAGGGAACACCCCGTGGGAAAAAAAAACCCAAAC
AACAGCCAGCAGCATCAAAGACCAAAACTAGATAAAACCACAAAGATGAGAAAAAAACAG
AAAAGAAAAACTGGAAACTCTAAAAAACAGAGCGCCTCTCCTCCTCCAAAGGAACGCAGC
TCCTCACCAGCAACGGAACAAAACTGGACGGAGAATGACTTTGACGAGTTGACAGAAGAA
GGCTTCAGAAGATGAGTAATAAAAAACTACTCTGAGCTACGGGAGGAAATTCAAACCAAA
GGCAAAGAAGTTAAAAACTTTGAAAAAAAATTAGAGGAATGGATAACTAAGGGAATAACC
AATACAGAGAAGAACTTAAAGGACCTGATGGAGCTGAAAAACAAAGCACGAGAACTACGT
GAAGAATGCAGAAGCCTCAGTAGCCGAAGCGATCAACTGGAAGAAAGGATATCAGAGATG
GAAGATCAAATCAATGAAATAAAACAAGAAGAGAAGATTAGAGAAAAAAGAATAAAAAGA
AATGAACAAAGCCTCCAAGAAATATGGGACTATGTAAAAAGACCAAATCTACGACTGATT
GGTGTACCTGAAAGAGACGGGGAGAATGGAACCAAGTTGGAAAACACACTGCAGGATATT
ATCCAGGAGAACTTCCCCAACCTAGCAAGACAGGCCAACATTCAAATTCAGGAAATACAG
AGAACACCACAAAGATACTCCTCGAGAAGATCAACTCCAAGACACATAATTGTCAGATTC
ACCAAAGTTGAAATGAAGGAAAAAATCTTAAGGGCAGCCAGAGAGAAAGGCCGGGTAACC
TACAAAGGAAAGCCCATCAGACTAACAGCAGATCTCTCAGCAGAAACCCTACAAGCCAGA
AGAGAGTGGGGGCCAATATTCAACATTCTCAAAGAAAAGAATTTTCAACCCAGAATTTCA
TATCCAGCCAAACTAAGCTTCATAAGTGAAGGAGAAATAAAATACTTTACAGACAAGCAA
ATGCTGAGAGATTTTGTCACCACCAGGCCTGCCCTAAAAGAGCTCCTGAAAGAAGCGCTA
AACATGGAAAGGAACAACCGGTACCAGCCACTGCAAAAACACGCCAAAATATAAAGACCA
TCGAGACTAGGAAGAAACTGCATCAACTAATGAGCAAAATAACCAGCTAACATCATAATG
ACAGGATCAAATTCACACATAACAATATTAACCGAATAGTACCTCACATCTCAATACTAA
AATTAAATGTAAATGGACTAAATGCTCCAATTAAAAGACACAGACTGGCAAAGTGGATAA
AAAGTCAACAACCAACAGTGTACTGTACTCAGGAGACCCACCTCACATATAAAGACACAC
ATAAGCTCAAAATAACTGAAATATAAAAGGATGGAGAAAGATCTAGCCAAGCAAATGGAA
ACCCAAAAAAAAAGCAGGAGTTGCAATCCTAGTCTCAGACAAAACAGACTTTAAACCAACA
AAGATCAAAAAAGACAAAGAAGGCCATTACATAATGATAAAAGGATCAATTCAACAAGAA
GAGCTAACTATCCTAAATATATATGCACCCAATACAGGAGCACCCAAATTCATAAAGCAA
ATACTGAGAGACCTACAAAGAGACTTAGACACCCACACAATAATAGTGGGAGACTTTAAC
ACCCCACTGTCAACATTAGACAGATCAACGAGACAGAAAGTCAACAAGGATACACAGGAA
TTGAACTCAGCTCTGCACCAAGTGGACCTAATAGACATACTACAGAACTCTCCACCCCAA
ATCAACAGAATATACATTCTACTCACCAGCACACCACACATATTCCAAAATAGACCACAT
AATTGGAAACAAAACTCGCCTCAGCAAATGTAAAAGAACAGAAATCATAACAAACAATCT
CTCAGACCACAGTGCAATAAAACTAGAACTCAGGAATAAAGAAATTCCCTCAAAACCGCA
CAACTACATGGAAATTGAACAACCTGCTCCTGAATGACTACTGGGTAGAGGGGCCCTCTC
TGCTCCACGCCCAGGCAGATCTCCAGGCATCTGGAGCACCCACTCTCCTGAATAACGAAA
TCAAGGCGCCCCACCCTTCCCGTGCAGAGAACTTGAAATTAAGAAGTTCTTTGAAACCAA
CGAGAACAAAGACACAACATACCAAAAATCTCTGGGACACAGCTAAAGCAGTGTGTAGAG
GGAAATTTATAGCACTAAATGCCCACATGAAAAAGCAGGAAAGATCTCAAATCGACACCC
TAACATCACAACTAAAAGAACTAGAAAAACAAGAGCAAACAAATCCAAAAGCTAGCAGAA
GACAAGAAATAACTAAAATCAGAGCAGAACTGAAGGAAATAGAGACAAAAAAAACACTAC

AAAAAATCAATGAAACCAAGAGCTGGTTTTTTGAAAAGATCAACAAAATTGATAGACCAC
TAGCAAGACTAACAAAGAAAAAAAGAGAGAAAAATCAAATAGACTAAATAAAAAAAGATA
AAGGAGATATCACCACAGATCCCACAGAAATACAAACTACCATCAGAGAATACTATAAAC
ACCTCTATGCAAATAAACTAGAAAATCTAGAAGAAATGGATAAATTCCTCGACACATACA
CCCTCCCAAGACTAAACCAGGAAGAAGTAGAATCTCTGAATAGACCAATAACAAGCTCTG
AAATTGAGGCAGTAATTAATAGCTTACCAACCAAAAAAAGTCCAGGACCAGATGGATTCA
CAGCCGAATTCTACCAGAGGTACAAAGAGGAACTGGTACCATTCCTTCTGAAACTATTCC
AAAAAATAGAAAAAGAGGGAATCCTCCCTAACTCATTTTATGAAGCCAGCATCACCCTGA
TACCAAAACCAGGCAAAAGACACAACAAAAAAAGAAAACTTCAGACCAATATCCCTGATG
AACATAGATGCAAAAATCAGCTTTTTAGTGCCTCACTCAATAATAAATACTGGCAAACCA
AATCCAACAGCACATCAAAAAGATTATCCACCATGATCAAGTGGGCTTAATCCAAGGGAT
GCAAGGCTGGTTTAACATACGAGAAGAATATGCAAATCAAGAAAAGAAATTCATAAAATA
ATACAAGAAACAAAGTCAAAGACAAAAACCATATGATTATATCAATAGATGCAGAACTGA
AAGCATTAGACAAAATTCAACAACAATTCATGATAAAAACTCTCAATAAAATAGGTATAG
AAGGGACATATCTAAAAATAATAAGAGCTATTTATGACAAACCTACAGAAAACATCAAAC
TGAAAGGGCAAAAAATGAAAGCATTCCCAATAAAAACAGGAACAAGACAAGGATGCCCTC
TCTCACCACTCCTATTCAACATAGTATTGGAAGTACTGGCCAGGGCAATAAGGCAAAAGA
AAGAAATAATGGGTATTCAAGAAGGAAAAGAAGAAGTCAAATTGTCACTGTTTGCAGATG
ACATGATTGTATATCTAGAAAACCCCATCGTCTCAGCCCAAAAACTCCTTAAGCTGATAA
GCAACTTCAGCAAAGTCTCAGGATACAAAATCAATGTACAAAAATCACAAGCATTCCTAT
ACACCAACAACAGACAAGCAGAGAGCCAAATCATGAATGAACTCCCATTCACAATAGCTA
CAAAGAGAATAAAATACCTAGGAATACAACTTACAAGGGATGTGAAGGACCTCTTCAAGG
AGAACTACAAACCACTGCTCAAGGAAATAAAAGAGGACACAAACAAATGGAAAAACATTC
CATGCTCATGGATAGGAAGAATCAATATCGTGAAAATGGCCATACTGCCCAAAGTAATTT
ACAGATTCAATGCAATTCCCATCAAACTACCAATGACATTCTTCACAGAATTAGAAAAAA
CTACTTTAAAATTCATATGGAACCAAAAAAGAGCCCGCATAGCCAAGACAATCCTAAGCA
AAAAGAACAAAGCTGGAGGCATCACACTACCTGACTTCAAACTATACTACAAGGCTACAG
TAACCAAAACAGCATGGTACTGGTACCAAAACAGACATATAGACCAATGGAACAGAACAG
AGACCTCAGAAATAAAACCACATATCTACAACCATCTGATCTTCGACAAACCTGACAAAA
ACAAGCAATGGGGAAAGGATACCCTATTCAATAAATGGTGCTGGGAAAACTGGCTAGCCA
TATGTAGAAAACTGAAACTGGATCCCTTCCTTACACCTTATACAAAAATCAACTCAAGAT
GGATTAAAGACTTAAATGTAAGACCTAAAACCATAAAAACCCTAGAAGAAAACCTAGGCA
ATACCATTCAGGACATAGGCATGGGCAAGGACTTCATGACCAAAACACCAAAAGCAATTG
CAACAAAAGCCAAAATTGACAAATGGGATCTAATTAAACTAAAGAGCTTCTGCACAGCAA
AAGAAACTATCATCAGAGTGAACAGGCAACCTACAGAATGGGAGAAAATTTTCGCAATCT
ACACATCTGACAAAGGACTAATATCCAGATCTTGGTTTCTAAATACCAATCTCCATTAGC
AGACAAGGAACTTACAGAACAACTCCTTGGAGAAATGGCTGATTCCAGGACTCTTACAAG
AAAAAGACATGAACCAACAAAATGAACCTGGACAACCCCATCGTGCCACAAGCTAAAAAA
GTGGGCAAAGGACATGAACACCAAGACACTGTCTGATACACCAGTCAAACTATAGAAAAA
AAAAAAAAAACTTCAAAGATAAAGAAGACATAAGCCAACCTGAACATGCGGCCAACAAAC
AATATGAAAAAATGCTTTAAAAAGCAACATCACTAAGAAGATCATCAGACTTCTCAACAG
CGACAATGCCAATAGTTTGAAATCAAAACATCCAAGGGGGCACAATAGAAGTGAGAAAAA
GTACCATCTTCAAAATTCTCACGGAAAAAAACTAACCAACCAGTCATTCTATACCCAGCC
AAACTATCCTTCAAGAATGAAGGAGAAAGAATGGCTATTATTAGACAAGAAAAAAGTCAA
TAAAAAAAATTTACAAGCCTTTCCCAGATGAACCCTCTGTACAAAAGAGCCCCCTCTGAA
GAAACAAACATGGATACGCGAGGATGATACCTGCTACCACAAAAACACACTTAAGTACAT
AGCCCACAGACCCTATAAAGCTGCTCCACCAAAAAAACAACAGAGAAATAGTCCTAGGAA
AAGAACACAGGATTGAAAGGAACACACTTTTACAAAAGAAAGAAGGAATGAAGACCAAAA
CTGTTGGTGGTCTAAATGCCCCACTTAAAAGACACAGAGTGGCAAATTGGATAAAAAAAA
AAAAAACAAGACCCATCCATCTGCTGTCTTCAAGAGACCCATCTCGAATGTAATTAGAAT
TAGTGGATCAACCATTGTGGAGAAGCAGGATATTTGCCTAGACAAAAAGCAACCACCCAA
AAAGAAAAAAAAAACAGTGTGGCGATAACTAATCTAGACAAAAATATAACTGACATTTCT
CCCTCAAAGAACTAGAACATTAAAGAAGCTAGAAACCTGTACCATAAAAGTGGAATTGAC
CCAGAGAGGCCAAAAAAATCCCATTAAACTGGGTATATACAAAACCAACATCCGAGCACC
CAGAGGGAGGAATATAAATGGAGGAATCATTCTAATTCCAAGAAAAGACTTAGACAGCCA
AACAATAGACAGCCTGTCAATATAAAATAAGACACATTGCACACGTATGGCCTTTAAACA
AGTAAACAAGCCACTTGCACAATTGGCACAGAAATAGAGTATAGAAAAGTCACAATAGCA

TAATACAAGACATAAACCAATGGAACGCAAAAAACAAACTCACCCAAATGGTATTCATAC
CAAAGAACCATAACCTCAACCCATCAATAAAACCACATATATACAACTAACCAATATTTG
ATAGATACTCTAAGATTGACCACATGCTTGGCCATAAAGCAAGTCTCAATGGATAAAGAA
AAAAAATAACGTGGTACAATTCTATACATGCTTCCATATGCTCAGTGCTCATCGGAATAA
AACTATGCAGCCATAAAAAAGAGATCTCTCAAAACCACACAAGATAAGACAGATGAGATC
ATGTCCTAACCAAAAACATGAGAGAAAAAAATAAACAATATGCAGGGACTAAAAAAACAT
TTGAAATGGATGAAACACTTAAATGTAAAGCTGGAAACTGAACATAAAAACCCTAGAAGA
TATTCTAGAACAGCAAACTAACGCAGGACATTGGCATAGGCAAAGATTATTTGGATAAAA
CAGGAAAACCAATAACAACCGCCTGGAGATTAGATAATGTTATTGAATCAATGCTCACTT
ATAAGTATGTTGGGAGATGTAAAATTAAACAATGAGAGAATCCCCATCTTTTAAAGAAAC
ACACAAAAACCTTGGACACAGGAAGGGGAAAGCAACACACTGGGGACTTACTGTCGGA
GGGTGGAGGGAAAAAAAAATAAATAAAATGGGAGGAGGGTGAGGATAGCATTAGGATAAA
AAATACCTAATGGGAATTACAAAGCTTAATACCTGGGTATAAAAAGAACTCAGATGGGTT
AATGGGAATACAGCAAACCACCATGGCACATAGTATACATATGTAACAAACCTGCACGTT
CTGCAAGAAGATATACATGTACCCACCAAAACTTTTTATAATATAAAAATAAAATTTAAA
AAAAAAAAAAAAGAAAAGAAAATAAAAA

>2_cons
AATGTATTAAAAAATTAAAAAAAAAAAAAAGAAGAGAGTTCCGGTTAGAGGGAGGCCGGC
AATGGCGGCCGAGTAAGAGGCACCTGCAGCCCGCCTCTCCCACTAAAAGGAACACCAAAT
TGAACAACTAAAAAAACTAGAAAAAATCTGCACCTATAAAACAACAATTTGAGAGAAGAT
CACAGTACCTGGTTTTAACTTCATATCACTGAAAGAGGCACTGAAGAGGGTAGGAAAGAC
AGTCTTGAATCGCCGACGCCACCCCTCCCCCATCCCCCGGCAGCGGCCGTGTGGCGCGGA
GAGAGAATCTGTGCACTTGGGGGAGGGAGAGCGCAGCGATTGTGGGACTTTGCATTGGAA
CTCAGTGCTGCTCTGTCACAGCGGAAATTAACAAAGAGAAGGAACCAACCGCCACAAACG
GAGAGAAAATTCCAGGAAACACCAGAAACATGGAAAGAAACCGTCCCAGCGGTCGGAACC
TGAGTTCCGGCAAGCCTCGCCACCGCGGGCTAAAGTGCTCTGGGGTCCTAAATAAGAAAA
ACAGCCAGCCCAGCCAAGAGCGGCCGAGATCCAGGAGAGAGTCCAAAATCTGTGCATGCC
CAAAAGCCGGGAAAAGCAAAGAGGACGCGACCTAGTGAGACACCAGCCTGAGTTCCCCAG
CGGTCCACCGCGCAACACACTTTCTCACCACGTACCCCCCAGCTCTGAGCCACGGGAGAG
CCACCCAACCCGCACGCGACCTGAGACTAGCATAGGGAGCAGAGGACTTTGTCCTGGAAG
CCGGGAAGCCGCACTGCTACAGAGAGATAGGGCACCGAGCAACTTCATGAGGCACCCCAC
GCACCGCCAAAACCCCAAGCAGAGCTTGCAGCCTCCAGCTTGAGCAGAGGGAACCCAGCT
CCCACCATAGCACACAAGACTGCATCCTGCCCTGGGGCCCCCTTCCTCCACCCAGCCCCC
ACTCATAGAACCCAGGCACGACCCAAGCCCCGGCAGCCCGAGAATCCTGGGGCCCTGATT
GCTCTCCCCCAGAACCAACCAAGAACGCGCCAGGGGCACAAGACCAGCTGGACCCAGCAG
TGCAGCAGGGTCCCCAGCACTCTAGCCCACACAGTGTCCCCGCTCACAGGCGTGCTGGCT
TCAAGAATGAACCATCCAGTTCACCACTGTGGTAACTATGGCACCACACGCATCCACCCC
GACAAAAGGAGGGGGCACAAAAAGGGAACCAAAGCGGCTACACCTCCCAGCCTGGGAGCT
GCCTCTAAGGCAGGGGCGTCACCCACAGAGAGGACCCCTGCAGCCCCCAGCTGCAGGGCT
GCCACACACCTGCACGCACCCTGAGGACAGGCTCTCCCTACCTACTGCTGCTGCCAGTGC
AAAGTGTATGCTCCCCAGAGCCTGAGGACCACCTGCCTGGGGCTGCGCCCACGGACAGCA
ACCCTGCCCCGACCCTGGCCCCAACAACAGCGCACGGCCGTGAAGGCGGGGCAAAAGGCC
AGGACTCTCAAAAGCTCCCCCCCTGCCTACTAATGAGACTCGAGGCATAGAGTGGACAAC
CCCAAGAGCCTGACAGCACCCGCCCTGGGCCGCGGCCACTGGTAGCAACACTGCCCCCCC
CCCCAGCAGCAGGGCAGCTGAGCAACGGCTGGCACCCCGAGGACAGACTGCCCTTCCCAC
CACTGCTGCTGCCACCCAGACACTAAGGCCTTCCTCCAGCGGCCTGGGGATCAGCCCGGC
CCCACCAACAAAGGCCACCGGAAACACCCTGAGAGCACCACCCTGGAGGCCTGAGAACCG
GACCCCCCGGCCCGACTCCAGCAACACCACTCCCTATCCGTGCCGGCTTGGGGCCAGAGC
AAGCCTAACCCTGCCCCGATCGGATGGTCTTTCTCTACCAACCCTGGTAGCAGAAGACAA
AAGACATAATCTCTTGGGAGCTCCCCAGCCCAGCCCACCACCACTGGAACCTAAGCACTC
ATCCCGCGAGCCTGAGGGCAAGGCCACAAAACCCACCACCACAAAGCGGCAAGAAACCAC
CACAACTGGCCCCCTCCTGAACGTGCCACCCAGGGGCCTGGGGACTGGCCCGCCCAGCCC
ACCACAGACAAAGAAGAGAGCAGTAAGGAACCACTCGGGCGCCAGAGGAGGGGCCCACAA
CTGCTAATGCCACTGCCCACGCCATGCCTGCTACCCAGGGGCCCGAGAACCTGCCCACCC
AGCAGGCCCACTGCTACCACTGCTCCAACCCGAGAAAGCCCACCGCAGGCCTAGGGACCC
CGTTTGTTTGGGGCCAGCCAAAAGAAGAGAAAACCAGTGCCAGCCTAGGAATCCCGAGAG

CCCCAGCACAGACCAACCAAGCCCACCGCTGCCGCCCCCAGACCAGTCCCCAAGAGCCAC
AGCTGAGAGACCCATAGATGGTTCACATCTACAACCAAGGACCCTCACAGAGTCCACTTC
ACTCCCCTGCTACCTCCACCGGCCCAAGAACTGGCATACCTGGCTTCCAAATCCACAGAA
AAACAACATCACAGAAATCCTACAACACACCCCCCAAATAACTACCAAAGCCTAGAAACT
CCACTGGGTGGCTAGAGCCACTGAGGAAATCACAGACACCACTGATACTGTTTAAAGACG
AAAAAAACCTAAAGGAACGGGGAGAGCACCACATCAAGGGAGCAACCCCCACAATGCAAG
AAGGCAAAAACCAGATCCAGAGTCTCCTATCCTCCCTACAACATAGGTACACCTAAAGGA
AAAATGTCCTCCCCTACGAAAGCAAATTCAAAAAATTGGAAGAAGCAACTGTTACACCAG
ATGCACAGATTTCAATAAAAAGAAACAAAAAATTACAAGAAATACGGAGGAAAGAAACAG
GAAAAAGGACTACTATGACCCCTCCAAAGAAGCAGAAGAAAATCAATCACCAGAAACAGA
CCCCGAGGAGACACAGAATCGTTAGTATTTGAATTACCAGACAAAGAAACTTTAAAATAA
CTATTATAAAAGATGCTCAATGAAATAAAAGAGAAACATGGATAAAAAAAAAAAAGGAAGA
CAGAAAAACAATAAAGGATAAGATCAGAGAAATTCAACAAAGAGATAAGAAATTATAAAA
AAGAACCAAACAAACAAGAAATTCTAAGGAACTGAAAAATAAAACAATAACTGAAATGAA
AAAACATATACACTGGAGGGGCTCAAAACAGCAGAATTGATACAATTGCAGAAGAAAAGA
ATCAGTGAACTTGAAGACAGGTCAATAGAAATTATCCAATCTGAAGAACAGAAAGAAAAA
AAAATTTAAAAAATTAAAATGAACAGAGCCTCAGAGACATGTGGGACACCATCAAGAGAT
CCAACATATACATACATGGACAGACAAAAAACATATGAATTATTGGAGTCCCAGAAGGAG
AGGAGAAGAGAGAGAAAGGGGCAGAAAAAATATTTGAAGAAATAATGGCTGAAAACTTCC
CAAATTTGACCCTGAAAAGACAATAATATTAATATACAGATAGATTCAAGAAGCTCAACG
AACCCCAAGCACACCTGGGAGAATAAACCCAAAGAGATCCACACCAAGGCACATCATAGT
CAAACTGCTGAAAACCAAAGACAAAGAAAAAAAAAGAATCTTAAAAGCAGCCAGAGAAAA
AAAGACACATTACCTACAAGGGAAAAAACAACAATAAGACTAACAGCAGATTTCTCAGCA
GAAACCTTACAGGCCAGAAGACAGTGGAACGACAATTCTTCAAAGTGCTGAAAGAAAAAA
AAAAAAAAAAAAAAACAAAAAAACACAGAATTCTACAAATAAAAAAAAAACTGTCAACCA
AGTAAATTCTATATCGAAGCAGCAAAAATAATCCTTCAAAAATGAAGGAGAAATAAAGAC
ATTCTCAGATAACGACAAAAGCTGAGAGAATTCATCACCAGCAGACCTGCAAACCTACAA
GAAATGCTAAAGGAAGAGTTCTTCAGGCTGAAACGAAAGGAACACAACAAACAAAATCAA
AACAAAAGAAAAAAAAAAACACACACACACAAAAAAAACACAAAAACAAAAGAAAAAAAC
GCAGAAAGACAAGAAAGGAATTATGAAAGAAATCAACAGAAAAAAAAAAAAAAGAAAAAA
AACAGAAGAAAGCCTACGGCACTTAACACACCAAAAAAGAAATAAAAAAAAACAATATAG
AAATAACAGAAGGAAAAGACATAGAAAAAAGAAAAAGAAAACATATTTAACAAAAAAAGA
ACAAAAAACGTCCCAAGGAGCAAAACAGAAATGAACATCCAGATCAGGAAGCTCAAAGAT
CCCCAATTAGATTCAACCCAAAAAGATCCTCTCTGAGGCACATTATAATCAAACTGTCAA
AAGTCAAAGACAAAGAGAGAATTCTAAAAGCTGCAAGAGAAAAGCATCAAGTCACATATA
AGGGAATCCCCATTAGACTATCAGCAGATTTCTCAGCAGAAACTTTGCAGGCCAGGAGAG
AATGGGATGATATATTCAAAGTGCTGAAAGAAAAAAAAAATAAAAAAAAAAAAAACTGTCA
GCCAAGAATACTATATCCAGCAAAGCTATCCTTCAGAAATGAAGGAGAAATAAAGACCTT
CCCAGACAAGCAAAAGCTGAGGGAATTCATCACCACTAGAACTGGCCTTACAAGAAATGC
TTAAGGGAGTGCTACAACTGGAAAGGAAATGATACACAGATAGAAACATGAAACCACAGG
AATCATGGAATAAAAGACTGGCAGTGATCACCGGAAACAGGTAAATACATGGGTAAATAT
AAAAAAATAATAATACTGTAATTATGGTATGTAATTCTTTTTTTTCTTTCTTAAATTCTC
TAAAAGACAAATGAATGTTTAAAGATAAAATGTACAAAAACAAAAATAACTATAACAAGT
GTCTGTAACTCCACATTTTGTTTTCTACATAATTTAAGAGACTAATGCATTTAAAAAAAT
TATTAGATAAGGATGTGGGGCATATAATATACAAAAATGTAGATATAATAAGACTGTAAA
ATGTATGACAACAATAACACAAAAATGGAAGTAGGGGGGGGGAAATTAGAAAGGAGAGTT
TTTTATAGTGTTGTAAGGTTTTTACATTTTATACAACTGGAAGTAAGATAATATCAATTC
AAAGTAGACTGTGATAATTAGAGAAATATAGGGTTAAGGATGTATATTGTAATCCCTAGA
GTAACCACTAAAAAAATTATAAAAAAAAAAGATATACAAAAAAAAACAATTAAAGAAATT
AAAAACGTAACAATAAAAAAAACATTCAAATAAACCAAAAGAAAGCCAGAAAAAAAGAAA
AAAAGGAACAAAAAACAGATAAGACAAAAGACAAACAGAAAACAAATAACAAAATGACAG
AAATAAGTCCAAACATATCAATATAACATTAAATGTGATTATGGATTAAAAAAAATGGCA
AAAGCTGTCAGACTAGAGATTTAATATATATAAATCCAAATAAATAGTTAAAATGATAAG
ACAGATAATACAAATATCAATAATAGGCTACATTAAATGTAAATGGACTAAACTCTCCAA
TTAAAAGACAAAGAGATTGTCAGAATGGATAATTAAAAAAAATAAAAAACAAGACCCAAC
TATATGCTGTCTACAAGAGACACACTTCAAATATAAAGACACAAATAGATTGAAAGTAAA
AGGATGGAAAAAGATATATCATGCAAACGGAAACCAAAAGAAAGCAGGAGTAGCTATACT

AATATCAGATAAAATAGACTCTTCAAAACAAAAAATATAACAAGAGACAAAGAAGGTCAT
TATATAATATAATGATAAAAGGGATAAAGGGGTCAATCCAACAAGAAGATATAACAATTA
TAAACATATATGCATAAATATATATGCACCAAACAACAGAGCCCCAAAAATACATGAAGC
AAAAACTGACAGAAATGAAAGGAGAAATAGACAAATCAACAATAATAGTTGGAGACTTCA
ACAACCCACTCTCAACAATGGATAGAACAACTAGACAGAAAATAAGAAAAGAAACAAAAA
AACACAACAATACAACAAAACAATAAACCAACTAGACCTAACAGACATCTAAAGAACATT
TATAGAACACTCCACCCAACAACAGCAGAATACACATTCTTCTCAAGTACACATGGAACA
TACACCAAGATAGACCATATCCTAGGCCATAAAACAAACCTCAACAAATTTAAAAAAAGG
AAAAAAATAAAAAAAAGGATCTCCTACCACAACAAAAGAAAAAAAGAAAAAAACAACAAA
AAAAAAACAGGAAAATCTACAAACACGTGGAAACTAAACAACACACTCCTAAAAAACCAA
GGGGCAAAAAAAAAAACCAAAAGAAAAATAAGAAAATACTTTGAGATGAATGAAAATGAA
GACACAACATACCAAAATTTATGGGATGCAGCTAAAGCAGTGATTAGAGGAAAATTTATA
GCTGTAAATGCCTATATTAAAAAAGAAGAAAGATCTCAAATCAATAACCTAACCTTCTAC
CTTAAGACACTAAAAAAAGAAGAGCAAACTAAACCTAAAGCAAGCAGAAGGAAGGAAATA
ATAAAGATTAGAGCAGAAATTAATGAAATAGAAGAAAAACAATAGAGAAAATCAATGAAA
CCAAAAGCTGGTTCTTTGAAAAGATCAACAAAATTGACAAACCTTTAGCTAGACTGACCA
AGAAAAAGAGAAGACTCAAATTACTAAAATCAGAAATGAAAGAGGGAACATTACTACTAA
CCTTACAGAAATAAAAAGGATTATAAAGGAATACTATGAACAATTGTATGCCAATAAATT
AAGATAACTTAGATGAAATGGACAAATTCCTAGAAAAAAAGACACACAAACTACAAAAAC
TGACTCAAGAAGAAATAGAAAATCTGAATAGACCTATAAAAATAAAGAGATTGAATTAGT
AATATAAAAACTACCAACAAAAAAAGCCCAGACCCAGATGGCTTCACTGGTGAATTCTCC
AAAAATTTAAAAAAGAATTAATACCAATTATTCACCTATTCCAAAAAATAGAAGAGGAGG
AAAAACTACCAAACTAATTCTATGAGGCCAGTATTATCCTGATACCAAAACCAGACAAAG
ACATAACAAAAGAAAAGAAAA

>3_cons
CTTGGCACTTAATTCACCAACTCTGGGAGCAGAGGGGATTGGACATATGAAAAGAGAGTG
ACGTCAGCAAAAAGGTAGTCTTTTAGCCAAGATGGCGGACTGGAGGAGCAGCCAGGGTCC
GCCCCCCCCCAGGAGCAGCACAGAGAAAAACCAAAAAAATCGAGTTCCCTGTTTCTCCCTG
GAAGGGGTTTATGAAACATTCACACTGCGAACACTTGCCAACAAAATCATCTAAGTGAGA
GCACTGGGGATTCAGCAAAGAGGCGACGGCAACCAGGTGGAGCACCAGAGACTGGAAAAG
ACCCAATGAAAGAGGAAGGAAGAACTGGATTACATTACCCACGGCACCCCTCCCCCACCC
GGGAACAGCACGAGAGCCAGGGGACACTCTCCCGGCCCCCAATGGCTCCTGCACGGGAAA
GGGTGAGGGACTGTGCCATGAGGGAGCCCCAGCGCGGGGCCAGAGACCCCACTTCCCCCA
CGGACCTAGCTGCAATCCTGGCCACAGGAGACCCCCCCCAGCCCCCCCCGACCACCAGGGTA
CCCACGCACAGAGCCGAAACAGACCCCAAAGACAAAGCCTCTGGCCTGACACAGGGAGCA
CAAGGTGTGCACTCCCCACCCCCAGTGACTCAAGCTGCTGCAGCACAGCCACCCGCTGCC
CGGGAATGCTGCAGACTCCCTAGACCACAGGCAGAGCCAGAGCTCCAGCAGGCGGAGAGG
AACATGCGCACCAGCCAAACCCCACAGGCCACACGGCTACTGCGCTACGCCACCTCGGAA
CTGCCCCTTGACCCCGGGGGAGCCTCAGCATCCCGGCAAAAAGACATGCAACCCCGCCAA
AACCGAAGCTAAGACGCCCCTGAACCACCCCAGCCAGCGAGCCTAGCCCCCGGAGCACGG
CAAGCAACGCCCATCCCCGCACTTCCCCCCAAGGCGCGACAGCCGAGAACCCGCACCACC
TACAACTTCCAGTCAAGCTACACAAACAGCGGCATACCCCCCCAGGACAGAGCTAAAGGG
GGCTCTGGCCCCCGCCAGCTGCCGCACCACGGGCACCCAGGCTGGCCCCTGTGGAATCTT
GCCTGCGGCACACGGAAGATCAAACCTGACCTTGTGGGGAAGGGATCCCCCAGCACAGCA
CAGCTGCTCTACCAAAACGTGGCCAGACTGCTTCTTTAAGCGGGTCCCAGACTCAGCACC
CCCCCCCCGGGCAAGACCACCTCCCCTGGGGCCAGAGCCTACCCCGCCCAGCCCCCACCG
CAGCCTCGGAGTTGAAACCTCCCTGGGACGGAGCGCCCGGCCGGGGGGGAGGCGCTGCACC
TGCCCCTTTTCCCGGGAAGTACACGAACCGCCCATCCCCATGTCGGGGGCCTGAGCACAC
CGGCCGCAAGCAACCCCCCACCCACCCGGGCGGCAAAACTCTGGGCCAGTGGCCCAGCCA
ACCCCGTACCCGTATCACAGCCACAGCGCCCACCCAAGGGGCCCGCCCTCCAGGCCGAGG
CCCAGACCCCCTCCCCCCAGAGTTCGACCACACACACAAAGGGTAAACAGCCCAGAGCCG
GGGCCCCAGAAGAACCGGGGAAAGAGCCTCCCCCAGTTCCCTGCACAGCCAGCACACCCC
CCCCCACCCCCCCCGGGCAGGCCGGTGCCAGTGGCCTGGGGGCGCAGCACAGCCCCCACGC
CCCCACAGCTGCGATCAAGCCAGGGCCCCAGACCCCTTTTCTGGGAAGGCCACAGGCCCC
TCCCCACCCGGAACTCCCCTCCCCTGCCCAGAGAGGCCCCACACGCGGGGACTCTCCCGC
CGGCCCGTGACCCAGCACTGAGCCCAGTTGCAGCTGCACCACCCTCTGGGGCCGCACCCA

CAGAGAGGGGCCCGAAACTCCAGCGGCCCTATCCCTGCCCCAGGCCAGCTGAAGGCTCCA
GGTACTGGAAAATCCGAGGCGACTAGGGACTGGAGCGGGCCCCCAGCATACCGCAGCAGC
CCTACGGAAAAGTGGCCAGACTGTTACGTGGGTGCCCGTTCCCATATCTCCTCACCGGGC
AGGTCCTCCAGGCCTGGGCCTCCAGCCACCCCCCGCCAGAGCTATCGAGCCAGTAGCAAC
TCGGCAACTCCCTGGACAGAGCCTCCAGGGGCAACTGGAATGCCCTCTGCCCCCACCCCA
CAGCCGTCTGCCAGACAAAAAACCACGCCACCCAAGCAGCACACCAGCTATACACACCCA
CGAACTACCCCCCGACCAACGATCCAACGACAGTCACCCACCCCCCCACCAAACCCCACT
CACACCAAGGCCACCACGAGCCACCGTTCCGCCCGCCCCGGACAACCTCCGTGCACCCCG
GCCCCAGAGAGGCACCCGCTGGCCCGAGCCTCAGGCAGGAGCACCACTAAGGAAGGGGGG
AGTGCAAACAGCCCTGCACCCCCTGGCTGCCGGCCACAGGCGATAGCGGTCCCACCTATC
GGAGGAGCACCCAGATGCCACGCCAGGCGACCCCCAGCACCCCTACCTCTAACAACAAAA
GAAAACACACCGGCCAGAGTAACAGGCCCAGAGCGGGCCCCATTGCCCGGCCCGCCTCCC
CGGCTGCCACCAACAGGCCCAACCAAAATAAACACTGAGATCGCCCCAGAGCTGCAGTGG
GCAGCCCAGGAGTGCCAAGCCACGATCTGCTGCCAGCAGACAGGCAACAAAGTAAACCCC
CCCCCAAAAAAATAAAGAGCAAAGCTGACCCCCACATTAGACAAAAAATGGAACAAGCAT
CTGACAGCCCTGACTCTTCCCACAAGCGGCCGCTCAGGACAAAGACCCCAAACTTCAACT
CTGCCACCCAAGCACACACCAGAGCACCAAGGCCAAGAAACCTGTGCTGACCCAGCCCCC
CCTGAAACCAAGGACAGGAAATTAGCCACAAATAAAGATCCTGCACAGAGCCTTGGCCCT
CTGAAAGCACCCAGAAAAGAAGCCAACAAACTCAACCCAACTTACACAGCCTTCCAAAAT
CAAAGAAACCCCCAAGGAGATCCAAGAATACAGCACAAACAAACAGGAAAGCGAGAAAAA
AACTCCACCAAAAGAACACACACTGCCTAAAAAAGACTGAAATGCTGAGCCCCTGAGTAC
CTGATGTTAAAACCCTCAAATGGCCCACCACTCCCACTACTAGCTTCTAAGCAAACATCC
ATAATGAGAAGAAACCAATCAGCAAAAGGGGCTCACACAAGGAATACAAAAAAAGCAGGA
GAGAGGAAACATGACACCTACCAAAGGAACACAATAAATCTCCAGAAACGGCACAAAACC
CTAAGGAAATTGAGATGTCTGGCTGAGTAAGAGCCTAATTGACAGACAAAGAATTCAAAA
ATAACGATCATAAACAGAAGCTCAATGAGATACAGGACAAAAGCTATAGAAAACCAATTC
AAGGAAACCAGGAAAACAATCCATGAAATAAATGAGAATTTCAACAAAACAGAAATAGAA
ATTATAAAAAAGAACCACAAACAGAAAAATTCTAAGAGCTGAAAAATATACAATAACTGA
ACCCAAAAAAAACCATTAAAAATTCAATAGAAAGTATCAACAGCGCGCGCAGAATAGATC
AAGCAGAAGAAAGAATTCCCACTGAACTTGAAGACAGGTCATTTGAAATTATCCAGTCAG
ACGAACAAATCCACAAGAAAAAAGAATGAAAAAAAATTGAAGAAAGCCTCAGAGACATAT
GGGACAACATCAAACAAGAGAACCAACATATGCATAATGGGAATCCCAGAAGGAGAAGAG
AAAAAAGAGAAAAGGCTCAGAAAGCATATTTAAAGAAATAATGGCTGAAAACTTCCCAAA
TCTGGCGAAAGAGATGGTATATAAATCAAAAGCCAATAACATCCAGATACAAGAAGCACA
AAGAACCCCAAATAGATTCAACACAAACAAAATAATCAACATATACACACAAAAAAACCA
AATAAAAAAAAAGTCAACAAGAAAAAAAAAATCTTAAAGGCAGCAAGAGAGAAAAGAAAGA
TCACATACAAAGGAATCCCAATAAGACTAACAGCAGACTTCTCAGCAGAAACCTTACAAG
CCAGAAGAGAGTGGGATGATATATTCAAAGTGCTGAAAGAAAAAAAATCTGTAGCCAAGA
ATATTATACCCAGCAAAGCTATCCTTAGAAATGAAGAAATAAAATCATCCACATACCAAC
CAAGAATAATATATCCAGCAAAACTAACCTTCATAAATGAAGGAGAAATAAAAACCTTCC
CAGACAAACAAAAACTAAAGGAATTCATATCAACTAGACCTGCATTACAAGAAAACCCAA
AGGGAAGTCTATAAACAGAAAAGAAAGAAAAAGAAATCAAACCACATAAAAACAATAAAA
CACAAAACCACAAAGATATAAAAAAAAAATAGACAAAAAAATATACATAACAACCAGAAA
ACAACAAAATGACAGAAATAAAATCTCACATATCAATAATAACCCTGAATGTAAATAGAA
TAAACTCCCCATAAAAAAAAAAAATAATGGATAGATAAAAAAAAAAGAAACAACTAAACA
CATCCTAAAAAAAAAACACCCAACCCAAAAAGAAAAACATAGACGGAAACAAAAAAAATG
AAAAAAGAAACACCATAAAAAAACAAACCAAAAGGAGAAGAAGTAGCAAGAAAGACACAA
GACAAAAAAAAAAGCAAAAACAAAAACTAGAAAAACAAAAAAAAAAGACAAAAGAAAAAA
AAAAACGACCAATTCAGCAAGAGATATAATTGTAATATGTGTACAATTGTAAATATATAT
GCACCCAAACACTAGAGCACCCAGATATATAAAGCAAATATTATTAGATCTAAAGGGAGA
GATAGACCCCAATACAATAATAGTTGAGGACTTCAACCCCACTCTCAGCATTGGACAGAT
CATCTAGACAGAAAATCAACAAAGAAACATGATTTAAACTGCACCATAGACCAAATGGAC
CTAAATAACAGACATTTACAGAACATTTCACCCAACAGCTGCAGAATACACATTCTTTTC
ATCAGCACATGGAACATTCTCCAGGATTGACCATATGTTAGGACACAAAACAAGTCTCAA
CAAATTTTATCAAGTATCTTATCTGACCACAATAGAATAAAACTAGAAATCAATAACAAG
AGGAACATTCAAAACTATACAAATATATGGAAATTAAACAACATGCTCCTGAATGACAAT
GAGTGAAGAAGAAATTAAGAATGAAATTTAAAAATTCCTTGAAACAAATGAAAATAGAAA

CACAACATACCAAAACGGGAACAGCAAAAGCAGTTATTAAGAGGCAAGTTT

>4_cons
GAAGGCGGAACAAGATGGCCGAATAGAAGACTCCACCGATCATCCTCCCTGCAGGAACAC
CAAATTGAACAACTATCCACACAAAAAAATACCTTCATAAGAACCAAAAATCAGGTGAGC
GATCACAGTACCTGGTTTTAACTTCATATCACTGAAAGAGGCACTGAAGAGGGTAGGAAA
GACAGTCTTGAATTGCCGATGCCACCACTCCACCATACCCGGGCAGTGGCAGAGTGGTGT
GGAGAGAGAATCTGTGCGCTTGGGGGAGGGAGAGTGCAGAGATTGTGAGACTTTGCATTG
GAACTCAGTGCTGCCCTGTCACAGTAGAAAGCAAAACCAGGCAGAACTCAGCTGGTGCCC
ACGGAGGGAACATTTAGACCAGCCCTAGCCAGAGGGGAATCGCCTATCCCAGTGGTCGGA
ACCTGAGTTCCGGCAAGCCTTGCCACCGCGGGCTAAAGTGCTCTGGGGATCTAAATAAAC
TTGAAAGGCAGTCTAGGCCAAAAGGACTGCAAATCCTAGGCAAGTCCTAGTGCTGAACTG
GGCTCAGAGACAGTGGACTTGGGGGACACATGACCTAAGGAGACACCAGCTGGGGCAGCA
AAGGGAGTGCTTGCACCACCCCTCACTAAACTCCAGGCAGCACAGCTCACGGCTCCGAAA
GAGACTCCTTCCTTCTGCTTGAGGAGAGGAGAGGGAAGAGTAAAGAGGACTTTGTCTTGC
AACTTGGATACCAGCTCAGCCACAGTAGGATAGGGCACCAAACAGAGTCATGAGGCCCCC
ATTCCAGGCCCTGGCTCCCGGACAACATTTCTAGACACACCCTGGGCCAGAAGAGAACCC
GCTGCCTTGAAGGGAAGGACCCAGTCCTGGCAGGATACATCACCTGCTGACTAAAGAGCG
CTTGGGCCCTGAATGATCAACAGCGATACCCAGGCAATACTCAATGTGGGCCTTGGGTGA
GACTCAGAGACTTGCTGGCTTCAGGTGTGACTCAGCACATTCCCAGCTGTGGTGGCTATG
GGGAGAGACTCCTTATGCTTGAGAAAAGAAGAGGGAAAAGTAAAGGGGACTTTGTCTTGC
ACCTTAGGTACCAGCTCGGCCACAGTGGGATAGAGCACCAAGTAGGCTCTTGGGGTCCCC
GATTCCAGGACTTGGCTCTTGGATGGCATTTCTGGACCTGCCCTGGGCCAGAGGAGAGCC
CACTGTCCTGAAGAGAGAGTCCCAGGCCTGGCAGCATTCACCACAAGCTGACTGAAGAGC
CCTTGGGCCTTGAGAGAACATTGGCGGTAGCCAGGCAGTACTCTCCATGGGCCTGGGATG
GTGGTGGCCACAGGGAGCGACTCCTTTGCCTGTGGAAAGGGGAGGGAAGAGTGGGAAGGA
CTTTGTCTCGTGGTTTGGGTGCCAGCTCAGCCACGGTAGAATAGAGCACCAGGTAGATTT
CTAAGGTTTCTGACTCCAGGCCCTGGCTCCCGGATGACATCTCTGGACCTGCCTGAGGCC
AGGGGGAACTTACCACCCTGAAGGGAAGGACACAAGCCTGGCTAGCTTTTACAACTGCTG
ATTGTAGAGCCCTAGGGCCTTGAGCGAACATAGGCGGTAGCCAGGAAGTGGTTACAGCAG
GCCTTGGGTGAGACCCAGTGCTATGCTGGCTTCAGGTCTGACCCAGCACAGTCCCAGTGG
TGGTGGCCACAGGGGTGCTTGTGTCACCACTCCCAGCTTCAGGCAGCTCAGAACAGAGAG
AGAGACTCCATTTGTTTGGGGGAAAGTAAGGGAAGAGAACAAGAGTCTCTGCCTGGTAAT
CCAGAGAATTCTTCCGGATCTTATCCAAGACCACCAAGGCAGTACCTCTATGAGTCTGCA
AGAACCACAGTGTTAATGGGCTTGGGGTGCCCCCTAAAGCAGATATGGCTACATGACCAA
AAACTTAGATCAAAACACCCAAGTCCATTCAAATACCTGGAAAGCCTTCCCAAGAAGAAT
GGGTACAAACAAGCCCAGACTGTGAAGACTACAATAAATACCTAACTCTTCAATGCCCAG
ACACTGACGAACATCCACAAGCATCAAGACCTTCCAGGAAAACATGACCTCACCAAACGA
ACTAAATAAGGCACCAGTGACCAATCCTGGAGAAACAGAGAGATATGTGAACTTTCAGAC
AGAGAATTCAAAATAGCTGTTTTGAGGAAACTCAAAGAAATTCAAGATAACACAGAGAAG
GAATTCAGAATTCTATCAGATAAATCTAACAAAGAGAATGAAATAATTAAAAAGAATCAA
GCAGAAATTCTGGAGCTGAAAAATGCAATTGGCATACTGAAGAATGCATCAGAGTCTATT
AACAGCAGAATTGATCAAACAGAAAAAAGAATTAGTGAGCTTGAAGACAGGCTATTTGAA
AATACACAGTCAGAGGAGACAAAAGAAAAAAGAATAAAAAACAATGAAGCATGCCTACAA
GATCTAGAAAATAGCCTCAAAAGGGCAAATCTAAGAGTTATTGGCCTTAAAGAGGAGGTA
GAAAGAGAGATAGGGGTAGAAAGATTAATTCAAAGGATAATAACAGAGAACTTCCCAAAC
CTAAAGAAAGATATCAATATTCAAGTACAAGAAGGTTATAGAACACCAAGCAGATTTAAC
CCAAAGAAGACTACCTCAAGGCATTTAATAATCAAACTCCCAAAGGTCAAGGATAAAGAA
AGGATCCTAAAAGCAGCAAGAGAAAAGAAATAACATGCAATAAAGCTCCAATACGTATGG
CAGCAGACTTTTCAGTGGAAACCTTACAGGCCAGGAGAGAGTGGCATGACATATTTAAAG
TGCTGAAGGAAAAAACTTTTACCCTAGAATAGTATATCCAGTGAAAATATCCTTCAAACA
TGAAAGAGAAATAAAGACTTTCCCAGACAAACAAAAGCTGAGGGATTTCATCAACACCAG
ACCTGTCCTGCAAGAAATGCTAAAGGGAGTTCTTCAATCTGAAAGAAAAGGACGTTAATG
AGCAATAAGAAATCATCTGAAGGTACAAAACTCACTGGTAATAGTAAGTACACAGAAAAA
CACAGAATATCGTAACACTGTAATTGTGGTATGTAAACTACTCATATCTTAAATAGAAAG
ACTAAAAAATGAAACAATCAAAAATAATAACTACAACAATTTTCAAGACATAGACAGTAC
AATAAGATATAAATAGAAACAACAAAAAGTTAAAAAGAGAGGGGATGAAGTTAAAGTGTA

GAGTTTTTATTAGTTTTCGATTGTTTGCTTGTTTGTTTATGCAAACGGTGTTGTTATCAG
CTTAAAATAATGGGTTATAAGATAATATTTGCAAGCCTCATGGTAACCTCAAATCAAAAA
ACATACAACAGATACACAAAAAATAAAAAGCAAGAAATTAAATCATACCACCAGAGAAAA
TCACCTTCACTAAAAGAAAGACAGGAAGGAAGGAAAGAAGGAAGAGAAGACCACAAAACA
ACCAGAAAACAAATAACAAAATGGCAGGAGTAAATCCTTACTTATCAATAATAACATTGG
AATGTAAATGGACTAAACTCTAATCAAAAGACATAGAGTGGCTGAATGGATAAAAAAAAC
AAAACCCAATGATCTGATGCCTACAAGAAACACACTTCACCTATAAAGACACACATAGAC
TGAAAATAAAGGGATGGAAAAAGATATTCCATGCAAATAGAAACCAAAAAAGAGCAGGAG
TAGCTATACTTATATCAGACAAAATAGAATTAAAGACAAAAACTATAAGAAGAGACAAAG
AATGTCATTTAATGATAAAGGGGTCAATTCAGCAAGAGGATATAACAATTTTATATATAT
GCACCCAACACTGGAGCACCCAGATATATAAAGCAAATATTATTAGAGCTAAGAGAGAGA
TAGACCCCAATACAATAATAGCTGGAGACTTCAACACCTGTCTTTTAGCATTAGAAAAAT
CATCCAGACAGAAAATCAACAAAGAAACATTGAACTTAATCTGCACTATAGACCAAATGA
ACCTAATAGAAATTTACAGAACATTTAATCCAACAGCTGCAGAATACACATTCTTCTCCT
CAGCACATGAATCATTCTCAAGGATAGACCATATATAAGGTCACAAAACAAATCTTAAAA
CATTCAAAAAATTGAAATTATATCAAGAATCTTCTCTGACCACAATGGAATAAAACTAGA
AATCAATAACAAGAGAAATTTTGGAAACTATACAAACACACGGAAATTAAACAAAATGCT
ACTGAATGACCAGTGAGTCAATGAAGAAATTAAGAAGGAAATTAAAAAATTTCTTGAAAC
AAATGATAATGGAAACACAATATAACAAAACCTATGAGATACGGTAAAAGCAGTACTAAG
AGGGAAAGTTTATAGCTGTAAGTGCCTACATCAAAAAAGAAGAAAAACTTCGAATAAACA
ACCTAATGATGCATCTTAAAGAACTAGAAAAGCAAGAGCAAACCAAACCCAAAATTAGTA
GAAGAAAATAAATAATAAAGATCAGAGCAGAAATAAATGAAATTGAAATAAAGAAAACAA
TACAAAAGATAAATGAAACAAAAAGTTGGTTTTTTGAAAAGATAAACAAAATTGACAAAC
CTTTAGCCAGAATAAGAAAAAAAGAGAAGACCCAAATAAATAAAATCAGAGATGAAAAAG
GAGACATTACAACTGATACCACAGAAATTCAAATGATCATTAGAGGCTACTATGAGCAAC
TATATACCAATAAATCGGAAAACCTAGAAGAAATGGATAAATTCCTAGACACATACAACC
TACAAAGATTGAACCATGAAGAAATCCAAAACCTGAACAGACCAATAACAAGTAACGAGA
TCGAAGCCGTAATAAAAAGTCTCCCAGCAAAGAAAAGCCCGGGACCTGATGGCTTCACTG
CTGAATTTTACCAAACATTTAAAGAATTAATACCAATCCTACTCAAACTATTCTGAAAAA
AGAGGAGGAAGGAATACTTACAAACTCATTCTATGAGGCCAGTATAACCCTGATACCAAA
ACCAGACAAAGACACATCAAAAAAAAAAAACTACAGGCCAATATTCCTGATGAATATTGA
TGCAAAAATCCTAAACAAAATACTAGCAAACCAAATTCAACAACACATTAAAAAAATCAT
TCATCATGACCAAGTGGGATTTATCCCAGGGATGCAAGGATGGTTCAACATATGCAAATC
AATCACAATCGATATGATACATCATATCAACAGAATGAAGGACAAAAACCATATGATAAT
TTCAATTGATGCTGAAAAAGCATTTGATAAAATTCAACATCCCTTCATGATAAAAACCCT
CAAAAACTGGGTATAGAAGAACATAACTCACGACACAATAAAAGCCATATACGACAGACA
CACAGCTAGTATCAAACTGAATGGGGAAAAACTGAAAGCCTTTCCATTAAGATCTGAAAC
ATGACAAGGATGCCCACTTTCACCACTGTTATTCAACATAGTACTGGAAGTCCTAGTTAG
AGCAATCAGACAAGAGAAAGAAATAAAGGGCATCCAAATTGGAAAGGAAGAAGTCAAATT
ATCCTTGTTTGCAGATGATATGATCTTATATTTGGAAAAACCTAAAGACTCCACCAAAAA
ACTATTAGAACTGATAAACAAATTCAGTAAAGTTGCAGGATACAAAATTAACATACAAAA
ATCAGTAGCATTTCTATATGCCAACAGTGAACAATCTGAAAAAGAAATCAAGAAAGTAAT
CCCATTTACAATAGCTACAAATAAAATTAAATACCTAGGAATTAATATAAAAAAAAAAAT
GAAAGATATCTAAAATGAAAACTATAAAACACTGATGAAAGAAATTGAAGAGGACACAAA
AAAATGGAAAGATATTCCATGTTTATGGATTGGAAGAATCAATATTGTTAAAATGACCAT
ACTACCCAAAGCAATCTACAGATTCAATGCAATCCCTATCAAAATACAAATGACATTTTT
CACAGAAATAGAAAAAACAATCCTAAAATTTATATGGAACCACAAAAGACCCAGAATAGC
CAAAGCTATCCTAAGCAAAAAGAACAAAACTGGAGGAATAACATTACCTGACTTCAAATT
ATACTACAGAGATATAGTAACCAAAACGGCATGGTACTGGCATAAAAACAGACACATAGA
CCAATGGAACAGAATAGAGAACCCAGAAACAAATCCATACATATACAGCGAACTCATTTT
CGACAAAGGTGCCAAGAACATACACTGAGGAAAAGACAATCTCTTCAATAAATGGTGCTG
GGAAAACTGGATATCCATATGCAGAAGAATGAAACTAGACCCCATCTCTCGCCATATACA
AAAATCAAATCAAAATAGATTAAAGACTTAAATATAAGACCTCAAACTATGAAACTACTA
AAAGAAAACATTGGGGAAACTCTCCAGGACATTGGATTTGGGCAAAGATTTCTTGAGTAA
TACCCCACAAGCACAGGCAACCAAAGCAAAAATGGACAAATGGGATCAAATCAAGTTAAA
AAGCTTCTGCACAGCAAAGGAAACAATCAACAAAGTGAAGAGACAATACACAGAATGGGA
GAAAATATTTGCAAACTACCCATCTGACAAGGGACTAGTATCCAGAATATATAAAGAACT

CCTACAACTCAACAATAAAAAAAACAAACAACCCAATTAAAAAATGGGCAAAAGACTTGA
ATAGACATTTCACAAAAGAAGATATACAAATGGCCAATAAGCATATGAAAAGATGCTCAA
CATCATTAGTCATCAGGGAAATGCAAATTAAAACCACAATGAGATACCACTTCACATCGC
CCATTAGAATGGCTAAAATCAAAAAGACAGACAATAACAAGTGTTGGCGAGGATGTGGAG
AAACGGGAACTCTCATACACTGCTGGTGGGAATGTAAAATGGTACAGCCACTTTGGAAAA
CAGTTTGGCAGTTCCTCAAAAAAGCCTTAAACATACACTTACCATATGACCCAGCAATTC
CACTCCTAGGTATATACCCAAAAGAAATGAAAACAAGATGTTCACAAAGATACCTGTACA
CGAATGATTCATAGCAGCATTAGATTCATAATAGCCAAATAATTGGAAACAACCCAAATG
TCCATCAACAGGTGAATGGATAAACAAAATGTGGTATATACATACAATGGAATACTATTC
AGCCATAAAAAGGAATGAACTACTGATACATGCAACAACATGGATGAATCTCGAAAACAT
TATGTTTAAGTGAAATAAGCCAGTCACAAGAAGGATACATACTGTATGATTCCATTTATT
AAATATGAAGTATCTAAATATAATCAAAAAGACAAGCAGGCAAAACTAAAAAATATATTG
TTTAGGGATACATACATATGTGGTAAAACTATAAAGAAAAGCAAGGGACTCATAGAGACA
GAAAGTAGCACAGGAAAGATAAACAAGAAACTAATAATAGTGGTTACCAGGGGCTGGGAA
GGGTAGTGGGGGGGAGGGGGAGAGGTGGGAGTGGATGGTTAATGGGAGTACTTTTCAAAA
GGTTCATTTTGGGATGATGAAAATGTTCTATATAAACTATTCTGTATTTGATTGTGGTGG
TGGTTGCACTGTAACAGTGTGAAATATACTTAATATAATATTGAAATTTGCCAAAACCCA
CAGAACTTTACAGCATAAAGAGTGAACTTTAATGTATGAAAATTTTAAAAAAACAAACAA
GAAAAAGGGAGATCACAAAATGAAATACAAACTGAAACAAAAGAACCTAACTGTATTACA
AATGAATAACATAACCTCACTGAAGGGAATAAGGAAAAAAAGTACTAAACTAAATAACTT
TAGAAATGAGTATTTTGACTACACACTCTAAGACTAAAGACAAAAAGAACTGTTAAAAAT
AACTGAACCCTAATGAATAAGCTTATTTTCCACAGGGGCACAGGTTAACAATTCTGACAC
TACTATATATATATATTAAAATTAAATAAAAAGGTAAATATATTGAAGATAATGGGAGCC
AGGTTTCTCACTGTCGGAGTAGGGAGTTACAAATATGGAAAGGGAGAAGACTAGAATGAA
CCCTGTGGTATTGGATTGGAATTGGAGGTATCAGTATGAACTCATGCCTTTTAATATAAA
TAGATATACAGACAGACAGATATAGAAATAGATATAGATATATATGTGTATGTGTATATG
TGTATGTATATACGTACATATATTTCCTAGCTCTGTCCACTGAGAGGGCCTAGAAGCAAT
GACACCCCAGTAGCAATGAGCACACCTAGCACCATGATCTTGGTATGTAAATACCATTCT
CCACTAAAAGGAACCAGGGCTCCTTGGAGAAATGGCTGATTCCAGGGCGCACAAAGAAAA
GATACAAGATGAGCCTGGAACATCTTGCAGTGCCAGAAAATAAGGAAGTGCTCAAAAAAA
CAAGGAGACAGGTATGCAAAAGGAAAAAAGAACCAAACTGAAAGAGCTCCCAATGGCCAA
AGCTGGGACAATTTAAACAAAAAAAGAAAATGCTTGAGGTGATGGATACCCCATTTACCC
TGATAACACAGGTGATTATTATACATTCTATGCCTGTAACAAAACATCACTATGTAAATA
AAGAAAAAGAGAAAACTCTTCCTTACAGTAGAATGCCAACTAATAAATGTAGAAGGAATG
ATGGAATTAGAAAATCACCATTTGAAAAAAATCACAGTAACAAAAGAATCACAAAAAAAT
CATCAAGAGATCCTAAAACCAAGGAAGGAAAGCGTAACGACCAACACCATAAGTACAGAA
TCGAAAAAGAGACCCCCACAAAAAAAGCATAAAATACAAAGGAAAAAAAAGTAAATTTAC
AGTAGAGAAACCTGACAAACACCACCTTAACCAAGTAATCAAAGTTAACATCAACAATAA
TAAGACAAATCGACAGCATATACCCCCTGAGATAAGACACGAAAAAAAACACAACACAAC
TCCGGGGGCATCCCCCCAAAAAACCAAAAACCCAAACCAAACAATAAAAAAACATCAGAC
AAACCCAAACTGAGGGACATTCTACAAAATAACTGACCAGTACTCCTCAAAACTGTCAAG
GTCCTCAAAAACAAAGAAAGACTGAGAAACTGTCACAGACAAAAGGAGACTAAAGAGACA
TGACAACTAAATGCAACGCGGGATCCTGGATGGGATCCTGGAACAGAATTTTTTTTGCTA
TAAAAGAAAATAAGGGAAAAACTAACGAAATCTGAATAAAGTATGGACATTAGATAATAA
TAATGTATCAATATTAATTTACTAATTGTGACAAATGGACAGATGTTAAGAAAGAGAATG
TCCTTGTTTTTAGGAAATACACACTGAAGTATTTAGGGGTAAAGGGGCATCATGTCTGCA
ACTTACTCTCAAATGGTTCAGAAAAAAAAATATGTATATGAAAACAGAGAATGATAAAGC
AAATGTGGCAAAATGTAAACATTAAATATATACAGATGAAGGGTATACGGGAACTCTCTG
TACTATTTCTTCAAATTTTCTGTAAATATAAAACTAAGCTTAAAAAAAAAAAAGTAAAAT
TTAAAAATAAAAA

LTR consensus sequences:

>1_cons
GTTTTTGGCAACCAAGGAAGGGGGTCGAGGTAGAGAAAATGCTAGGCATTCAAAAATCTC
CTTTTCCTTTTTGCTACAAACAGGAGATAACCTCACGCTCTAAGCTCAAAATACTTTTCA

ACCTCGGACCAATGGGGCAAAGCGCCGAAACATGGAAGCAACTCCAGGTTTCTGGCCGTG
GCCAGTGAAACTAAGGAGTTTCCATGTAGAGAAAACCAACCACCACCCCCCATTTCACCC
AGGGGTCAGGAGTCTTTTCCTGTACTATTCCTCTCTTTTTCGAGTTCAACCTGTTTCCAA
CACAAGAGAGGCGGCTTCCTCCTACCTCCATGAAAATGGAAGGCAAGTAGCTGGGGTCAC
TCCCCAGTACTGCCTGAAGGCCTAGGAATGAATGGGAATAATTGCCCTGCCCCGAAGGGG
GAATGAAACTTTTATTTTTTTATCTTTTCCGAGTGTGGTCCCTGATCCCTACATGCGGCA
CAGCTCAGAGCAAACTCACACGTGTTTCAGGAGACTTAAACCTTCTTTTCTTATGCTAAA
TTCTTCCCTTATCGTACTCAACTGGCTAAGGAACAAAAAGGCCCACCCAGCATCCAGTTC
CTATCATTACAGTTCATGGCTATAACTAATGGAATGGAAAGCACGGGAAAGCGTGGCGTT
ATCAAATTATAAGAATGCTAAAAGATGAGGCCTCCATCCTGGAGCTAGAACACACCTCCA
AAAGGGCACCAGAGGCATAGCACCCAAAGTGGCGATGGTAGCCACCTATCCTCAGAGTCA
TCCCCAACCCTAGGAAAGGAACCCAATTGGGGTTTGTCTTGGAGATAATCCCTATCGGAT
AAAAATTAAGCTTAACTCTCTAGAAAAGAAACCATTAGCACAAGAACTAAAGATTAATAC
CCAGCCACTCTTAAAATCTTCTTTCGAAGTTGCAATACTGTGTGGACCCCATATTGTTTG
GAATCTGAAGTTTACTGTTGAATGAGAAAGCGAAACAACCATATAAGTTGCGAGACCATT
GGGCTGCCTTAACAACCAGGGGGCCTGGTGAACGTGTTAGATCTTCCTTTTGTACACCTT
GCCACCAGTTTACTTTTGAGTCTGACCAACTTTATCCTATAAAAAAAGTATTGCCAGGGA
AGAAAAACGGCCAGCAAAAGGAAAAAATTGAACTTGTTTACTTTGACTACTATACCCGAA
AGAGTGTTACAAATCATATGTGAAATTGTCTGTGGAGTCGCGAAACACCCACAAAATGTA
ACAGTTTGGAGGGCTCAGAAAGAGAAAGAAAATGTTATGTTAGATTTTAGCTATAGAGAC
AAAGCTGTAAGAGCTTTGGCTATGAATGTATATACGGCTCGAAGCAAATTTATATGTAAG
CACCCATATTGTTTTACGTTCTCACTGCCACAACCTGCATCCAAGTAAAAAAGCAAACAC
AAATCAAGTAAATAAGTCAGAGCAATTTTCAAGTTCACATGACTATAAGTATAACTTTAC
TAAACAAGCAGCTTTATAAATTATTGAGGAAATAAAAATAGAAATGCCTTCAGAATTGCT
AGCATACGTTTTGATCTAAGTTTTAGTTTAGTCTCTGCTACTTATGAAGAAGGATCATGG
TTTGGCATAGAAAGTTATAAAACTATAAACCCAGCCAAAACAAAATGATCTTGGTTTGCC
TGCCCTTTTTTTTTTTTGACAAATAGTAGCAATTTAACATGAACAATAAAGTCTGAGCTG
TTGGCCAAAATATCTATGTACTTAACTTTGAGGCTCTTACTTAGGTTTGGGTGAGCACCT
GATGTTCACTGGCTATTAAAAATGTGGTTAAAAAGGAAAAAAAGAACTTAAAAGAATAGT
GTAAAAGAGCTCAATGGACAAAAGTAAGCTAAATAAAAAGGCTAAAATCCAAAATGTAAG
TACATATGAATGCGATAAATGTTTTAGGTAAACTTTTTGTGTAAATTAAAATCTTAAAAT
TATATTTGATGCTCATTAAATATCTGGGTAATCTTCCATAAGGGAAGGGTTGCAATCTGG
AGAAATACATGTTTGAATTTTGTCGAATGTTGATTCTCTTTAAGTCCCCGTAACGAATCG
ACCAGAATTTCTCTCTAGACAGAGAAATTAATTCTAAGTTCTCAATAAAGTTTTAAGCCA
CCAAGGACAAAATTCCAGTTAACACATAATGTACCAAACAAAACGAACCAGAAAAGGCTT
CGATATTAATGAGAAAAAGAATAATTTTATCTAATTCAGAAGTTATATAAAAGTTAGTTC
AAATTACAGAATTAAAAAGGTTATTTATGAAAAAATGTACTAAGAAAAAAATAAGTAGGG
GAGAAAAATGTGGAAAAAGTTTAAATAATAAAATATTCTTTAAAACCTGATAAAGAATTG
GAGACATTTGACTAATTAACATTTTCATAGTTAAAGCTGTTAGTCTTGAATAATAAAATA
AGAAGTATTATATAAGAACCCATAAACAACTCAGAAAATACTTCATACAAACACAGTTAA
ACAATAAACTTAATTTAGTCCTGAGCCAAACTCAAAAGAAATAAATACACAGATTCCCCC
CAGGTTTACTGTTTTGCGTGGATAGTGATAGAACTTGGCTGTTTGAGAGAGATGATTTAG
GAAAACACCTTAATTGACTTTTTTGATTGAACAGGATGTATAATGATATTGGTAGACTTG
AGGAAACAGAAAAGCGTACAAGGGGTGAAATGTGAGCCATGTGAGTTTTTTTGGGGGACC
CTAGGTAACACTATAGTCTCCAAGGTAAATTGAGTAGGAAAATTTAGGGTTGGTATCCTG
TTTAATTGTATTTGCTTCAAATTTTCATTCGTTTGCTGTTTATTTCTCCTCTGGCTTTAC
TTGTGTGTGATCGTATATACATAAAACAATTGATTTTTTTGATAAGTTTTTTAGTTACTA
GTGGAAGGCTTTTATTTGGTTCTGTGGATAGTTATTTTGTTTCCTATGCTATATCATTAC
TAGCAAGTCATCATTTGTTCCATTTATCTGGAATGCAAAAGAGACCACTGTTAGGCCCGC
AAGAAATAATGTAGTACACAAACTTTTCTATCCTATTATAAACTAACTTTTTGGAATTTA
GGCTTCCTGATACTTTAAGGAACTTAAAAATATTTTCGTAAAAAGAATTAAGCTCAATAT
TTCTCTTTCTCTGCCTAATTTCTCAAAAATGCATAAACTAGTTATGACTATTCTTAATTC
ACCCTTATGTGATTCTTTGCATACACAGTCAAGCAGGGATGCTAGGGCAGCTCAGGGAGA
GAGAACCCAAGGGCATTCAGGTTTGGAAGAGAAAGAGTGAAAATTTCTTTCTAAATAGAC
TCAGTCCTCTTTCTCAATTTTATCATACCTCTCTAAATAAATTAATGAGAGAGTAGAAAT
CACTGTTTAACTCATCTCTAAAACCACAAAACAATGAACCCCAGAAATTTAAGGCGGATA
TTTAGCTATAAAAATTTTCATCTGCTACATGTATCCTTCTATATTAAACTGTAGAAAAAA

GGGGTACATTAGGATAAAATGCGTGCCTAGGACTCCATAGGCTTGCTGTTCAAGATGGCC
CAGCAAACTGGACAGTCATGTCCTTGGGAGCTTGACCTCGTAACCATGTGGCCATGCTTT
CATACCCCTAGGTCACTAAAACAGCTCAACAAAGAGTTCAATAACTAAATTAGGAAATTG
GAAAACGAAATATCATAAAGCTACTGGGTCTTCTTCTATCTGTCTGTGTAATACGTCTAT
GTATGTATGTGTTGTGTGTGTAATGTATCACTACTAAAAATACAAAAAAGAACCTAAATT
GATCTGCATAATAAAAAAAAAAAAAATTTGAATCAAATATTTTATGAGGAGAGAAGAAAG
AACGAGTCAAATGCTTTTTCAAGTTAATGAAGTACTTTATTAAGTTCATGTGACTTAAGT
AATATTTAAGAGAAAGAGACAGCCTAAGGTTAATCCCCAAAGTATTAGAAAAAAGATATC
AAAAATGTCTTAAAAAATGTAAAAATACATTTTGGTCTAAATTATACAGATCAAATACTT
TCATACTTATCCCTGCCAAATACTATAAAGGTGTCAAAGGTTGGCTAAAATGTTTTAAGG
TTATAAACCCAGCCCAAAACAGAATGATCTTTGCTTGTTTGATTTTTAAAAATTATTAAA
TTGATATTGGAATAATGAAAAAAGCTACATCTTGAATTTAGTAAGATTACCATAACTTCT
AACCTTGTGGCTTTAGGCGATATTTAAATGATGACTATCGCAGTTTTCATAAAGAATCTA
GGTAAGCAATTAATAAAATAATTAGGTAAATGTAATGGGATAAATACCTGGAGACAAACT
TGTCATAATTTAGAATATAAAGTTATATTAAATTAAATAATAGATAATTAATTATTTGAG
TATTTTCCAATAAAAATATATTGTAGGAAAACATTCTTACTTAAAAAAAAGTGTGTCCTT
TTTAAAAAAATGGTGAATAAGTTTTGTCTAATTCAAAGCTTATTTAAAGGTTATATATAA
AACAAGGTAAAAGGAACCAGGAAATAAAAAAAATGTAAAGAAAGTTATGAAAATAAACAC
GTCACCTAGCTATGCAAAAAAGCTGAAAAAGAAAAAAAATCATATGAGAAAGAATCTTAT
ATGGTAAATTCATAAGAAAGAATCTTATATGGTAAATTCTTGTCCTAAAATAAAATAACA
GGTTGTTAAAAAAGAGAGATGTTTAGGACAAATCAGAAAGTCCAAGCATATTATAAATGG
TCTGTGTAAATCATAAAAAAATTTACAAAAAAAGAAATTAAAAAAATTTATATGATTAAG
TTGGCTATAATTAAAAGAAAATTATTTATAATAGTCTTTCTAGAGATTGAAGTTTGATAT
TAAAAATACACTAATACACTAAAAATATGCAAGAAAAACAAAACAGTCTAAAAGTAATGA
GGCATCCAAAAAACTACAAAAGATGTTAATGTGAATCCAAAAATTCAACATCTATTCAAC
CTCACTGCTTTAAAGCTAGGTATACCCCGTAAGAACACATGAAATAACAACTCCCCCCCC
AAACTCTACTGTCAGCTCCTGTAATTTTTTCCTCAGGTTCTAACTGATGTTGTGTACTGA
TGCTGGAAAGGGTCAAACCTAAAGGTCTAAAAGAAATGTTTTCTTCCAATATAACATTCT
GTACTCAACTTTTCTTGATGTGTCTGAACTGCTCCATGAAACCAAAAAACCACACCTAGA
ACACTGGAAACACTCTTCCTTGTCTAATTAAAAAAACCACAACTTACCAAGGTTTACATT
AAAGTTAAAAGTCGACAGCAGTTCCCATTATAACAGACAAAGGAACCTAGTGAAGAAAGA
AAGCCTTGGAGAAAAGGCCCAGGGTCCCTGTGACGGAAGTGGCCAAAGAAAAAGATTTTA
TGTTTTATCAAAAAAAGTTTGGTGTTTTTAGGAACAGGTTATTAAGAAGCAAAGGAAAAC
TGAATTTTTAATTGTGCAAAAAAGGGTAAAAGCATCCATGTATCTTTCTGTATTGCTTTT
AAAGTCCTTATTGTTTTAAGTTAGAGAACATAAAGCTAAAGGTTTAAACAGGTCGTGGAA
GAATTGTAAACAATTAATCTTGAAAAAAATTAAAGCCACATCTTCAGGCCCGTAGAAGAT
GCCAATCAAAATAAACTGCATTCCTGAGACACAGGAAATTAAAGCTATTCAACTCCTCAA
GGCCCAGGGACTATCCAGAAGAGGTGGGTATGTGTGAACATGATGTCTAATATCCAAAGA
TAAAGTTATTTTATGGTTTCTCTGTAAATTGAACATTGAAAGTTTCTCTATAAATTAATC
ATTAAAATTAAAAGCACACTGATGCAAGACCAGCATATGGGCCCCTGTGTCAGATTAACA
AGGTTTTCTTAAAGCACTAATCTGCTCTTTAATAAAAATTTATAAAGGGTTATAAAATGT
TTACGAAAATCTCATCCTATGGTCAAACTGATTAAGATCGGAAAGATTAAAATATAAGAG
ATTATTTAAAAAATATTTCTGAGATTGACATTAATAGTACACTAATGCAAGGGTGAAATG
TGGCTTTCTCTCCTGAACAAGATTTTCAAACAAAATTAAAAGACACCAAAAGATTTTTAT
TAGCCTTTTGAATAAACTACCAACAAAAAAAGAAGGGAAAGACAAGAGACAGATTGTTTG
GAAAACTAAGTCTTCCCTCTCTCAAAGAATGAAGGTTTTTGCCCTTTAAAAAAGTTTTCC
TGGAGCAATCATTTTGGCTAAATGAATGACTTATTTTAATGTAACCTGCAATTCTATTTC
ATAATATCAAGTGTTTTAAACCTATAACATATCTCCTCAGTCTCCCCAAACCTTCAGTAC
AGTCTATGTCTTTCTGACCAAAATTGTCTTTTCAGATATCAGGCTTCTTAGAAGCATCAG
AAGGCCCCACGAAGAACCATCCAAAAGAGAGGTAAAAAGGATTATTTGACACATTTAGTT
ACATTTCTTCCCTGCCAGAAAGCATTGACAAAAACGAAAAATGTTTAATCTTCTTTAGGT
TATATTTTAATAAATAAGTTATTGATATATGTTCCAAAATTGTATGGGATTTCTAAAATT
CTAAGATGTCTGAGTATATATTATCAATCATAATTAAGGTTATTATGTTAAATTATTGTA
AACCACAGAAATAACAAAACCTTTTGATCTGTGTGAGTTGTGTTTTTAACTGTAACTATT
CTAAGAATTTTCCACAGTTATTCACAGACAATTGTTGTATTGTTTGAAACCGTTTCAAAG
ATAGTTTATAATAAGCTATGGTGTCTTTTAGGAAGTTGATTAAAGGATGGAAAGAACTCA
AAAAAGGGGGCTGAGATCCCACACAAGGTCTCGGACAACGCGGTGGGAGATTGTAACATC

AGAAAAGAGAAAACACCTACAGGACCCCGAGAAGATCCAAGTGACTCAGAAAAATGCCTA
AACCAAACTCCAGCAAAAGAAGCAGGAATTAAGAGCCAGCCCGTGAAGGTGACCAGGAGA
GAACATGAAAAAATCTTTTTGACTTTTTGCTTAAAACATTGCTGATCCTTTGTTTTGTTT
TTCAGAGTCAAGAAAACTTTTATTTTGAACTATTTACAGCCTTTAACAATTGAGTAAAGT
ATACTCCTATGAACAAAATTTGGAGCATGTTTGTTTCTCTCTGCCTGGTTCCCTGAGAAT
TCGCCCTGATTAATACTTTGTTACGTCATTAAAAAGAGCCAGTATTGGGGAGGGCACTGG
GAACCCACTTCCTGCAATATAGTGACTTGCATAAAAGACAATAAGAATCTAATTTTCATT
TGCCACAGGACGTGATATGAGGGATTGAAGATTTGACTGGCCAATTTTATTTAGACCTTA
AAGGGAAGGGTTTGCTTTCCTGTAAGGAATCAATCTTGACATGTAGAGCCAATAAAAGCC
CTATGGAAAAACTGGCCTCATAACCTTATATACACAGTCCCTGTACAAGCTTTCTGACCA
GTAATCAGCAAAGAATGTCACTTTCTGACAGGCCCAGGAGCCCCAAGTTTATCTTGGGAC
CTCAAGAGGAGAGGGAATTCACCCAACTCACAGGTATTTGAGGACACAAACCCATGGCTG
GGCTCAGCTTTAAAAAAGTCTTATCTGAGATTCCTTCTATGGAAAAAAGTTCCATCAAAG
CCAATCTAAAAAGACCATATATAAGAAATAATTATTCTTGCTGCACTTTATGCAAATAAT
CAGGCCAAGTATAATAAGAATAAAACCTATTTTACAAACAAATCAGTCCTAACATGATTT
TTTTTTACAAAAATCAGGAAACTGGAGAGAGAAAAATTATGTTTCAAAAACTATAATACA
ACTGTCATTAGATTCTAAACCCAGAAGTTGTTTTTAAGTTTTTGCCTACATGTTAGACTA
ACCCTGCTTGTTCCTGTGAACCAACCAGCAATCTCCGGCTGCAACAAAAAAAAAAAAAAA
GGGATCGGTACTGCGGACATTTCGGTTACAATTTTAATTCTCAACAATTTTCGCGAACAC
CCTACAAAGGGATAATCTTATACTCTGATAGATAAAAGATGAAGACCCAGAATAAATAAG
AAACCCAAAACCTACAGACGTCCCCTCAGAAAAAGTAAGAAAAAGAAAACTAACCAAAGC
CAAGTAACCTGAACACAAGTCTTAAAAAGAATAATTACAGCAACCAGTTATCTGGGTATG
TCACAAGACATCCTCTTCACTCCCCTATAAGAGAAGGACATAATTTATCAGTTTTACCTT
CCATTTTAGAGGATGAGAAAGAAAACATGCAACCACCAAAAAACAAACCCTCACAACAAA
CTCAATACCAAACTCTACGGCAAAGCCCAAAGAAACAAAACAGAATCTAAATTATCCTGA
GAAAGATACACTAAAAAATAGAACTGGCCAAAGAGAAAAGCCACAATGCAGATGAGCACA
GCCAAGGCCTACCGAGGAAAACAAGAGGCCAGTATCATAGAGAAACAGCAGGAGAGGATC
CACAACAAAAATAAGGTCTAGGGAATTCAAGGCTACTGACAGCAGGGGAGATAGGGCATA
AGTGAGTAGACCGCATAACTACCACACCCTAGGACCCACTGCATCATAGGTGCAAGCCGC
TTTGACACCCATGGTGGCACCTGCCAAGGTCACGGGGACCCTGGAAAACAAAAACACAAG
AAGAAAATAGTACCATCCTCCTTATATACATAACAATAGTCTACCCGGGGCATCGGATAC
GAACAAAAGGGAAACAGGCATGCATGCAGCCATCTTTATAATGACAAATAGCCTCTCTTC
TAATGTATGGACAACAGACGAATGAAACCTGTAACCACAAGAACACTGTAAAAAAAGAAC
GGCATGCTGACACCAGAAACCCAAAATCACGCTAAACGAAAAGGGAACTAAGGCTCCAAA
CTATGGGAAAACCCCCACAGAAGCAACAACCAAACCAAAAAGAGGGAATTTTCCAAAACA
AAAACGGGGTAGGCCATTGCTTAGGACTGAGCTCAGGCAAAAGACCTCATCAGATCAAAC
CAAAACCAAAACAGACCGGCAAAAGCTAAAACTTTAAGGCAATACATATGGATCTTATTA
GTTTATTTGTTTTTGTTTTTCTCCCAGAAAGCCCTATAACAAACATCCCCCAGAGCAAAA
GGATAAAACGCCTGAAGATCCTTGAGCTGTCCTTACCCCTCCCATTGTTTCATCATAAGA
CATGAAATCTAATAACCTGATTAATATCATCTTTCCTAAAGACCATCAAACTCCAAATGA
TCATGCAACAAAAGCCCCAGACAATGCCACCTGAAGACACCAACCCTGGTCATTAAGGAG
CTACCCTGTCTCCATTAGAGAGAGCACGCTAAGAGAGATGTGACCCACAATTCCAAGAAA
CAACACCCCCAGCCAGCAGGAAGCAGTTAAAAAAGGAGGACAGCCAGACCCTTTTCCTTA
TAAAAAGGTCAATAGAAATGACATCTGAAAAGGGGGGAA

>2_cons
AAATTGCAGCGGCGAGCAGGGTCCGGCCGGACAGAACCATGCCATAATGCCGAAGCCAGA
AGGGCGATACCCCCCAGTACCCGGCCCACAGCTTGAGCCTAAGCTGCGCGCTCCATGGGA
TTCAGACGTGTCTGGGCATGTCTGGAGAACCGGCCCGTCAGAAGTGGGGCGGGCCCGCCC
CACCTCCCCCGCAGATGTGAGGGTCTATTTGCATGAGCGAATGAGGAGAGAAAGGAATG
TGGAAACAGGAGAGAGCAGCGCACGATCCCCTGGTTGTTGGCCTCCTAAGACAGAGCTGC
AGACCCGGGCATCCTTGGGGAAGAGGGAGCCGGGAACAGCGAGCTCAGATAGACCCACAC
GCATTGGTGCAGGAACAAGGGCCCAGGCGTCGGAGGGTCCCTGAGGCATCTGAGAGGGAA
CCAACACGTACAATGCGGGCCTTCACAGAAAGGGCGATTCACGAGCTAAAGGACAGATAT
GGAGAGAGGACAGCACGGTGTCTGTCACTCCACAGGATTGGTCTCCCGGTTTGATAAAGG
GACATTGCAGATCCAAATGTCTCAGGCTGAATGGCGCAGTTAGGAATGGGGCCAAGGTCT
CCCAGCCCCCACTGAAGCCCACCGCCCCTATACTCCTGTCAAGATAAAAGGAGAAAAGAA

AGGAATTCAGACTTTTCTGGGGTTCTTGGATACTGGAGCCCACATGACAACATGTCTGAG
TCCCCTTAGGGGAAAAATTAAACTGATGACATCGGGAGGTTTGGGGACAAACATGGTGAC
CCATGGTGCTTATTTGCTTATGGTGCTTATCTGCTTGTGGGTGGGGCCCTTTGGGCCATT
TCGGGTGCCAGTGACCATGGTTCCCACCGCTGAGTGCATTATAGGCATTGACATTTTGGC
TGCTTGTGGCACAGAACATCACCGCTGCCTGAGGGGGTATGCCCCCTCACAGCTAAGAAT
TCGAGCCATAACAGCGCCGCAGACCCACAACTGCCTGCCCCCTCAGCCTGCCAACTCCCT
ATGGGTTATTCAACAAAAGCAGGACTGCAATCAAAGCACAGGCAAAAACTAAAAGAGCTG
ATTTCAGGACTTGCTACAGATAAAAATGTTACAAACCACCCTGTCACAATGTAACAGCCC
AGCCCGGCCGGGCGAAAAACCCCGTGGGACATTGAGGACAACAATGGACTGTCGCAAGCA
GCCTGCTGCTTTGGCCCCCTCCACACAAGCGGCGCCGACATCACCACAGTAATTGAACAC
ATCATGGAGGCTTCCAACCAATAGTGTGACACAGTTATTGATCTGGCTAATGGATTCTTC
TCAAGCCCAGGGGGGGAGAGGGGCAGAGATCAATTTGTATTCACATGGCAAAGTATACAA
TATACATTTACAGTGCTGCCACAGGAGTATTTGAACTCACCTGCCATATGCCACCAGTGG
GTAGGATGGGATTTCGCCACTGTGCTTTTGCCTAAAGTGGTCATGTGCATTCATTACATA
GGTGACATCCTCAGCGCGGCCCTGCAGGAGCCCATCACAAAAAACACCCTGGACGCCATG
AACACAAGCACGAGACAAACAGACTGGGAAGTTAACCCTAACAGTCCTGGGATCAGCCAA
ACTGGTGACCTTTTTCCCCCCCACTTCGGCGCGGAGCCAAAGAGGCAGCGCAGCTTCGGT
CAAGCAAAAATTGTTCGCCCCGGGGGCACCCACTAATAAAAAGGAGACCGGACAGCGGGT
CGGCGCCCCGGGGGACTGCAGACAGCATATACCTCACCTGGGTGTTCTTTTGGCCCCCTT
AGTCAAGGTGACCAACAAAGCCGCCAACTTTGAATGGGGCCCCTGGGAGCAGCAGGCCCC
GGAACCCAGCCAACAGACAGCGGCCCAGCGACCAATCGGCACGGACTTCAGCGCCTTCGC
GGACGGCAAAACCCATGGAATCACAGGTGTCCGCAACCTCCATGCATGCTGGCCGGCAGC
AGTGGCAACGGAAAACTGCCACTGGGGTGCACCAGCCTCTCAGATTTTGGACACATAAGT
TGCCTGAGGCAGCCACCAGATATACCTCTTTTGAATGGCAACTCCTTGCTTGCTATTGGG
CACTGGTGGAGACTGAGCATCTTACGGCCGGAGCGCCACGTGTGACGCTGCAACCCGAAA
CGCCCAGTCTCACAGCGGCGCTGCCCAACCCCACCAGCAAAACAGGACAGGCTCAACAGA
GCTCAATTATCAAATGGAAATGGTACATTCAAGATCGGGCCCAGCCAGGACCCCAAGGGA
CCAGCGGGCTCCATGAACAAATGGCTAGCTTACCAGAAGGGACCAAGCGACCCGTAGGGG
ATGCTTTGGCTCCTCCTGTGGCTACCTGGGGCCCAAGATTCAGAGACATGCCTACCGACG
GTATGGCATGGGGTTTACTGACGGCTCTGCGAAACAACAAGCAAGAGGCTCCACCGGGCT
GTGGACATCAACCAGCCAGTGGATGGCCATCTTTTGACTGAGACTGGACATGGACGTTCT
GCCCAATGGGCCAAACTACATGCAGTGGTGATGGCCATGCAGGCCGCCCCTACCACCATA
TCTTGCTACATTTTCACTGACTCATGGGCCATTGCCAACAGCCTAGCCATCTGGTCAGGA
GAATGGCAACTGAGTGACTGGACTATTAAAGGATCCCCTGTGTGGGGACAAGGACTATGG
CAACAGCTTGCTGCCTGGAAGGGACAAATATATGTCACTCATGTGGATGCTGGGACTACC
ATGGCCACCCTTGAGAGGAATTTATGTCATGTTTTTGGATACCCCATGGGACTTCACTCT
GACCAAGGAACATCCTTCACTGCCCAAGCAACATGACAATGGGCACACTCTCATGGAACA
CGATGGACTTTCCATGCACCCTGTCATCCACAGGCCAATGGAGCTATTGAGCGCCAGAGC
GGCCGAGGAAGAGCGCAACTGAAGAAAGGACATCAAGACGACCTGCTAGCGGAGAGGAAC
CACCAACTCAAGAGGCCAATATCGACACAAAACACTGCACTCCAATGCAAGGGAAACACG
GCACTGCAGCACACGTCGAGAAACACTGACCGCGGAGAGGAGCGAGACACACCAGGCAGC
CGCCGAACTAGGCTACCCCTCAAAAAACCCAATCTCAATCTTCCCAACCATCCACTTGCC
TGTGTACCTCAACAGTCCACGACCCAGGACAAGAACGCGGGACAGGCCACCAAAGGACCC
CACAAAAAGACCCCCAACACAAACAGGGAGGAAACACGCCCCCTGGGGACTCCCCTGCGA
GCAGATCCCACAATAATTGGGGAAGAGACCAAGCCTTTGCAGGGGGCTGGGGTCCTGCGT
GACGCAAGTGGGGGCAAGAACCACTGAAATTTGGGTTATTGCGGTTAATGTGACCCCATT
TGTAAAGTTTGTACAGGACGCCACCGCCCTCCCCGGACCCGACCCCACAGGCTGAAAGCG
TGAGGCCTGGGGAAACACCAAGGGCAATGGTGCCCACAGAGGTAGTAGCCTCAGGACAGG
GACAGACAGACTGGGTCGCTACACCAACTCAGCCCAACCCCTATCTGATAGGTAGGGAAC
ACCTGACACCCTGGAAGAGAGCCGGCACCGGTGCCGGCCTGTCAGGCTGCTCCACCCGCA
GCAACATGCAAGCAGCCGACTGGAAACAGAACCCCGACCCCACGCTCGCCCAAACACCCC
CCACCGCGCCGAACCTGACAAAACCCTGGATCAGCCATCCCAGACAGCACGCTAGCACAA
ACCAAACGCACCCTCCCAGCCCGACAGGGCGAAACGACACCAGCAGGCCCAAGCAAAAAG
CGGACAAAAACACAACCCGACACAGGGGTTACAGAGCAAAAAACCGACACCCCAAACATC
CCAGGAAACCGGAGGTGCCCTGTTTTAACTTAACTGACTTAAGGTGGCAAAATGTCACGA
CCACAACTAACAAAACCTTGGTGGGCTGGTACTTTGACGCACCACACTCCTTTGATTACA
TGGACGAGAAGTGTCCCAGTGGCGACGACGAAAACAAGGACCGGACTATTGCTAGCCCTC

TGTGTAGGGGCTTCATGAACAATATTGTATGGGGAAAACTGAGCTCATGCAACTATGCCA
TCAATGAGACTTGGCTGGTGAATGCCAATGCCTCCATACCCATGAATGGGTCACTGAACA
ATAAAATGGGAAAGGGTGTGCTGTGTGCACCCGAGGGCTACATCTTTCTCTGTGGGCGGT
CCGGGAGTGACCCAAATACGGGATGGGCAATGTCATGCCTGGAAAGCTGGCGGATGGTGG
GATCCTGCACGTTGGGCGTGCTGGGGGTGCCCCTGGATATCACCCCTGGGAATGAGATGC
ACCATTGGGCCAGCAGCCTAAAGCTGTACACCAGGCTTACTAGGGACCTGCCAGGAGGTG
TAACTGACTCTGGGTTTATGTCCTTTATGAGATCTTTGGTACCATACATAGGAGTCAGTG
CTCATGAAAAAATGATAAGAAACCTGTCCCTGACCATGGCAGATATTGCTTCCTCCACTG
CCACTGCCTTGGCAGCCCAGCAGACATCCCTCAACTCCCTTGGGAAGGTTGTTTTAGACA
ACAGAATTGCTCTAGACTTTCTTTTAGCCCAACTGGGAGGAGTGTATGCAATTGCCAACA
CCTCCTGCTGTACCTGGATAAACACCTCAGGTATCGTAGAAACACAAGTAGAGGAGATCC
GGAAGCAGGTTCACTGGCTGCAGACAGTGGGGCCACCTGAAGGATCCTTCTTTGACCTCT
TTAGCAACTTCTTACCTGGATCACTGGGATCCTGGGCTAGGTCACTGCTCCAGGCAGGCC
TGATCATCCTGCTTGTGGTAGTAGTCCTCCTGGGCCCAGTGAAATGTATTCTGGCTATGG
CTCAATGATGTTGCACTGAGATTGTGTCAGTCAAGGTGCTACATCAATCTGACAAGACAA
ACCTCTGCCTCCAGATCCGGGGAGGTCGGTGGGCATATGAAATGGACTAGCTTTGCTAAG
GGGGATATCTGGGTTGGGGG

>3_cons
CAAAACCGCCCTGTTTGAGAAAGAAGAAATGGACACCCTGAAAGCACCTCGCGACCCTAA
CGCAAGACAAAAATACGTCAGGCAGACAGTGAAGGTAGGCCCCCACAGAATGGGAAATGG
CACCTTGGAGTCACTGAGGGTAGCCTCCCTTAGAAGGGTTGAGTTCTTTCAAACAGCCCA
ACAGCCCACAGCCACAGACAGAAAATCTAAGTCACCCTGAGGTATGAAAATAAAAAAAAA
AAACTCAGGCGGAACTTGCACCGGGGGGCTGGCCCAAGACAAGCAAACAAGGAAGGTCTC
TCCAGATCAGCAAGGCTGCACAGCCCGGGGCTAGCCCAGAAGCCTTTTGTTCTTTGTGTA
ATTAACATGCCCACAGGGGAAAATTCCCTCCCCTTTTCAGACACATGCATAGTGGGCTCC
AAAGGAACATAAACAAATATGGAGGAGCAATACCAACACCAAACTAAGGGTCATACAAAC
AAGAGAAGCAGCGCTTTGTGCCAACCGAGAGGCATCCTTACCAGAGCATAACAAAAATGG
AGTGGAACAATCCTCCCTCAAAGAAACACTGCGCCTTATGCAATAAGAAGAGTCGCTCAG
CAAGCGCCACACAGCACTAGAGGACCTTAAAATCTGCAGCACTAGTATTCAACAGCGACA
CCATGTGCCAATGAAGAAAAATGGCAATAGTGGACTGATCCCACTTCCGGGCTCCCCTCT
CTGCTGCAGAGAGCTTTCCTCTCTCAAGAAGGAGACAAGACATGTTTATAAATAGCCAAA
ATGCAGCTTAGTAAACTTTCACTCCAACCTCACACTTCGAGGGAGAGAGAAATAAACGAA
GCAAGAAGCCTCAAGAGGAAGAGATGTGTCTCCATGCTCATTAATAACGCCGGCACTATA
TTTCCTGATTGTAAGAACAAAGAACTCCGGATAATACCTCACAAGAAGAAATTTAGAAAC
AGCTAATTTGAGGAGCATTGGCGACACCTAAACCTGTAACA

>4_cons
TGTGGCAAGGATAATATTTTGAGATATTAATTTATGTTTTGTTCTCCTCTTGGTAAACCT
GTATTTTCCCCTTCCCACCTTCCCCCATTATGGCCCCAAGCAGGTAGCCAGGCCCTTGAT
GACTCATTGCCTCGGGGGAGGTATGTGCCGCAGGGGAAGTCCATGCCAGAAAAGCAGAAA
GAATGCTGCATGAAGTCAACAACTCTTCCTGGCTTTTGTTTTCCAAAAGCCTAAGCCCAT
TGGAGGGAGTATGCTAGGAGACCTGGGGAGGAGGGGGAAAGAGGTTAGAGGCTGAGAGGG
AACCTGAGAGGAGGCCGGGCTCCCTCCCCCCCAGACAAAGAGGAAGAGATTCCCCTGGGC
TGAAACCCAGGAAGGAAGGCGGGACAGGAGCTTTAAAAAAACCCAGATAGGTCTGGGGGA
ATCCTGGGAGGAGAGCAGGGAAGGGGACCTGTGCCCTGCTTCCCGGCAGCGCAGCCCGGG
GAGGCGGCAAGACTCTCAGAGAGGCCCTGCATTTGGCTCGGCGCCACGTCCAACATGGCG
CGAGAGCGGTAGAGCAGCAATGGCTGCGTGGTGTGTCTAGGCAGACGGGACCGAGGGCAG
CCCCACCGGGGCTCCCAGCCTCCCATGGCCTGCGTGTGGCACGGAGGGGAGCCAAAGATT
CCCGAGTGCCCCAAGCATGGCACGGAAGGTGGGCTCAGAGAGAGCCGAGGCAGAAAGCAG
CAGAGAGCCGCCGGCCCTGAAGGGGCCATCACAGAGGGGCAGAAGAGGGCCTACATGCCT
GGGAGAAGGAGCCCGGCGACCGGAGGCCAAGAGGGAAATGGCCACCACCGCGGACACCAG
CAGGGAGCGGAGGCGCATCACGCCCAAGGACCAGACGGGACAAGGGACATCTCAGCGGTC
ACCAGCACAGGGAGCAGACCAGACCAGCCACTCCGCAGCAGAGACCAGTGTGGATCCCGA
TGACGCCACGAGGACCAGAGGCCCCCCCCCGATCCCCCAATGCCACGGCAACTGCGTAAG
CCCACACTACCCCCGGAACCTTGGCACAACCCTGGGGAGAAGGGGAGGAGAGGGAGGGGG
AGAATCCTGAATTGACTGAGTATTTCCCCAAAAATGACTGAGTCATAAAAAAGAGACTAA

TTTACCTAAAAGAGACTGTTTAAATCACTGGAATTGACTGAGTTTACCTGGAAGTGACAA
GATTAAGTCGTTCTCACGCTTCCCGCCAGCCCGGCGGGATGGGGGCTCGGAATCAGAAAT
TAAGTTGAGTTATAGAAAATAAAGAAATGTTACATTTTCCTTGCACACCTGAGTTTGTGG
CGAGTAAGATTGCATACCCGCTACA

SVA consensus sequences:

>1_cons
CTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCCATGGTCTCCCTCTCCCTGTCCCCTCTTT
CCACGGTCTCCCTCTGATGCCGAGCCGAAGCTGGACTGTACTGCCGCCATCTCGGCTCAC
TGCAACCTCCCTGCCTGATTCTCCTGCCTCAGCCTGCCGAGTGCCTGCGATTGCAGGCGC
GCGCCGCCACGCCTGACTGGTTTTCGTATTTTGTTGGTGGAGACGGGGTTTCGCTGTGTT
GGCCGGGCTGGTCTCCAGCTCCTAACCGCGAGTGATCTGCCAGCCTCGGCCTCCCGAGGT
GCCGGGATTGCAGACGGAGTCTCGTTCACTCAGTGCTCAATGTTGCCCAGGCTGGAGTGC
AGTGGCGTGATCTCGGCTCGCTACAACCTCCACCTCCCAGCCGCCTGCCTTGGCCTCCCA
AAGTGCCGAGATTGCAGCCTCTGCCCGGCCGCCACCCCGTCTGGGAAGTGAGGAGCGTCT
CTGCCTGGCCGCCCATCGTCTGGGATGTGAGGAGCCCCTCTGCCCGGCCGCCCATCGTCT
GGGAAGTGAGGAGCGCCTCTGCCCGGCCGCCATCCCGTCTAGGAAGTGAGGAGCGTCTCT
GCCCGGCCGCCCATCGTCTGAGATGTGGGGAGCGCCTCTGCCTGGCAACCGCTCCATCTG
AGAAGTGAGGAGCCCCTCCGCCCGGCAGCCGCCCTGTCTGAGAAGTGAGGAGCCCCTCCG
CCCAGCAGCCACCTGGTCCGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCG
CCCCGTCCGGGAGGGAGGTGGGGGGGTCAGCCCCCAGCCCGGCCCGCCGCCCCGTCTGGG
ATGTGAGGAGCGCCTCTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCACTTTGCCCGG
CCAGCCACTCTGTCCGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCC
GTCTGGGAGGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTCCGGGAGGG
AGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCGTCCGGGAGGTGAGGGGCGCCT
CTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCCCTCTGCCCGGCCACCACCCCGTCTG
GGAGGTGTACCCAACAGCTCATTGAGAACGGGCCATGATGACGATGGCGGTTTTGTCGAA
TAGAAAAGGGGGAAATGTGGGGAAAAGATAGAGAAATCAGATTGTTGCTGTGTCTGTGTA
GAAAGAAGTAGACATAGGAGACTTTCCATTTTGTTCTGTACTAAGAAAAATTCTTCTGCC
TTGGGATGCTGTTGATCTATGACCTTACCCCCAACCCCGTGCTCTCTGAAACATGTGCTG
TGTCCACTCAGGGTTAAATGGATTAAGGGCGGTGCAAGATGTGCTTTGTTAAACAGATGC
TTGAAGGCAGCATGCTCGTTAAGAGTCATCACCACTCCCTAATCTCAAGTACCCAGGGAC
ACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTTCAC
TTGTTTATCTGCTGACCTTCCCTCCACTATTGTCCTATGACCCTGCCAAATCCCCCTCTG
CGAGAAACACCCAAGAATGATCAATAAAAAAAAAAAAAAAAAAA

>2_cons
CTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTCTCCCTGTCCCCTCTTTCCACG
GTCTCCCTCTGATGCCGAGCCGAAGCTGGACTGTACTGCTGCCATCTCGGCTCACTGCAA
CCTCCCTGCCTGATTCTCCTGCCTCAGCCTGCCGAGTGCCTGCGATTGCAGGCGCGCGCC
GCCACGCCTGACTGGTTTTCGTATTTTTTTGGTGGAGACGGGGTTTCGCTGTGTTGGCCG
GGCTGGTCTCCAGCTCCTAACCGCGAGTGATCCGCCAGCCTCGGCCTCCCGAGGTGCCGG
GATTGCAGACGGAGTCTCGTTCACTCAGTGCTCAATGGTGCCCAGGCTGGAGTGCAGTGG
CGTGATCTCGGCTCGCTACAACCTACACCTCCCAGCCGCCTGCCTTGGCCTCCCAAAGTG
CCGAGATTGCAGCCTCTGCCCGGCCGCCACCCCGTCTGGGAAGTGAGGAGCGTCTCTGCC
TGGCCGCCCATCGTCTGGGATGTGAGGAGCCCCTCTGCCCGGCCGCCCAGTCTGGGAAGT
GAGGAGCGCCTCTGCCCGGCCGCCATCCCGTCTAGGAAGTGAGGAGCGTCTCTGCCCGGC
CGCCCATCGTCTGAGATGTGGGGAGCGCCTCTGCCCCGCCGCCCCGTCTGGGATGTGAGG
AGCGCCTCTGCCCGGCCAGCCGCCCCGTCTGGGAGGTGGGGGGGTCAGCCCCCCGCCCGG
CCAGCCGCCCCGTCCGGGAGGAGGTGGGGGGGTCAGCCCCCCGCCCGGCCAGCCGCCCCG
TCCGGGAGGTGAGGGGCGCCTCTGCCCGGCCGCCCCTACTGGGAAGTGAGGAGCCCCTCT
GCCCGGCCACCACCCCGTCTGGGAGGTGTACCCAACAGCTCATTGAGAACGGGCCATGAT
GACAATGGCGGTTTTGTGGAATAGAAAGGCGGGAAAGGTGGGGAAAAGATTGAGAAATCG
GATGGTTGCCGTGTCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTTCTGT
ACTAAGAAAAATTCTTCTGCCTTGGGATCCTGTTGATCTGTGACCTTACCCCCAACCCTG

TGCTCTCTGAAACATGTGCTGTGTCCACTCAGGGTTAAATGGATTAAGGGCGGTGCAAGA
TGTGCTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCACTCCC
TAATCTCAAGTACCCAGGGACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGA
AAACCAGAGACCTTTGTTCACTTGTTTATCTGCTGACCTTCCCTCCACTATTGTCCTATG
ACCCTGCCAAATCCCCCTCTGCGAGAAACACCCAAGAATGATCAATAAAAAAAAAAAAAA
AAAAAA

ERV consensus sequences:

>1_cons
AATTTGGTGAGCCAGCCAGGAGCCGCTGGGACGGTGATGCATTCAGCGGCCAGCGGCTTG
TGACGAGACAGTCTTCAGGAGACTCCCAGCAGCTGCTGGGTGAGATTATCCAGGGGACTC
TCCTGAGGGCTGTCCCTTGGACAAAACCACACATCCCTCTCACTACTGGGGAAGAACGGA
GGTCAGGAACGGACGTGCTCAAGGGGTGAGTAAAACTGAACCTAATAAGGGACTTATTAT
TTTCCTATCTGGGCTTGTTAAGCCATTTGTCCGGTACCACCAGGGAGACAATAGGGCTCA
TTTGTACACCCGCTTTGCGTTTGGTTAAAATCAGGTTTTGAGTTGGTTTTGAATCTGTTA
TGTCAGCAAGACACCCAGGAGGTAGACCCATAGGTCCGGAATCATCCTACGCTTGTTAAT
GGATCGCTCTAGGGATTGCGGTACCAGATAGGACCAACGGTGAGAAGACAAGATCCAGTG
ACAGCAGGGACGACGGTGGACAGAAACCAGTCGAGAGGGAAAAAGTAGCGATAGCAGCTG
AGAGCGAAAAGGAAAGAAGAAGACCATTACACCGGCGGCAGAGACAAGTAGGAATTAGTA
ACTGCGCGAAGGGGTGTGAATGCACAAACCCTGAGGGATGCACCACTGTTTAAGGTGGCC
CCGGCCCAAGACAAGCCCATAGAAACCGACTCAAAAACCTACGCCCACACACTGGCCTGA
CGGCACCGCCCCGAAGGCAAGGAGCACAGACACGCCCGGCACGAGCCCCGGAACAACAGC
GTAAACCCCCACTTCAGCGGAGGGAAGGGAGGGACAACCACAACACACACCCCCAGTCCG
TCCGAAGCCCCCTCCGGCCTACTGCCCCTGGCAACAGGCAGGGGCCCCCCCATGCCACGC
CAGCATGAACGCCAACAGGATCCGGCAACGCAGGCGAGCCGACCCGCAGTCATAATGGCA
CCCTGTGCTTACAGGGCCGACACTCAAACGCGCCAGCGCAGATCCGGACCCCTCGCCGGC
CGACGTCAGCGGGAACATCCACGAGCCAAGGCGTTACCGGGGTGGGATGAGCCGGCCAGC
AGGCGGGACTACTCCACCTTTCGACTTCCGAGTGAGGGCGCTGGCAGCCCTCCCAGCCCC
CGCACGCTGTCGGCGCCGGACCCGCGTCGGCGCGCACTCTGGCCGCGCTAGAGAAGCCCA
CCCTATTCCTGCAGGCCGGTGGATGCCTGCTCCCCCCGCCTGGCCTGGCCCAGCACAAAG
CCCGGGCTCCGAGCTCGGAGGAAAGTTTGGGCCGAGCCCGGGCGCAGTCACAACAAGAAC
GGGTATGCACACACTCGTGGTAGCGCTGAAACCCCAGACACCAGCCGCCTCGGCCACCTC
CAGACATTTGGCGCCACCGAGCATGGGCGCCTGGCCAACGAGGGGCGGAGGGTAGCTAGG
CGCAGGCCTCCAGGCGCACCTTGGCCTGAACAGCCAGGGTGCCATGAACAACACCACAAA
ACTCACAATAGCGGCATTCCCGAGTCCAATGCACACATTACTAAGACTTCTCCGAATTAA
CCATGCTTACACCCACCGCTTCTTGAAAGGTTTAAGGAGGCAGACAAATGCCTGGGCCCC
GAAATCACTATACCGGGCAGCACGAACTTCCACAGGGGGCCACCCCATGAGGCAGAAAGG
AAAGTTAACATCAGGGCATCCCAAAGGAATGACACCCAAGCAAAGCTGTACTGCAGCAAA
CTTGAGACGGAATGAAAAAAAAAAGGCAGGGAAAAAACACTCATGGCAACAGCAGGAGGA
AACCCAAGAGGGAAGGGACCCGAAAAGCGGAAAGGGAAAATAGAAAAGGATCAGTGCGCA
TACTGCTGGGAAACGGGGCATTGGAAAAAGGATTGCCCAAAGTTAAGCCAGAAAGAACCA
AGGCCAATAATGGCAGTTAAGCCCGGGAATGAATCTGAGGAAGATTGAGGGTGCCCAAGA
CTCCCAGCAGCTCCAACTCTATCTGACATCAAAATTTCCCCACCGGGTCCCTTACTAGAC
ATGGGGACAACCAGGAAAAGCCCACAGGACGCAAAAAAAAAAGTGGACTTCAGACAAGACA
AGCAAACTAAAAATCAGCTAAGAAAAAAACAAACTAATGAACTATCGCAAAACCCTGTCA
GTATAATACGAGTAACCACAAAAGGGAATTCGGCGAAACCCCACCTTCCATCCTGAGGAA
AATATCGCAAAAATAGAATTACTCAAAAATACCCAAGAAAGATAGAAAGACCGCTTCTCT
TTCCAGGCCCACATAGGCCATCCAACTAATAGTCTCATATCGGCTACCAGTCGGAGATAC
GGCACCTTTTCCGCCAAAGACAGCCTGAAGCCTGAAAGCCAGGCTACCAGTTCCAGCTGA
GACCCGCGGCCCAACAAACTGATTGTGAGTTGATCAAACAGCAACACGGCGACAAGGTAA
AACCAGAGGTCTGGGCATGAGACCGACCCGGGAGGGCAATTAATGTGAGTCCGGTAAAAA
TCAAACTGAAGGAAGGGGCCCAACCTATCCGGAAAAAAAAATACCCCTTAAAGAGGGAAG
CCTTGGAAGACATCCAGCCAGTCTTAGTCCAGTTCTTGCAGTATGGCCTAGTGAGAACAG
GCATTCCTTTTCTGCAATACTGTCTGGCCCCATTATCAGACAACAAGCCCAGAATACCGG
GCAAACAAAGCCACCCTCAAATGAGAAAAGATCTACCAATTATCTAAAACATGATTTTGG

GAGTGCTTTCCTCTGAAACCCCCCTACTTCTCACCTATTATATGTACACCCCCCCCCGCA
TGCTTCGTTGTTAATACTGTCTTTTCCCCCCCTGAATGGTGTCCTAAGGCAGAACTAGAT
ATTATAGATGACCCCCTTTTACAAGGGCCACCTGTCTCTCAGGGTGAACAGCAACCGCCC
CCATATAGCCCCTTGCCAAGTGCTCCTGAGGCTAAAACCCAGGAGCAAACACCGGGGACC
CTACTAAGTCCCCCTCACACTCGGAGGGGAACACCGTATTCAACCCTTTTTCTGGGCCTC
AAGATAATAACCCCAATCCGAAAAAGAGAAATTGGAGGCAGCCACTCATGGCCGCCCATG
GAGACAAGACACTAACCCATCAGCCCGCACTCCCCCCTCTAACCGATATACAACAATGCA
AGGAAAAACTAGGAAACTATTCGGAGAATCCAAGGAAGTTTAAAGACGAGTTAAATAAAT
TGACCATGGCCTTTGATCTACCCTGAAGAGACCTACAATTAATTCTATAGGCCTCCTGCA
CACATGAGGAAAATCAACGGATTGTAGTTCCACAGCAACCGGACGCCCAACCTGAGCCCA
TAAAAACTGCCCAGGCTCCACAATTCTCGTCATTTACTGTGCACCAGGAGCAAAACATAC
ACACCTAGACCAAAACTAAGAATGCCAAAGAGACTACCAAGAGCTTGAAAAAAAAGGATA
AATTACAACAAGATCAACATAAAGGAATGAGAAAGTAAACAGGTAAGCAAATAAATTGGG
ACAAAGGCAGAGAGCCAGACGGGGGGAAGCCCAGAGGAGCAGCCAGGCTGACAGGCAGAG
AGACGACCGAACGTAGGAGAGATGCGAGGGAACGCCCGGGGTAAGGCCAGAGGCGAGAGC
TGCGGCGAGGGGGCCGGACCCCCAAATCCACCCAAGAGACAACCAGTAAGCTGCAAGAAA
CAGAGAGGGGCACACAAAATACTAAAGACCAACATTAAAAAACCGCCTTTATCTTTTAAA
AGAAACCGACCATGAAGAAATGAAGAGAAAAACAAAGAAAAGAAAAAAAACAGAACAAAT
TTGTGGCAGCCTTAGCCCCCCCGCTCCCTTAAGGCCACATAGCTCCAGAGAGCCACGAAA
AGTTTCCGTCCAAAAAGGACAGAAAAGAGTACCCCTTCACCTAGCACAAAGGTGAGAATA
AGTATAACAACTCTAAGCATCATGGTCAAGGCACCAAGGATTTACCCGCCGGCACCAGTC
AAAACGGAAAGCACCGAATCAATAACAGAGCTAACATCCTCCCAACAGCTCACTCAGACA
AAACTTTAGCAGCGAAAAAAACCAAATTACAAGAACATTGACCAAAAGGACAGCCGCCCA
TAAACGAATACAAGGGCCAACGGCAACAATTGTCTGCAAAATCCCCACGACCGAGCAGGT
CCGCTAACAAGAAGGCCACACCAAAGGCCACCATTGAGAAAAGCCAGGCACACCCCCCCC
CCACCAAGGTGAGGCAACGGCAGACTCTTCCCGGCCTCTTCCGATATTCCAGACCAACCG
GGCCCAACAAAACTCAAACAAGAAAACCAGTCCAAAAGTTAACAAAAAAGGAGACGGCCT
GGGAGTCGGACGAACAGCCCCAGACCACCAACATGGAAATAATAACTCATTATCTTCAGG
GAAAGATCACCTTCATACCAGGCTGTCCCCTTCCGGGTCCCCTCTCCCTACCGCATGTTA
AAACCGGCCCAAATGGTAAAGGGTGAAAATGAACGAACCCCACCAAGGGCTTGAGTTAAA
CGGAAGAAAAAGGGTACAAAAGATAAAATATACCTCAGATATTTTATACATTGTAAAACA
ATCAGCTAAATACTGCATAAGATGGAAAAAATGCCCAGCGGTCGTTAAATTCGCATGGGA
CCTTGGTTCTTGGAGGTGGTCTAAACATAGCTATTAAAAGTAAAAATGTTAAACACATGT
GAATGGAATAGATGCTTAAATAGTGAGCTTATTGTACGCCCTGAGAGCTGAACACTCGAT
GAGACAACCCGCCTACAGATAAGATATACCTTCAAAGAAGCGGAACACCACCGGATCCCT
ACCCCCAGGTAAGGGACGCCACAAACCGACACTTCCTACCGGGGAGAAAGAATAACCAAT
AAATAGGAACTTATAAAAAATAAGCTTCAAAATATAAAGAAACCAATACTGAAAAATCAG
AAACATAACTAACACAACCGAGACAAACTGGATAAACGGCAAAAGAAGCATAATAGAAAG
GAAGCGAAAAGCAGAAAATCTCTGGGAGACTAAATACTTGTAAACACACCAAGCATCTCT
TCCTATCTGTCTCATCACTCAATAATCCGGGAAAAAAGAAAACACCAAACACCCACAACT
GACCCAAGACAAAACCATGGGCTCTTAAAACCCTGCAACCCTCCTCCGCACAAACATTCC
TCAAGCCTGTGATAAAGCTCTCCCGTTCATAAATTGTCTGCGGAAACAAAGATAAGGAAA
GGGAATCGGAATTTCTAAAGAGGGTCTTAGCAGAAGCAAACAACGCGGAGCATGTAAGTG
ACCTCATGGGGACTTGGAGAACTTTTCAGCAAAAGGTTTGAAAAGCTCTTCCTCAACATC
CTCCTAAACAACACTTGCATGACCTAAAACATGGAGATATTCTCTAAAGGACGCCGGACA
AGAACAAACCGAGCACAGACACAAAACGAAAAGATACCGGCTGCTCAAAAACTAAAAACC
GAGATTACCTAAGGCATTTATTTATGAAATGGCAAAAGGCTATGACTTCAAAGACAGTAC
ACTACGACAGTTGAGCACTTCCTCGAGGCTTTACGGAAAGCCTTAATTCCTCTAATTGAT
CAAAAGGGCTGGATGTTCACCTACAGTTTTGGGAAAATGGCATGATGCTAGAGTATTTGG
AATTCGATCTTAAAAAGAACGGAACTATAGAACATTATGAAGGCCATATATATAAACCTG
TGCATAGCATTGAAGAAATGGAAAACAAAGTGCATAAAAAGAAGTTAAAGGCGAAAGACC
AATCGCTCACCTACATACAGAAATATTCAAGAGCGGTAGCCCGGCAAATACACCTAGAAC
TAGTACACCCGAGGTGTAAACTGAGGCCAAAGAACAATAAGAACTAGCTCAGGCCATGGC
AGGCAAAAGACGGTTGGAGAGGCAGCCAGATGCCACAATTATGGCTTATAGCATATCCAC
AAAGGCCACCAACAACAGGGGATTAAAAAAAAGTCACAGTAAAGACGAAGACCTTCTCAA
GATGGAAAAAGGAATATCCAGACCGGCTCAAGGGAACCCACCAAAAAGGGACCACAAAGC
AAGCTGTAAAACTCTCCGCAACCTGCCCCCTGCCTTCAGGGAATAACTAGTTGGGTACAA

CTGTCCAGGATTAAACCTGTTTCTTATGAGTCACAGGCACAAAAGGAGGACACCATGACC
TACATCTGTGAACCTTTGGAAGACTTCCACTACCTATTTAAAAGAATCAACACTCAGCCA
GAAGTGGTAACGTGATGCTGACGGAAGCAAAAGTAATAATAACCCTTCACTAACTACTGA
ATTAACTAAATATGTAAAAAAGAAGAAGGGAATAACATTGAAACTGCAAACACCGAAATT
ACCATAATAACCCGGACAAACCGCCAAAATGACCAACAAGTTAAAAAGAATCAACTCCAA
ATCGTGTAAGCACACCCAGTTAAAACGAACCAATGTAATTGGAATCGAAAAACTCAAGAA
AATAATACAAAACAAAAAACAGAAAGAAATAACCCCCAATGAAGTTGAATTAGTCATAAA
CATAACCGGAAACAAGTGTACGATCACCGAAACGATCCCAAAAAACCAGCAAACAATTCA
AAACAATCCCGATCAAACCAGAAAAACCCCTAACAATCTAAACACCTTTTCGTCAATGCG
AACCGTCCCAAACTCAGATGATTCCCCCCAGCCGTTCAAAACTGATAGAGAAAGAACCTG
GAAAGAACAGAAGCAACAGCTTCCCGAGCAAAAGGAGGAAAACAATTGGACCGACCGGAG
TGCTGAAGTCTCGAAAGCCGAAGGGTCGCAGCAATAGGAAGCCGCAACCCCCGGAGCCCC
GTGCCAGGGGGCGACACCCGCCCTGAAAAGGGGAAAGGGCCCAACAGAGCTGTAACCATT
AAGCCGTCCGCGGACGGCAAAGCTAAAAGAGAACAACAACAGCCACACAGATCATAAATA
AACATGCATTCTGGGTCATGCACATGCCAAAAACTGCAAAGAAAGGGCACCACAAACTAC
CAAACACTAACCTATAAAGCAAGTGAAAGAAATTCCAAAGCCAGTGTAAGCAATACAACA
GCCACCCTCCGGACCTATAAAATATTAGAACAACACCCAATGTGTCTGGACCCGAATAGA
GCCGGAAACAAAGAGGGCCACACGACCCAGCTGTGCCAACCCCCTTGGGCCTCAGTTTGT
GCCCCAACTGGGCTCATTTTTAGGAGAACACATAAAAAGGCAAAAGCCCGCGGCCGCTGG
CACCCCCCCCGTAGATGCCGCCTTTTGCGCCTCCCTGTGATCCTCGGCACGACCTCGCCGA
CCATATGAAAAGAATGGAACAAACAGGAAACTAACCTGGAAACCCTAGACCCCCGGGCTG
GATAAGATAATAACCACGAAAAGAACACTAATGAGTCAGAAAAACCCACAATCAATATTT
CCAATTACGAAGAGCACCAAAACTCACCCTCGGAAAAAAAAACAAAAGAAAACATAAGAA
CCTGAGAAACTCCCGTACCAGAGCCCAAGGCAGAAGCCCAGACTGCCAGAACTCAAAACC
ACCAATATAAACCCTGCTTCCGCCACGCCTGGGCCAGCAGCAGCCTCCCAGGGAGCCGGC
ACCCGTGCCTGAACACACCTACCAACAGATGATACAAGAAAACAAAACAATAGATTGAGC
CGGCGCCTGGAGTCGCTCGCCCCACCACAGCAACCGACCCACACGGACAACGGCAGCAGC
ACAACACCCAGACACAAAGACACCGAAGATACAAATATCAGTTCCAGAGACCTGACCATT
TTTGGCGAGTGGTGGACACAAGCACTAACACAATAGTGCAAACTTACGAGCTTAGCAAGG
CGGAGGCCGGATAATTAATAGCACAATATGTCTCCCATTGGAGACACGGAACATATATAA
AAACGATACTAAACTCAAAAAAGACAATACGGTGTCCCAAGAGTAGGCACATGAAAGTCT
ATAGCAACTATGAACAAATGCCACCATGCTGCAGGTACACAGAAAACTCTGGACAAACTA
AAAAAAACACACAAGACAAAAATTATCACGAAACAACAGAAACCACCAACACCGGACCCC
CAAGGACCCTGCGAAAGAAACGAAAAACCCCGAAAGCAAAAACTAACATAGACACAATTG
GGACCTTATAACTAACGATATCCATAACAGACAGAAAAGTGACTTACAAAAACAAAACAG
GCTTGGCATTATACCAGCTGAGCATAGAACAATCTGCGCACAACTAACTTCTGGCGCTCG
CTTGCATGCTCCCTCCCGCTTGGGGCCTAGCTCGGCCTCGCCAGTCGTGAAGCCAGCGAC
GGTACGGGCACACAAACTCCAAAAGAGATCGACTACTCAACCAGGGCATCACAAACCTCG
GTCACCGAATTACATCCTACTGGGAAGACATAACACAGGCAAAGAACTACAACCAAATAG
CCCGACACATCTTAAAAGAAGAAAGCCAGCAAAACAGGAAGCCCCGGACCCAAAACCGGA
TGGTCCCAGAAACAACAACACCTGCCAAACGGGGCAACACCGCAAGTACAGGAACACAAG
ACAGCAAAAAGATAACTAAAAACCTACCAACTTTCCTTGGTGTCTTTGTTGCATAGTTAC
TGCGGGCTGAGCAAACCAGGCACCCCCTGAGAATAACAATGACAGCCACAGGGGATCACG
GCCGGTCATATCCTAAGAGAGTTGGGCATACAAATATCCTGAACAATCAAAGGCTGTAGA
CATCAGTGGGCCAAAAGAAACCAACGCAGCACAGCAAAACTCAGAGAGATGTACCACTGA
CCGCAAAGGTCTGCAGCTCGCTAATAGGACCAAAAATCGGATAAACAAACTCAGTTCAAC
TCGCCCACACACCATCAAGCGCCAAAAAATCAACCAAAAACGCTGACAACAAACTCCTGC
CAAGCACTTGCCACGTGACACACCACATGAAGACGATCCCCTGTGATAAGGGGCGGGTGC
ACAGCCACTGTCAGTCATCAACCAGGATCCAACAAATGATCATCCAACAAAAAGAACAAG
GATCCTGCCCCTTCCGGCGCGCGCCAAATCTGCCGAAAGTAAGTATAAAACAATGACACA
AGAAAAACAAACAGTAAACAGGACTCAAAAAAAGAATTAAACAAATCACAAATCGACAGT
GGGAAATCCTTCAGATCGGGATGAAGAATGACCAAAGGAGAAAT

>2_cons
TACTTTGGTGCCGCGTGACTCGGATACGTTCCCTAGTGGTAAGAAACCTCTATGCCTCGC
CTTCTTTGGCTGGAGGCGTTCAACCCCCGTATGCAGTTTTCTTCTCCCCCGGCACCCCAC
CGCGGACCAACCAAACCCCAAAAAAAAGCCGACCGGCCACAGTCGCTCCGCTCCCCCGAG

CCCATAAACCACACAACACCCAAAAGGAAAGAAAAATACTGGGAATGAGCATGGAAGCCA
AACCGAAAAAGCCAGGGACACCAACAAAAGACACGGAGAGGAGGCTCACGGGAGAGGAAA
GGGTAAAGCCAAAAGCATCCCAAAATCCGGGGTGCACTCGTGGAGTCTAAAGAGAGTTGT
CCCATGTAGCAAACCGGACACCACCTGGGAGCCAAGTCTTGACCAGCAGGGAACTCGAGG
TCCCACTCGGGACCCCAGCTCGCTGGCAGCAAAAGAGCATCATCCTTTGTTCTCCCGCCG
ATTCCTAACCTACGCGTATTGGTGATTGCCTTTATCCTTTTTTTTTTTAGCTCATCTCTC
CTCCTAACTGGGATTAGGAGTCATGACAGTGCCTGGATTGGGAACAGCAACGGACCGGAT
TTGGTAATTAACCGCCTTGGTATCAGTTAGAGGCCCACTGAACCTCTGTGGGTCAGAAAA
AAACTGGAAGTTAAAAGAATTAGTTCAGGGGGTGTGACAGATTACAATAATTTACCAAAG
ATTTTTCAAACTTTCCATTCCGACGAGAGAACTTCTCTTGTATCTTTGGGAGAAAACTTT
CCCTCGGCAAATTCCCCCACCCTGCACCCAAGTAAGGACCCCACCTACCCGCCACCATGG
AAGGAGCAAACATAGGCCAGGAAAGAAAAGGAGTCCTTCGACATGAGCGATCAATTAAAC
CTAACAGCAAGCCCAATATTAGACCAAAAAATAAGTTCCCAGGACTCACAAACACTGCCT
TCTTTAAGGCAAACACGGCCACAAATGTTAGCACCTGAAGGCAGCGAAGCAATACAGAAA
TTGCCACATCCCACTAAAGCTAAAATAAGGCGGAACGTACCGAATGCTTAAAAACTCGGT
CTAGAGCATCTGGTAAAAAAGACTATCAACCAAGCCTCAGAGGCGAGCCCCAGAAACAAC
TCCAAAACTTAAGCCACACTTTCGACATCAACATCAAAAGGCTCATATAACAGCGTAGGG
AAGACCAGCTCCTCAGTAGGAGTTAAATCCTAGATCTGGTTAGCTCTTACATTACTCACA
AGACTGTTCAAAACTAGGGAGAAAGAACCACAATATGGCTGACACACTGTCCCTGGTGGG
AGTCACTTTTGTCGAGGAAAACCAATGATTATGTGAGATCCCAGACCCCTCATTAAAAAA
CACAATATTCCAAATAGGCTGCACAATCACAGCTTTAAAGCTATCTTTTCTAAGACTCAG
GCAGACATGAAGTGACTGATCATTATGACCAATTAAATAATCAAACCTAAAAATGAACAA
GGGGCCTAGGTAAGATCCTACCCCTCTCCAGCCCTAGAAAAAAACATACTTAAAAGTTTC
CTCAAAACCTAAGCGTCCTCAAAAATTACTCAACTGAAGAAACCAATAAGGACCACTAAT
GTAACTTCAGTCCATCAGACCCCTTTAAAACAAGCCAACGCAACCCCTTTAATTCCAATG
TATTCCTCGGGTCCCCACACTCAAATCAGTAACCCTATCCATCAACCACCAGCACAACCC
TAAAGAAAACCCATAAAAACACTCTGCACATTTTCACTCACAAACTGGAAAACATGAAAC
AGCCATTCAAAAGCCCTAACCCACTCTAAACTACATTGGCCCCCGAAACCGACTCCTTCT
GTCTCTAATACAACACCTAATTAATGCCCTTCCTACACATAATCATTTACACTTAATTTC
ACGCCAACACGAGAGAGACCCCGTATTGAAACACCGTCAGCCCATAGCATGGAAATCAGT
CATTTGGCAAACATTGAAAATCTATTCAGACATGACTGGAGCCCCAGTCAGGAGGCGCCG
CAAGGTGGGTTCCCAAGGCAGCTAAGGAGAAAACCCCAGAGAAACCTCCCGCGGGGGCAA
AAGGCCAAGCATATCTTACCAATCCAGCTTCTCCCACAAGATCCAACCATCCCAGAAATC
CTCATTTTAGGCAGGTCCCAGTCTCAATTCCCCTCCTATGCAATCTCGCCCATAGCAAAC
CAAAACTCAAAACTTTAACCTTAAATAGCAATTTAGGCCCTAAAAAACTTGATTCCTCTA
TCCCATCTCATGAACTTAAGCAAATCAATAACGATCTTGTTATGTCTATGAATCACCCCG
ATAAAGACATGTACCAGATCTGTAATGTAGCCATAGCATTTGAAGACACCTGGAAGGACA
GATCAGTCATCTTGACTAAAACCCTGACTAAAGAAGAAAATTAAACCACCAGGAATGTGG
CCCAAAATTTTGCAGATGAAATTCACACGACTAATCCTAATGATAATAGAGCTGGGGCAG
AAGCTTTATCTAATTAAGAGTTCTACAATAAAAATCCTTTATGAAAGCTAAAAAAACAGC
AGCTGCCTTAACCAAAAAGAGATTCTAGAGGACAAAAGAAGGGCCCTATATTGGATAAAG
AAGGTGACTTAGGAAGCTGGAAGAGAAACCACATGATCGTGTGCTTCCTCGAAGGAATAA
AAATGGCTGGAATTAAACCTTTATACTATGTAAAACTATCCAAAATAGATCAAGATCCTT
ATGAAAACCCTACTGCTTTCCTCGCTCGTCTATGAGAGGCCATGAAAAATTACTCCACCT
TTAACCCTAACACGGAAAGAGAAAAAAACATCCTGCAATAACAGACAAGGCTCAACCCGG
AACTAGGCAACGAGGACTCTCACAAGGAAATCTAAGAAACCCTCACCAATTGACTTCTAT
TCCTCAATCCTCCTCCTTTGGCTCAGCCATCACCAGATGCACGACACAAACTACACTTCC
TCCAACCCACCCTTGAAGATTCTAGATTTTGTAGAAACTACTTACCACCTCTTTGAAAAT
ATCTCGTACACTCGTGGTTAAGTAATAACCTTAGTTGAGGCTTGTTGGTTTCACCTGGGA
GGTTACTTTTGGTAAAGTTCAAAAGCCAGAAATATTGGCCGTTTGGCCCGGCTAAAGTCG
GGCAAAAAGCCAGCTGAACGGACTCAAACCCCACTACAGGACCTCCTAAAAGTAGCCTTC
AAAGTCTTTAAAAACCGGGACGAAAAAAATAAAAAAAAGAAACCTAAAAAGGAGAAGCAA
AAATAAAAGCAGAAGCTCCAACAAAAGGCTGCCCTATGGGCACACAACCCAACTCCAGGT
CGCCATCAGATCACCCCCCCACTCAACAGCCACCCATGAGCGGGACCAGGCAAAAATGCA
AAAGCCAACTGCCCCCCTTAAAGACCTCCAAAGAGCTATGGGCCGCCGCAGTCGTCAAAG
ACATACTAGCCTCCCGGATCCACCTGGTCCCTGTAGTAAATAAGGCCACCTGAAACACTA
AGCTAAAGATGGCCCAAGTAATAGGAAGTCCCCCAATTTTGAGTTTTGGGGGTATCAGAA

ATTACTTCGCATTATGAGAGAGCTTTGGTGTGTAATAACTAGGTAGGAAATACACATTTA
GGGATGGCTAATGGCAGTTATGGGGGATACTCGGCTCTTTGCACATTTGGATAAGAGAAG
CATGCTCTTGGCCACCTGGAAGGTATGAAAATGCCCTCCCCCCTCTGTAGGATAGAAGAC
CACTGGAAGGGCGACGGTGCTCACAGACAGGGTTAACGGGGGCCGGCTCCCGTTTCAATG
AAAGCCCAGCAATAAGCTCTCCACCGTCTTCCCCCAGAGGATTCACCACAACCTGGACAA
CCACAAGCCAAAAGAAAAACACCACCAACAATATTCCGGGACTACGAGCAAGCCACCGAA
ACCCACCCACTAGACCACAGAAGGGGCCTTGAGGCCACAGACTGTCCCCCAACCCCTATT
CCCTGTAAAACATACCCATGAAGTGCGAAAGAGTTCTGAGCACAATCCCAATAGCAACTA
AGAAACAAGTATAATTCTTTAAATTTATACAAGAAACACCAAATCCGTTTTACACAATCA
ATTGGGAAACAAGTTATTCTAATTAACCACAATAGTCGGAACCCATCCACAAGTCTCCAA
ACCCCAAGCCAAGAGGCAAAACCCTCTACTCTTGGCTTCTCCAAACTAACCCTCTTTAGA
GCAAGCTCTTTAAATAATATAAAAAAATATTCCTCTCAGCACTTCACTTAAAGCTGCTCA
GCCATACTGCTCAATAATATATTCATAAATTTCCCAGAAAACAAAAAACGTATTGTCTGA
AAAGGGAATGTTTCTTATATGGTTTAGCCAATACTTCAATCTCCCACCCACAAACTTGAA
TCTTAAATAAAAGTGAAATATACAAACAAATTAGCTGTTGGAAAAAATTCGAAACCTGCA
GAAGCATCACCTTACACACCCGTCGACATAACCTTTAAAAATCCCACCAAGTTCCTAAAA
AAAAAACAATATCCCATGATCCCAGAAGCTAAAAAAGAGCTAAAAACTATAATTTTTGAT
CTTTTAAAAAAAGGATTACTCAGCCCAGGCAACATACCCTGTAATACACCTATTTTAACT
GTAAAAAAACCAAACAAAGAAATAGTAAAAAAAGGAGATGTAATAATGGAAAGTGGATAG
ATCATACAGACTAGTTCAGGACCTTAGACTTATTAATAAAGCTGTAATACCAATTCATCC
TGTAGTTCCTAACCCTTATACATTACTTTCACATATTCCCTCAAGTACAAAATTATTTAC
TGTACTTAAGCTAAAGGATGCCTTTTTTTCTTAACCCTTAAACACCATGACCCAAAATAT
TTTTTTATTTAAAAATGGCAATAACAAAACAAACACGTTTACACACAGATCACATGGTCT
GTTCTACAAACAATTGAACAGGCACATAAAAGCTCTATAAGAAAGACATTTTAAGGTCAA
ACAAATAAAGATTGTATAATATTCTCTAAAAAATATTTAATAAGAATGCGAGCGAATATA
ATAGTATTTGAACCAATAGATGAAGTTAAACAATCAATTTAAAATGGAAAGCTCAGGGAT
TCAGAGATAGCCCCCACCTATTTGGACAGGCCCTTACAAAAGACCTCGCTGAAATACGCC
TTACACAAGACACACTTCTTCAATAAGTCAATGACTTGCCTCTTTCTATGCCAAAAGGTT
AAATAACCTTCCAAAACAGTAAGCATTCACTTAACTTAATGGAATGAACAAAAAAAACCT
AGAAAGCAACAAAAGAAAACCAGTCATCTGGCACCTCTAGCATCCTCACCAAGAGACACA
AAACCCCCTTGAACAAATCAGGTGTTGATATTTGGTCACATCCAAGCTAGAAAGGACCAA
AATACAGTATTAGTACTAAACAAACTTGCTGATTGTGGGTACAAAGTATCTCCTTCTAAG
GTACAAACATCCACACAAAGAGTTCAATTTTGGGGCCCCACTAAAAAACCACAAACAAAC
ACCCCCTCAAGCACCCATCTACAGAGGGCTCCTGAAGCATCCAAAAGAGAGAAAAGAAAG
AATCTTATATAAGTCATGCCGCCACCCAAAACTAAACAACAAAAAAGTGAATTTTTACCT
ATTTCAGGATATTTTTGAATATGTATTCCAAACATTGGTTTAATAGCTCAACCCTTATAT
CCTGCATTACAGGTGTCTCAGTAGGTACAATCAATCACCCATAATATTAGAAACAAGAAA
AAAAAAATGCATTACAAATAATAAAACATGCTTTCATAACAGCCCCAGCACTGGCACTAC
CAAAACTCATTAACCCAATTAATTAGTTTGTACAAAAACAGTGTGGAAAACCTTTTCAAG
CACTCAAAATCCTAGAATATGGTGTCTTCAAGGAGGTTCATGAAAGGATGAAAAGGACCC
CGAAAAGCACTCTTGAATACAGGTTTCTAATAACTTTAGAATCACATCATTTGGACTGGG
TAAGAATTCCCGGAACTCTAATGAAAAGACTGACTGGTTTATAAAACTGCTAACCCAAGC
AGAACAAAAATTAATTGAATACCAAGAAAATACTTTGCCAGATTATCATGCTAAATCAGC
CAATACTGAAATTGTTTAGATATACAATTTGAATGAACTCCATGGTCAAAGTCAAATGAC
CTATGATAACCCATCAGTTATCAGGCCAAACAAACGGAGCCGCCCAAAAAACTGTTGGCT
AACTTAACAAAAAACTTAAATTTGTCGTCCAAGGATGGCCAGCCTGTCTGAGAGCCTTGG
CAGCCGTGGCTCTTTTAATCCTAGAAACTCAAAAACTTTACATGCGAGAACACATAACTG
TTGAAATAACACGTAACATAGAAGAAATGATAAATATAATCAGGTTACTACTTTAATTCT
CTAAAAGACGTTTATACCAAACGGTAGTCTGCCTCTTCCAAATTCATGCCAAAACCATTG
AAAGTTCCGCAATCTTTCGGCCAACAAATAGGCCACTTAAACCTGCATACATGCACCACA
ATCAAGTGTCCTATTCTGAGCATGTTTTCTCGTTCATTAAAATCTTACCCACCCAAGAGA
ACCAATACCATAACACTACCTTCATTCACAAGAATGTAACTTGTCTTGTAAATGTAAGTT
ACTCCTGAGATAAAGAAGGAAGACATACAGCTGATTATGCAATTGTAACCCAAAAACAAG
ATATTGAAAGCACATATCTCCCATCAGACACCACGGCTCAAAAAGCTGAACTAATGGCAT
TAACTAGGGCCCTTAAATTAAAAAAAAGAAAAACAGTTAACATTTATACAGACTCTAAAT
ATGCATTCCAAGTACTCCATTCATATGCCATTATCTGGAAAGAAAGAGGTTTCCTAACTA
CAAAAGGAACTCCCATCAAAAATGGAAATCTCATATGCAAACTATTTGCAGCGCTCTAAC

TACCACAGAAAGTGGCCATTATCAATTGTAAAGGACACCATAACACGGCCAATCCAATAA
CCGAAGGAAAACAGTTAACAAATTAAGCAACAAAAAAAGCAGCACTAAATTCTGAGCAAA
GAGAGCCACTGAACTTCCCACCAACCCTAGAGACATGAGAATTAGAAGTAAAGCAACCAT
TTTTAAACACACAAATTAAATAACAACATGAGTAGTAAAACAACGTGGGGTGGAACCTTC
CATTAAAAGAATAAAAAATCTGACGCATCCATTCACACTCTCGCCCCTGTACACCCCTTA
ACAACCTTAAGTTAAAACTTACCAGACAAATTCCTACCCAGAAAAAGATCGCATCCTTGA
TGTGAACAGCTTATAAACAATTAAAAGAAAAAAAACTTGCCAGTAATAATAGAGACTTAT
TCTTAAAAATTAAGACCCCCAATGAACAACCCTGGGCATGATGTTTCTTTCTTTCCCCCC
ACCCCTGCCCTTTTCAAAGATATTAATTAGTCTTATTCTTTTTTCTACACAATAGTAACA
AAGGAGGCCATTCGACTGCCTCCCTGGATCACACACTACGGCGCAGAAGAATCAACAGGG
AAACTTAGGCAAATAAACCTTACTTTCGCCCTGGATGTAAGATGAAGGCCCAATCTGGGT
GAAAAACATTAAAGGTATCTAGTATCTTCCGTAATTAGTAAAAACCATCACTTGGGGGAT
CCAAGCACTATCGAACCCCTCCCTCAAAGGAGACACTAAAACAATGCCTTGAATCTTCTA
TATGTCTAATAATTCTTAGGCCTCCCTTTCCATGGTGTCAAGCAAATGACCAAATGATTA
GAAATAGATCCATCATAATAAGATCTATAGCAGATACAACCGTAAAGACTTTAGTTTTAC
CAAATGCCCTAGATCTGCCCGTACTTTTAACGGGAAGAAATACAAGTGCCACCGAATTCA
AACTGACTCAACAGAGCCGTTTCTGTGCAGTTGAAGGAAATTTATCCCGTACGTGAACAA
AAACATTTCAGAACAAGGAAACTGAGCTGGAAGAAATTAATAAAAAATTCCTTGAAGGGC
ATTAAGCACAAGACCGTTTAATAAAACGAGCAGACTCTCCTTATAGCTCCTTCATTGACC
CTTCTGACCCCAATCGTCTGGATAGGTTGGGAACCCGACCAAAGACCTCAATCCACACCT
CAGAAAACAGTCTCAAGAAACTCATAATCCCAAGCCTAAAAGTCCACTCTTTACCCTCGA
GAGACTCTACCAAGAACAATACGAGACGTCATAAAAAAAACCCTCAACCAAGCCTCCTGT
CCAAGCGAATAGATGGGGGCCCCCCCTTTTTACACCAACTAATCCAAACGAGAAAAACAA
ACCTTCCTTTCCACACTGGATACACCATTCCAAACTAAAAAGAGCACCAGATCCACATCC
AGAAATTTCCTCACCCCCAAATTATTCTCCCTCCCTCACAGGACCAACCTCACTGCACTT
AACAAGAATTCCAGAAGTTGCCAATCCAGAACGCCCTGGTCCATAACACTCTCTGCCTCC
AATTTCCAATCTTTTATCTCCTACTTTGTTTCAGATCTTTCCTGGTATCCCTTCCCCATG
TCCCTGGATAGTCCACGCCAATTTCTCACACTAATCCAGGAGATATGGCTGCAGGGCACC
TTCCAAAATTTCACTCCTACTCAAATCTCCTTTTTCTCCTAAGACCAAAACTGACTGTGC
AAAAAACTTGGTCCACATACCATTAATCGTGGAGATAAATGGCAGCCTTCAGACAAGAAA
ATTACCTCTTAAATCAGTCACACTCCCCTCTTTCTTCCAACTGTTAAATTTGTTTGTCCA
CACAAATCCAGCAGTTCACAGCCCTTCCTGTCAACCTCGCAACATAAACCAGATACAAAA
AAAAGCCGAAAAACAAAACAGACGATATACTCCTCCCTATGCCTACCTTAATATTTCTTC
TTTTGAAATTAGTTTTCTCTTAGTCACTGAGTTTTTGCATGCAGCCATCTCTAGAATGTC
AAATGGTCTCTCTTCAACTGGAATGACAAGAGCTCAAAGAAATGTATGCCCACAACGGCA
CCCTAACCTATGACTAACACCCTAATACCAGAAACCCTAAACTATGATTCCTAAGAGTCG
CACTCCAGCAATAACCTATGGACACCCTCTCAGCTATCAGAGAACCTGACTAAAAATGGG
GCAAACTCAAAAACGCAACTTTGAGACAACAAAACTTTGAACCAAGCTCGTGCACCACGC
ATCCACAGACCTCAACAAAAAAAAAAAGGACTAATACGTTCAAAAAAACACAAATATAAAA
AGATACCCAACGAAAACAGATTCATACTCCCAAAGACCACCATTCCTTTACTTTCTGTCA
GGGGGGCTAACCGGGCTAGCACACACCCAAATAGAAAAAAAATAAAAGAAGAAAAAAACA
GGCGGTTCAAGAGACAACCACCAACCAAACAGCCCCTGCCACCATAACCTACAACTTTTG
TCTGTCCACCCCCAGTGTCTTCTTCCTGTGTGGCACAAACTCCAGGTTTTGCCAGAAGGC
GGGCCACCCTACACCAACAAGTACAAAAAAACAATCACCAAAAGTCATCGTTCCCACCAT
GCAAAACCCAATGTTCGACTGTTTTCCAGCAAATCTAATACCCCCAATAACTATCGGAAA
TTACACGATAACATACAAGCTCAAACAAAAATCGCTTCAGACATCCGAAACGGACCATAA
CATCCCTTAAAAAATCAGTAGGGATGGGTACCAAAAAAACTTCTCGTCTTGTACATCGGA
TTTCTTCAATTAATTAACACACTGAAACTAAGAGACATTGTATTCATCCATACTTCACTT
AATGGAACTCCAGTTGCCTATACGCCCTTTTGAAGATAGGCACCAATAGACAGATTAAAA
GGTGAAAGCAGCCCGACCCTCAGTATGGTCCTGCTACAATGACAACCTGAATATCAGCCA
CTCTCACAAAATTCAGTGAAAGACCAAAAGGCGCGAATTCAAAAACAATAACCACAACAC
AACACCCAAACCAAAAATGAGCAGGCAAAATAACTTCCCTTGTCCCTGTTAAATTTACAA
AAAAATAGAAGGGAAGAA

>3_cons
TTTCTGGGGTCCCAGACCAGGAAGCGAGACAACAGATGGACTGTCTCCTTTGCCTGTGGG
GAGGCAGCCCCTGAGCCGGCGGAGACCTGCCATCCGAAGCATCCCTGCGGGGAACTCCAG

CCAGCAGGAGAGACCCGGACCCCCAGAACCCTCCCAGCCAGCCAATCACCCCAACGGAAC
GCCTCAAGGGGCTACAGGACGATACCAGCAACAGCGTGCTAAAGAACCGCGGTAAGGAGC
AAGGGCCCAAGGAAGGAGTGCCAGGCCCATATGGACGAAAAAGGAACTGGATCAACACCC
GGAAGGAACCGACAAATCCAACCCAGACATGCGAAACGTGGCAAGACTGGCCTGCGAAAC
TGGACCAACCCCATACCCCCCTAACAAGTGAAAGTGGTTCACTGGTGGAGAAAATGGGCC
GATAGAGCGGCAAGTCCAACAAAGAAGAGCTTGCTGGCAGGGTGGCAAGAGTGGCTTGCC
ACCCCAACTGGGAGTGGGGAAGTGTGTGTGGACCTACCCAGGACATGAGAGAGGCTCGTT
TCGTCCGATGAGGAGTCCTGGGGTAGGAGTGGTGTGTGTATGTGTGTGAATGTGGGAGCC
TAACTAGGCTCACCCGGGACACGGGAGAGGCCTATTACGTCCCAACAACAAACCGGCGAA
AGGCAATGAGTGTGAACGTGGCCGTATCACACAAGGCCAACGTGGACAGGAACGTTGGGT
GCTGTAGGTCACTTGGCTTGGATTGTTTGCTCCGAGGGGAGTGGGTGGAACATCGGTTCT
ACCAGCGAACATCCGCCAGGGCAATTACATATGGCTAACAGAAAAAACAAACAACTTAGC
GACCGAGGTGGCCTGCGTAAGGGGAAGGTACGTACCCAACAGCAAAATTGGGGGCTTTCC
CAAGGATCACATTTTCGTCCTGGGTCGGGTTTCTGGTAAAGTGTAAAAAGGGGAGTGCGC
GCCCCACACCCCTTCTTTGGGTTGCGGTTTGCATCAGGAGGGAGCAAGCTGGATCCACCC
AGAGGTTCATGAAAAACACGAGAATAACACCAACTCAATAAGAAACTGTGTTGAAAAATT
GTTAAAAGGGTTCTGAAGGAGACTGGCGAATAAAAATCACTGTACACAAACTAAGGACGG
AAAATCCAATAGAGTGGCCAACGTTTATCGTTGGTGGGCCGGCCGAAGGTAACATGGACG
GGGCTGTAATTCGCCGTATGTTTAAGATATGGTATGAGGTATGTTGAAAGCCAGGGCACC
CAGACCAGTTTCCGTACATAGACTCTTGGCTACAGCTGGTCCTAAACCCCCCAACCCCGA
AGTTGAAGATAATACAGTCCTCCAAACGCCAAGCAGAACCAAAAAAAGACCTAGGAGAAA
AGAGACAAAAGAAAAAGAAACAAAAGAAAAAAAGCGTGGAGAAACACCAAAAGACAGACA
GCAATGCCAGGCAAGAAATTGCAAGACGGAAGAAGCGAGCCAAACCGAGCCTCAGACAGA
AGAAGAAGCAACTACGAGTAGGGGGAGAAAAGGAGACAGTAACCCTAGAAAAGACAACAA
AAACCAGATTTGCAGAAACCAGCAAAGGAAAGACAAAGACTGGAAACTTATAAGTCCAAA
CGTAAAAGAAAAGAGAGAAAGGGATGAAAAAAGGGATCCTATGACACACCTCCAGAAGCA
AAAAAGATAGTCCCATAGGTTTAATGATTAACCTATGGAAGAATCATCAAAGCACTAATC
ACAAGAATAAAAAACAGATAAAACAATACTGTGCCCTTATCTGGCCACTAGGTCCCCTCC
ACAGCCCCTCAATCTTCTAGCCTAAGTGGGGGTCAAATGAGGATGTAAGGGCACAAAATG
TAACCCTATCCGATACTGAAAAAAGTAAAGGGGAAGAAAAAGAACAAGGTTACGCACTCT
GTTGGATCTAGGAATTAAACCGCCTTGAGCCTTCCACTTCAAGAGAAGAAGAAACCCAAA
TACCATGAAAAAGTAGGAAGCCAACTAGGAGGAATATCCCATTTAAAACCACCAGCGCTT
GTGACCCCCTGACCTTTTTTCCTGGAAATACAGCCTCCATTCCTACACCAAAAAGCCGCA
GGCTCTCATCAAGCAGCTGCAATCGATTATGATGACACAACACCTCACCTGCCCACACTG
CCAACACCTTCTGCTTATAATCCTTAAAACAGAGGAGAGTAGCAGTGCGACCCAAGCAGT
CCTTAACAGCCTATATACCATTACCCAAGAGGACACCCTCAACAACTACAACTATGTAAA
AGCGATATTCACAAATACCGATCCCCATTCGGACCAAAAGAATGAAGAAGAAGTGTATTC
CCTATAAGGGGAGTGAAAATAAAACTGGAAAAGGCTGAAATTAGTTTTATAAAAGGACTT
TTAAATGAATCAGAAGTCAGAAATGCTTCAGAAGGAAGTAGAAAGTCCCCGAAAAGTCCA
TTAGAGAGGAGAACATAAAAATGAACAATTATAAGGGATCTGAGGCATACACTCGGTATG
CCCCAATGGCCCTCGTGGACACTCAAAGTCAAGGTAAGGAAAGAAAGGAAGATGTGGAGG
GAGGCCAGGAAGAGTCAGGAAAGTCAAAACCCTCCCGGCCAAAAAGTTCCAAAAGACATA
CCCAACACAAATTCATTGCCAAATCGAGACATCAGGTGAGAGCTTCAAAAGTCAGAGGGC
TTTGCAAACATGAATGTCAGTAGGCAAATAATAATAGAAGGAAACGTAGAATCAAAACGC
AAAAAACAAAATCTAATGAAAGCATTTGGAAGACAACAAGAGAAAGATGAAGAGCCAACA
CGATGCCTACACAGACAAGAGGAGCAGATGAAGCAATAACTAGAGACCGATATGGAGGGC
CCAGTCGAGGAAGGGATCAAAAAAGTCAAATTCATGAAGAAAAGTTGAACTGGCATAAAA
ATAAAGGTGCAAAAAGTAGAGAAATAGGAAAATAGACCAATAAGTGAACAATTAAGGGAA
AATAACAAGATATTCATAGGTAGAAAGAAAGAAAGACAAAAACGAATGACAAAAGCAATG
TTTCCCACTGCCTAACAGATAGCGACAAAATCGGCTTAACATAGACAAAGGATCAAGAGA
ACCAGAAAAGATAAATTGCCAGACCTTAACCACTCTAGGGACCCAGGACCCCTATAAACA
AAGGGAGGCACAAAGTCCATGGATATTATGCCCAATACAGTGCCTGAAAATTCAGTAGTA
AATCTTCCAATAACCCAGCTATAAAAAAACCATAATACTTTTAACAGGGCTACAGGGCAA
TCAGAAAAAAAAAAAGATTGCCCTTCAAGGAGAAATCTGATAGCAGGCCTAGAACTCATG
AATGTTCAAGTAGACCAGGAAAATCATCGGGTCCCTCTGTTTGAAAGAGAGTCCCACCAG
AAACTGTTGACAAATATAGATTTTGCACCAGAACGTGATATAGTCACACTTTTGGTTCAC
ACAAGGGCTACTGTGTCCACTATTACTGTCCCAGAATCTAATCCTGACTCCTCTGCAGAA

AAACTGACAGTCTCAGAGGTAAAAGGGGAAGGATTTACAGTGAAAATATTAAAAAAAACA
GAAGTCAGATAAAGACCAGGATTACACAACTCATATTCAGCGGTCATTCTTGTTAATCCC
TGAAGCAGCAACTAATTTACTAGGAAGAGACTTAATGTTAAAGTTGGGATTAGGGCTACA
AGTACGATAAACACAATTCCTAACTGTAATGGGCCTACACATAATGAGCGGTCTGCATAA
CATGAATCCTGGTGTCTGGGCTGAAGAAGACAACCCTGGGATGGTGGAAAACCCCCCCAC
CCAGATAATAGAAATAAAACCACATGGAGAAGTAGTAAGAAGAAAACAATACCCCATTCC
CAGAGAAGGCATTCAAGGGATAAAACCAATTATCGAGGGCCTCAATAAAGAAGGGATGAT
AGTTCCCTGCATGTCCCCATATAACACTCCAATTCTGCCTGTGCAGAAGCCAGATGCCAA
TGAAGTGTAGGTCATACAGGCTGGTACAGGACTTAAGGGCGATTAACCAAACGGTAAAGA
CCATCCACCCTGTAGTACCTAACCAAGCCGCACACAAAAAAAAAGCAAAGAATAAAATGC
AGATGGTTTACACCCTTCTGCGAATGATTCCATCTAATGATAAATGGTTTACTGCAATAG
ATCTGAAAGATGCTTTTTTTTCCTTTCCCCTCAGGATTCTACAGCCTGGAGTGCTAAAGA
GAGCCAGGACCTATTTGCCTTTCACTGGGAAGATCCACACTCAGGGAGAAAACAACAATA
TACCTGGACTGTCCTGCCCCAAGGGTTCATGAACTCCCCCACCATCTTTACAGGGAGCAA
ATTTTGGAGCAAGTACTTGAATAACTTTACCCTTCCAAAAGAAATCTGACTGGTACATTA
CTTCGATGACATTCTTCTATTTGGAGACACTACGGAAGAGGTAGCAGACAAAGGAACACA
AATTCTTAACCATCTCCTGGACAAAGGAGGGGATAAGAGTCTCAAGAAAAAAGGCTCAGT
GTTTAGATCCCGAAGTTAAATATCTAGGCTTCATGATAAGTGAAGGCAAGCGAAGAATAG
GTCCTGAAAGAAAGGAGGCAATTTTGTCTCTGCCCACGCCTAAAACCAAGAAAGAAGTCA
GAAAATTTTTAGGGGTAGTTGGATTCTGCAGGATATGGATTCCCCAATTTGCACTAATAA
CTAAGCCCTTATTTGAGAGGGTAAGAGATATCAAGAAAATGGAGAGGAAATCTCATGCAT
GCAACCTATTATACGAGACTCAACATAAGAAAACTTGACCAAACTGAATAATGGAATTAA
AAACCTTAAAGCAACATCTCTATGATGCCCCAGGCCCCCCCTGGAGCTACCTTACTTAGA
AAAGCCATTTGTACTGTTTGTGTCAGAAGGAGATAGGATGGCTGTTGGAGTCCTTACTCA
AGCCCTCAGAGGCTGGCGGCAGCCCGTGGCCTTCGTCTATCAAAACTACTCGACCCGGTC
ACCTGAGGATGGCCCCAATGCCTTCAAGCCCTAGCAGCTACTGCCCTACTAGTACAAGAA
AGAAACTGAAAAACCCTCCTTCCTCTTGCCATCCTTGATAACCTTTGGCCAAAACCTAAA
AATAAGCACACCCCATGCTCTCATGAAACTATTGCTTTCTGACCAATAAAGCAAAAAGGA
CGGCGATGGCTAACAGACTCGAGAATTCTAAAGTATGAGGGGATTCTGTGATTGGGCCCG
AGCAGGCCCTGAAGGCACAAGTGAAAATCATGATATAACAGTTGAAACTTACAATACCCT
GAACCCAGCCACCTTGCTATCTCTAGAACAGCATCCATTAGTGGCCTCAGAACAAAAATG
TTTAGATTTAATTGAATACCAAACAAAAGTAAGACCAGATCTAAGAGAAACACCATTTCA
TGACGGGTGATATCCATTTTTAGTTTTCCTGCAGCACAGCCACAGGACCTCTTTATAGAT
GGGTCCTCCCAGGTGATAGATGGAAAATGATACACTGATGGTTCTGCATTCAATGGAGAA
AGACTCCCAAAACCAAAAAATAGAGTCATAAAAACCCTTCCTGGGTTGCCCCAGTATTAC
AGGGGTGAAGGCAGGTCTGCTCAAACGTGAGAACTGTTTGCACTGAACCTAGACCTATAC
CCCTTAAAAAATCAGGAAAGAACGCAGAACTAATTGCATTAACGAGAGCTCTGTTACTGA
CTAAAGGCAAAAATGTGAACATATATACTGATTCAAAGTATGCCTTTGCAGCGGTGCATA
CATATGGAGAAATATGGAATGAAAGAGGCCTAATAAATACTAAGGGAAAAAGACATTATT
TATAAAAAAGAAATCATGCAAGTATTAAATGAACTACGGAAACCAGAAAAGATAGCTATT
ATCCATGTAAGTGGACACCAAAAAAGGACCTCATTTGAAGGTGAAGGAAATAAACAAGCA
GATAAGGTGGCTAAACAAGCAGCTATGACTAAGCCTCAAACTCCAGTCATACCCCTAACT
CCCCATCTCCCTAAAGAACGACCTTAACCCCAGTTTCCTAATCAATCATAAAGGAAAAAG
AAGTATTTTCAAAAATAGGCGACAAAAAAAAATAAAAGAAGGATAATGGATGTTACCTGA
CGGGAGAGAAACGATATCTAAACCCCTAACGACAAAATTAGTATAGCAACTACATATGTG
CAAGGAACACACCGGGGAAGCAAGGCTCTAGGTAATGCAGTAGTTCGGCATTATGGATTT
ACACGGATTTATGCCCTCGCCAAAACGGTTACAGATAATTGCCTAATTTGTCAAAAGAAT
AATCCAACACATGGTCCGAAATGTACCCCCTGGAAGAAAAGATACTGGAACAAGACCATT
TGAAAATCTGCAAGTTGACTTCACAGAGATGCCTAAAATAGGAGGTCTACAATATTTACT
AGTGCTCCTATTTCACCTTCCCTGAAGAAGGGTAGAAGAAAGATAGGAAATTTCTTTCCC
CACCTGAAATGCAACTGCAAAAAAGATAATCAAAGCACTAATAAAAAACATTATACCCAG
ATTTGGACTACCACAAAGCATTGATTCAGACAATGGAACTCACTTTACTGCAAAAGTAGT
AAAGCAGTTGGCTCAAGTACTTAGAAATAAAATGGAAATTGCACTAATCTGGTCTTACCA
TACTCCCTACCATCCCCAGACATCAGGAAAGGTAGAAAGGATGAATCGGACTCTCAAAAA
ACAATTAAAAAAATTAATTCAAGAAACTTAATCATTGAAATGGGCCAAAGTTCTCCCAAT
GGCCCTCCTTAGAATTAGATCTACTCCAATAAAAGAAACTGGTTCTTCCACTTATGAGAT
ATTATTCGAAAGGCCACCCCCAATCATAAGTCAAATTAAAGGTGATCTACAAGAGTTAGG

AGAAATAACATTAAATTATGTTTACCCTTAGAATAAGAGGCAAATGCAAAGCTTTAGGAA
TAGCAATACAGAAAGTCCAAGGCTGGGTAAGAGAAAGAATACCTATAAGCCTAACAGACC
CAGTACACCCATTCAAACCAGGGGACTCTGTCTGGGTCAAGAAATGGAATCCAACCACTT
TGGGACCCATATGGGATGGGCCCCATACTGTAATCATGTCCACTCCCACTGCTGTTAAAG
TTGCAGGAATCACACCTTGGATCCACCACAGCCGGCTGAAACCAGCAGCACCAGAAACCC
CCGATGACGACAAGTGGACCAGCCAACAAGACCCAGACCACCCCACCCGAATAATCCTAC
GACGAAACCCAACCACCACTAAGAGACGACAACAGCCCTGCTCCGACCACACCGGAAGCT
GACCAGTCTACGCACGGCCGAAGCTTGAAGAGCCAACAAGCCCTGCTCTAGTCACACCCC
GGAAGCTGACTAGTCTATGCATTGCCAAAGCTAAACTCCACTTCAAATGTCCCTAAAATG
AAAATAAAAAATCTGATTCCTATAACCTTCCTGATAATACTAATTGTTCTACTATTACGC
TACCACTGCAAATGCTGCAAACCTCCACCCCCAGAGAAAGACCTCCCATGACCCTGACTG
GTATAAGCATGCTACTATTAACCAAAAGCTCACCACGGAAGAAACTACACATAAACTGAT
GATTTTACCATTACTGACTTCACGAAAAGAAGATGAAAATATACTTGGTTTATCACCTAT
TATATCCTCCATCCAGCCTGGACAGCCATTAACAAAAATCGGCTGCTACCTAAGTTTTAC
TATAAATGTACAGCAACCTTCCCATTAACCTGTATCTGCCTTAAGACCTTAGTTCCAGAA
GCACCAACCTCCCCATTCTGAGACCCAAAAAACACTAACCCAAATGTATGTTATGACCCT
AATCTCTTACCATCAGACACCAGGTTTGAAATAAAAATAAAAACTAGCCAACTCCGAGGG
GATACAAATGGAATAAAAGAAGGAAAACTTATAGATCAAACCAAAGAAGACCCTCCCTCC
AATAAGAGGCCTATATCCTTGTACTTTGATGATTGCCATGCAACATATAAACATAAGCAT
AACAAACCAAAAGCAACCAGAAAATGTTTAAATCAAAAAACAACTATCAATAGTGGCCCT
AAAAATATAAGTTAAAAAAAACAGATTGCATCCCTGAACCGTTCCATTAACTGGGAAAGG
AGCTACAGAATAAAAAATAAATATGTTTGTAAGAAGTTCTGAGGTTCTGAGCTGGGCCCC
CAGGTCTGAGCTCTGCTAAATCCGACCTGAAAATAACATTCAAAAGGGCCAGTTTTTCTA
ACAAAAGGCAAAGACAAAAACTACTGTGCGAACAGCAGAAAGAGGCCGGAATAACCAAAC
ACGAGTCAACCAACTCCTCCACAACGGAAAAAAGGAAAAGAAGCAATGTAATAAAGCAAG
TCCAGAGGCAAGCTGGAAGCATTTTTTTACTTTATTGTTTGAAACACTGCTGTTTTTTTT
GTTTTATTTTCTTGAGTCAAGAAAACTTTTTCTTTTGACCAGACTCCCAGGTGAAACCAC
TAATAAACGAAAACTCAAAAAAAACCTAAAATGAACCAATATTCCAAATCTCTTATTATTA
ACTAATTCTGCAAGCATCAGAAATTCTAAAAAAATCAGAAAACCACATTTACAACAAACA
TATTACTCGTTCTTCGTTTCCTTCTCCAACTAGCAGAAAATGGAGATTAATGTAACCAAG
TCAATTCATGCTATGTATATGGAGTTTCAAGCATGGGAGATCAATCCCCGTGGGAAGTGA
AGTTAATTTTGTAGAGCCAACAAAAAGTCCTTGGGAAAAACTGTCTGCACAATCCGGAGA
GTCAATGACAGCTCTCTCAGTCAGAAAAGTACCTTTAGTCCTACTAACCCTCCTCATGAA
GGACAATAGAAAACTACAACCATCTAGTCCTTAAAGAACTCCATAAAGGCAACATACTGC
ATAGCTAGAGCCTTGAATGACCCCACAAGGCCTATAGGAAGTATAACCTGAAGTCTCTGG
CCTAGGCCAATAGTCTGTTGACGACGCAATAGAAAGAATGGACTGTCAAAACAAGCTTTT
ATATAATTACAAGCTACAGAAACTAAAACAATAAATTCAATTAATCCAAGGGCCAACCCC
CCCAAAAATTAATCCTACTAGAGGGACACCCAGACACAAACGGTGAATCTTTCTGAATAA
AACCTATCCTTCTTTTTATCTTCTCTCAAACCCGCTTGGTACCACCCAGATTCTCAACCG
GACTGGAAGGCTCCCTCTGGTATATACGGGATATGGTATAGGGGCAGAGGTTACAGGGAA
AGACCCCATTGGATTCTTTGAAAACCGTTTCTCAAATACACTAATCAAAAAATCAACCTC
CAGCACCTGCCCCAAAAGATGTCTCCAACACCACTCACAACTAAAGAAAACCAAGCATTC
GCCCACCATCTATCAAACGACCCGACCAAGATAACGATCGTAGAGGTTAAAGACTTAAAA
CAAACTTTAGCAAGAAATTGAGACAGGATACCAAGATGTAAATGCCTGGCTAAAATGGAT
CAAATATTCCATCCGCACCTTGAACAAAAGGCATAGTTACGCTTGCGCACAAAGTAGACC
GGAGGCTGTGTACTTGGCACAATTAATCCCTTTTCCTTTATAATGCCCCTCAAAACAGGC
GAACACATGCGGTACCCTGTCTATGTTGAAAATAAAAAAAAAAGAAAAACCAACCAGGAT
AAGACAGCCTGGGGTGACAAGACATGCAAAGCTATAGCAATGCTGAATCACAAAGAACGG
CACCCTGAAGGTAAGATCCCATGGGATAGCCCACCTCCATCGCCACAAGAAAGATCAAGG
GCACATCGTACCCCTATCTACAGGCACAAACCTATCATACGGTTATAGGCGATCTGATTG
AAATAATAACCAATGAATAAAGGAAAGCAGTGAGTTTACTAACTCGGAAAGACACACAAA
TGAGTAAGCCCTTCTTCATCCCCGAGCGGATGTATGGTGGTATTGTGGGAAATGCTACCT
ATAAGAAAAAACTTGCCCTAGACCTCCTGCTGGACGTGTACTCAGGTCCATTTGGAAATT
AACTAACGTGTATTCATTCTCACTGGTATTTGGTCACAAATAAAAGAAAAAAGAAAGACA
GTAAAAAACAAGAGAAGCCACATATCAATCTTGGGACACGTACCTGTACATAGATGGCAT
TGGAGTCCCACGGGGACTACCTTTGGAATTTGGTGCCCAAAATTAGATAGATGCAGAATT
TTAATCAATGTTCTTCTGGTTGTAAAAGGAACCTGCTTTATCAAACTTTTTTTTTTGTTT

TGATAAACTACATCTATTACAACCAACAGCGATTTATTAACTACACTAGAGATGCTCTTA
AAGGAATAGCTGAACAATTAGGAGCTACTAGCCAGATGGCCTGGGAAAATAGGATAGCCT
TAGACATGATATTAGCAGAAGAAGGAGGACTCTTTCTCATTATAAGAACCAAATGCTCTA
CCTTCATAAACACACAACATCACGCCCAAACAATACCGCCCCTGATGGAAACATAACAAA
AGCATTACAAGGTCTGACTGCTCTATCCAATGAGCTAGCCAAAAATTCTGGAATAAATGA
CCCCTTTACAAAATGGCTAGAAAAGTGGTTCGGTAAATGGAAAGGAATAATAGCCTCAAT
CCTTACATCTCTCGCAGTCGTAATAGGAGTACTTACTCTTGTAGGATGCTGTGTCATACC
ATGTATCCGAGGATTAGTACAAAGGCTCATTGAAACAGCACTTACTAAAAAACCTCCCTT
AACTATCCTCCACCTTATCCAAATAAGCTGCTTACTTGCAACCTTTTAGCAGAATGAATC
AGCATAATATTATCATTTGCATAAAAATAAAAATGTTTTAAAGAGATACGAAAAGAGAAA
TAGTGAAAATGAGAATTACAACCAATGGTATAAATAGTAAAAAGGGGGGAAAAGGAAT

>4_cons
GATGGTACCACAAGGGCCCCGAAAAAAAAGAAGGCGCCAAGACGGGGATTAGGAGTCGAA
ACAAACATTTTCAACCTCGCAATGGGGACTTCATCACACGTGGTAACTGCAGCATAGAAA
GCATCCCACAAAACCTAGCACCCCAGTCAACCTTGGGTACCGTGTGAGAGACACCGAAGG
GACTCCCCCCGCGAGGCAAGCGGCCCCTCGCACCTCTCCCTACATGCAGTCTTTAATACC
GCGGGACTCTCCCTCTAGAGAAGAACAGCTCACTATCCTGTCCGTGCCTGCCATAGAGAT
GAAGGGCTCAAGATTGCTCAACTGCCCGCCAATTAGAATACCAGTGCCAAAATTAGGCTA
GAAAAAGAAAAACTTCCAAAGGACCTAAGCTATCGAACCTTCTTTTTTTCCTAAGCAACC
ATGTGTTAGCGCCCTCCCTTAATCCAAGGTTAGGTGTCTTGATCAACCTTCGGTTGTGCT
GTCAATTTAGGAGCTTTATAGTCGTTTCTATCCCTGGGGAAGGGCTCTTTAACTATCCCC
AACCTTTTCGCGTCTTAAGTAACGGTTTGCTACGAACAGGTTTTCTCTGCTCTTACTTTT
CTGTGGAGCGATGTCGCACACCTGTCAGGAAAAACATAACGTCTTTGAGTCACCCAGAAT
TTCCCAATTGAGCCAGAAGTTGAGCCCGATCTGTCCCTTGACATATCAGCTGACCTCTAG
CGAGCTCGCTACTATGGTTCCTCTGTCTTCTGCCTCCTGGGTCCAGATCATGACGAGCAC
GAAAAGGTAGCCACAAAAGCACCAGGACCACCGAAGCGACAACCTCGCATCACAGACTGC
CTGGAACTAATGACTTCCTCTCTCATCTCCTCTACAAGGTTATTCCTGCTAAGAAAAATC
AAGAAGCCCTACGCAGAAGCCTTAAACACCTAGGCTAGGAACCAAAAGATCCCTGTCCCT
GGTGCCCTTCCAGATTTAGGCATAAGACTCAATTCAAGGGCAAATTTGAGGGACCAGTTC
CCCACCATAGTGGACAGGCCCCCACATCTGTAAATGGCTAAGGGAAAAGAGAGACAGAGG
AGAGAGAGAGAGACGGAGGAGAGAGAGAGAGAGACAGAGAGGAGAGAGAGACAGGAGAGA
GACAAAACAAAGAGAGAGACATAGAGGAGAGAGAGAGAGTCAAAGAGAGAAAGAAAGAGA
AAGAAATAGTAAAGAAAAAACAGTGTGCCCTATTTTTTTAAAAACCACAGTAGCATTAGG
GCCTATCATCAATTATTCCCCAGAAAGACTTCCCCATAACACCAGGCCTCTCAAATACAA
TCTTGTTGTCAGTGTAAACAAGGGCGTGGAGCAAGGGTACAGAGACCACAGACAATCAAT
TGCTTTCCAATCAAAAATCCTTAACCCAGTAACCCGCGGATGGCCCAAATGCATTCAGTC
AGTAGCGGCAACTGCTTTGCTAACAGAAGAAAGTACAAAAATAACTATTAGAGGAAACCT
CATTGTGAGCACACCTCACCAGTTCAGAATTATTCTAAGTCAAAAAAGCAAAAAGGTAGC
TTACTAACTCAAAAATCTTAAAGTATGGGGCTATTCTGTTAGAAAAAGGTATAATAACTC
CAACCACAGAAAACTCCCTTAACCCAGCAGATTTCCTAACAGGGGATTTAAATCTTAATT
ACCATACAAAGGTCCGACCAGACCTAGGAGGTACCCCCTTCAGGACAGGGCGATAGATGG
TTACTCCTTCGTTATTGAGGAAAAACACCAAAATAGGGAAACAAAAAGCAATAAAGACAC
TCTCGCGAAAACAGAGGAAGAAAAATTGCCAAATAACAGCTATCAGCAAACACAGGAGAT
GTTCGCACACACCCTAACTTGAGAATACTCACAACACCAATAAACTATAAAAAACATGAC
GTAAATGATTCCAAATTCCCTCTTTCAAGGTCATTTCAATAAGAACACTAATTGATGGCC
AGGCCTACAGAAAAGGCAGCTGAGCTGGCATGAATCATTCGTACCCCCCACCCAGGACTA
GTAACCACCTCGGAACGCAAAGCATGTCCTCCATAAGGTGTAAAAGGACACCTAGAATCA
AAAAATTCTAACGCCCTGTCTATGAGGCAAAATATGCGAGACAACAGGGACCAAACACTC
TTGCAAATTCGTTTGCATGACTAAAGGCAGCTGATAAACAGATGATTGAAGGCAGTGAAG
AAAACTGTTAAAAAATACACCTCAGGCCCATAGGGGACGCTCTAAGGGAACTTCAGACCC
TAGACTAACTAAAAATGCGAGCATAACTATCTGTATACTTTCAGATGAGAACTATACCTC
TCTGAGAACAATCTCCATTAATATGTATCCTTAACAATTGGGACAAATTCAAACCTGAAA
TTTTAAAAAAAAAAGCGGCTGATATTCTTCTACAATACTGTCTGGCCCCAAAATTATCTTC
TTTTAAGAATGAAAAACATGGCCACCTGACGGAAGTATTAATTATAACACTATTTTACAA
ATAGACCTTTTTTGTAAAAAAGAAGGCAAATGGAGTAAACTCACATATGTCAATACAAAC
TTTCTTTGCATTAAATGACAATACTCAAATATGCAAAAGGTTTCCCTCTCGTCTTGTTTT

ATGTCCTTGGGAGCTTGACTTTGTGACCATGTGGGGGTACTCTCTCTTGGTCTCCACCAT
CCAGAAGAGGAAAAGAAGCAAAATAAGAAAAAGAACCAGACATAAAAGAACAGCCAGTCT
TGCAACTTATCCCCCACAGACAGACCACCCAGCTCTCCCCCTTGCCCTAGTTCCCTCCCA
AACGCTCCTCCCCCAACTAATGATGGTTCTTCTACAGCCCCCTTACCAAAAGGGCCACAA
GCAGATCGTCGATCGAATAAAATATAAGCCATACCCCTTTCCTAAAGACCCCCTTTTGGC
CTAGGTGATCCTAAAGAAAACCTGTCAATCATAGAGAGTGGAATAAAAAAGCCTTAAAAA
TTAAGGCAAAAGTAAAACTAAAAACACCATAAAAAGAAAAAGCACTTTAGAGAACTCATA
GGCCAAACTTTCATACTTATGAAATATTATAGGTGTAAAATTGGCATAAGTTAAAAATAT
CTAAAATTGTCGATTTTGTTTGCGTGGAGAATTCAAGTGGTTTCATGCTTACACCAGAAG
AAAAAAATAAGACCTCAAAAAGAGCAACAAAGAACATAAAACGACAAACACAGCCCAAAA
AACAATGGACAAAGAGCGTCTTATATGTGAAGGATAACAGGTACCTTAAAAACTGTTTAA
AGCACAATAAGAAAAGGCAAGATAGAACACATATCCTAAAACTGCCTCAACAAAAATGAT
AGCCTCTCCCCTTGCACATTGAAATTCAAAGAAAAAGCATAAATACTAAGACAAAATCCT
AAATAATAAAAAAAAGCCTTCTAAGTGTCTTGTAACTTGTCAGTGCCAGAATAATAAACTA
AATTACCAGAAATTTATACCTCCCTTTTATCAAAATTGAAGAGATATAATTGTTCAGTTT
TATCTTCAAACAAAAGACCAAAAAGAAACATGTATTCTATTTACAGAAACCTGAGAATTT
AAAGATGCCTGGTATCTTAGTAAAGTCCATTATATACCTCAAGCCGAATAAAGAGTAACC
TTAGAAACATTCAAAAAGTAAATCCCAAAAGTAGAACCCGACTGCCATATTGAATGAGAT
AATGGTGATTGGAGCTGAAAAAAATTAACCGAAAAAAGGTTAGTACTGTAAAAATAAGGT
CACTCTCATTGTAGACGGGTTTATGAAAAATTGAAAAAATCCTAAGAACTATTCAAAAAC
AAACACTAATACTAAGGGAAAAGAAGAATATCCATTTGCTTTTTTCGAAAAAACAAAGGA
GCCGATGAGAAAAAATAGCAATCTGAAAACTGAATCCATTGAATACTAAAAGATATAAAA
AGATTAATCCAAAAGCCAATCAGCTGCATATATTATTAGTAAACTTCAAAAAAATAATAT
GAGCCCAAAAAAAATTTTTTAAAAATAAAAGATTATTATTTCTACATTTAATGCCGTCTA
GATTCTTGCCACCCTCAATGCCCACAAGAGAGGACCTAAGGTAATTTCTGACAGCCTGGG
ACTCCTTGGGAAAAACAGTGGAGGTGCCACAGACACTATTGAACCTGGCAACCTCGGTGT
TCTATAATAGAGACCAAGAGGAACAGGCCAAAAGGAAAAAGCGAGACTAAGAAAAAGGCC
GCAGCCTTAGTCATGGCCCTCAGACAAGCAGACTTTGGAGGCTCAGAGGGAACCAAAAGT
GGAGCAGGACAATTGCATGGTAGGGCTTGCTACCAGTGAGGTTTGCAAGGACACTTTAAA
AAAGATTATCCAAGTAGAAACAAACTGCCCCCTTACCCGTGTCCAATATGTAAAGGAAAT
CACTGAAAGGAGCACTGCCCCAAGGAACAAAAGTCCTCTAGGACAGAAGCAAACAACTAG
TTGTAATAACAGAAGAACTTTGGATATCTGGAGCAAAAGATAAAACAAGACATTACTATC
ACAAAGACACTAGGATTCACTACCATTGAAGACCAGGAAAATGACTTCCTCAGGGATACC
AGTACGGCCTTATCAGTCTTACTCTAATATCATTGACCTCTGTCTTCTAAATCAATTACT
ATCCTAGGGATATAAAGTAAACCACTTAAATGGTATTTCTCCAACCTCTTAATTTGTAAT
TAGGTAACTTTGCTTTTATAAAATCTTTTTCTTTTTATGTCTAAAAGTACCAACCTCTTA
TTAGGGAAAGTTATAATAACATATCTTATAAAAACAGTAGCTACTATCAAAATGAACTTG
TAGAAAAAATTACTAACTTAAAGTCTCTAACTTGAGAAAGGATATATACTTGTGATTTTA
ATTTATATTTAATAAAAGCTCTTATGGAAAACAAGGCTAAAAATATCAGTGAAGAAAAAT
TCTAAAAGGCAAATAATGATCATCAAATTCTATTTAAAATTAAAGAATATATCACAATTC
CTTAAAAAAAACAATAACTCCTTCTACCTAAAACTTTAAAAAGATTTAAAAAGATAATAA
AAAATATAATAATACAAAGTATATTAAAAAAATTTAATTACATAAATCTTACAAAATTAC
AATCTTTTGAATACAAAAATCTAACTGTAAATCTATATGTCTTAAAGTATAACATTATAA
AAAACAGTATAAACTAAAAAAGTTACAAAAAAAGCAGTACATTTATAACAAAAAAAAACA
AGATTTTAAGATAAATATCTTAAAATAAAAAATAACAAAAATAAATTCTTGGACATAAAA
TTCATATAATATTATAAAAATGATTATATAAAAAATCTAAATTATGGAATACTTAAACCA
TATACTCCGTTTTCACATATACCAAAGAAAGTAGTATGGTTTATTATTAAAGATAATAAA
GATTAAATTTTTTATTATATACAAATATAATTACAATTAATAATTGTTTTATTTGAAACA
TTGATTTGAAGATCATAAAAAACTGGCCTCTCATATAACTTGGAAAGTCTTACCACAAGA
GTTGATAGATAGCCCAAAACTAAATAGCCAAACTTTAGTCCGGGATTTAAGCAAGTTCTC
AAACATATTACCTAACCCTCATATCATTCAATATGTGGATAACATACTTTTGGCTAACAG
TTAAGAATAGACCTTGTACTTTCAACTACACATAGTAAACAAGTTACTCAAGAACCCTCT
ATAAGAATAGAATCAACATTTGCTTAAAATTTCTAGATAATCAAGGGTACAAAATTTTAC
TTAAAAAGACTCAGATTTTCTTTTAATAAGTAAATTATAAAGAATTAGGCTAATCGAAAG
ACTTAAGGGCCCTAAGCGAAGAAAGAATCAATCTTAAACTTGTTTATCCATGTTTCAAAA
ACATGAAAATATTTGAAGGGGATCTTTAAAATAACTCAATTGTGTAAACAATTATTCACA
GGATACAGAGATATAATCCAGTCATTCTAGACCCTGAAAAAGGAAAAAAAAATAGCAAAT

GCGTATTTAATAGAATGGAAACCTTCATAAACCACATTCAGATAAAAACAGTCCCACTAA
ACACAGAAACAGCCTATAAAAAGCTTATATAAATCTTACCCTAAGATCCAACATAACAAA
ATAATTAAATTAGTTTAGTTTTAAGAGTTTATTAAAACAAGCTTGAGCGATATAAGACTA
CCCACAGGACAAAACTTACCTTTAGAATACAGCCATCCACATGAGGTAAAACCAGCAAAG
TACAATAATACACCTATTTTCAAAGAAATTACAGGTATAACCTTACATGAAGTTAAAATA
AAAGTTATATAACGTCTTAATATCCAGTAAAATAAATTAGTAGAAAATTTAAAGTAATAG
AAATGGGTTGGCTTCAATGTTTAAAAATTGTTTCAACAGTTTCAATCTTAATAATTGATA
CATTTAAAATAATTCATGAAAAATATCTTACTTTGAGGATATTTTACGATGTTAAAGGTA
TATTAGATGCTAAAAATAATTTGTGGCACAAAGATAACCAATTACTTAAATATCAGTCTC
TACTCCTTGAAGGACCAATATTTCAAATATGAAATTGTAAAGCTCTTAACCATGCCACTT
TTCTCCCAGAGGATAAGGAACCAATTAACCATGGCTGCCCAAAAATTTTTTCTCAGACAT
AAACCATTCGTGTGGATTTTTTGGACGTCCTTATAGATAATCTTAACATCAACTAGAAAA
CTGTAAAAAGTGCTTTTGAATAAAATTTAATGCAAAAATTAGAGTATGTTATAGTTAGTG
ATACGAAAATAATTGAAAATTAGTAATGAACTGAAATTATCAAACATAACAAAAGGAACA
GAAATATAGATAATAAAATGAACCAAAAAGAAAGCAGAATAATAATTAATTAATTTGTAC
AAAAATAGTTTAAATCGGTGAATTTTAATTATATATTTAAATAAAAGTTTACATTTCTAT
TTCTTTATGCTATAAGCATCAATATTTCTTATATGTTATAGAGTCAAAATAACTATCTAA
TAATTAAAATCTGACAGAAATTAAAGTATTATGCAAAGAAATTAAAATAGAGCATTCCTG
ATCCCCTCTAGGGGAACACCCATTAAACACCACAAAGAAATTATGAAATTATTGCAGGCA
CTACAAAAACCTAAAAAGGTGGCAGTCTTACACTACCAGCGTCATCAAAAAGGTGAAGGA
TAAAAAGTAGAAGGAAACCATCAAACAGAAAGCAAAACCAAAATTGCTGAAAGGGAGAAA
CTTCCTTTAAAAATAAATATCAAAGGACCCCTAGAATGGGACAAACCCCGCAAGGAAAAA
AATCCACAATATTCACCAATAGAAATAAAATTAGGGAGACTTTCACAATAACATAATTAC
AAAAATAAATAAACCCTCAACAAGGTTAACAAAAGAAGAAGGAAAAATATTTATTCCTGC
AGCTTTCCTCCAGAATATAGAAACTATTCGTGAGCATTCTTATCTTATGGCAATGTAATT
ATTTGCATAAATCAAATAAGAATATGTTTATTTTTGTAACAGGACATAATTGTAAAAACC
GGTTAGAATGACCAAGGCTTTCCCTGCAAGACATTGGCAAGTGTTAGAGGAAAAGAGGAA
AGTTTATCCGAAAGTGTTATTGAAATCCCCCGGAGAAGAGACAATGGAAAGGTCTGTTGG
ATGTTCCAACTTTAAATTAAAAATCAAAATTATCATTAGAAGGCAATAAAAAAAAAATAT
GAAAGCCATAAAATAAGGTATTACCCGGGAAAGGACAGGGAATCAGACCTTACGCAGATT
ACCAAAGTCACGATAATACAGGACAGGCACATAAAAAAATTCAGTATTTATTAGTCTGTG
TTGATACCTTTACTGGTTGGATAGAAGCCTTCCCCTGCAAGACAGAGAAGGCACAAGAAG
TAATTAAAGCACTAATTCATGAAATAATTCCTAGATTTGGGCTTCCCAAAAGCTTACAGA
GTGACAATGGTCCAGCTTTTAAAGCCACAATAACCCAAGGAATATCCACGGCGCTAGGAA
TACAATATCACCTTCACTGCGCCTGGAGGCCACAATCCTCAAGGAAAGTCGAAAAGGCAA
ATGAAACACTCAAAAGGCATTTAAGAAAACTAACACAGGAAGCCCATCCCCCTTCTCTTT
CTCTAAACCTCACGCAGAGAGCAATTTAACCAACTCATAAGTTTTTTAGGGGAAAAAGAC
ATGCCTTTCTTTAGCAGGCACAGGATCCAAAAGATGCCCATAGACTTAGTCAACTAAGAA
TCCTACAAGTCTTTAAAAAGCTCTTATAAAAAATTCATTATAAGAAAAAAAATTCCATTA
AAAACAAAATAAAAGGCATATTGGAAAAATAAGTATTCTTGCTAAACTTTTTGCAAAAAA
TAGGTTTAATAGTAATAAATATAAGAATTGTATCTCAATCCAAAATTCTAAAGTTTATGT
TCCAGCAAAGCAAACCTTAAAAGAGCCTATGTGGTCAGTCACTATTCTTGCTGCATTTAT
GTAAATAATCAGGCCAAGTCTAATGAGATCAGACTTATTTTGCAAGCAACCATTTTTAAA
CTATTGACAACATGCTAAAACAGGAAAAAACAAATGTGGACACAAAAAATATAGCACACC
TGTTGTTAAATACTAGTATTGCCTAAAATTTTTCAAACTTCAGCACTGGCTTTGGGCAAA
ATATCAAAAAAAAAGCAGGCAAAATCTTAAGACGAAAATAAGCCTCAAAAGAAGACCAGA
GTACAATCATTACTAATAAATGTTATTGACGTCTTAATATCAAACAACACTATAAAAAGC
ATGTCACAGAGATAAAGGTATAATCCTTTTTCAACTAAAAGTCTTCAAATTAATCAATTC
AATTCTCTCTAACCCTTCATCCCTGGAATCCCCACAAAATTAGGAGTCCCCCTCCTCAGA
TTTGATGCAGACAAGGAATAATCCAATCTACCCCAAAAACCCTCTACGTAGAGATAAAGA
CTCAAACCATGCAGTTAAAATCGAAGGAGTGGAATCATGAAAAAACCACACCTGACTAAA
ACATTAGACACTGCCTGAGATAACTAATAAAACATCAACCCCGTAAAGACAAGGTCAGCG
AAAGAAAAATCAATACACATGGAAACCAAAGGAGGACCTGCAGCCTCCTTTGAGAAGGCA
ACCTACATGAAGGAAAAAATATAGTCTGTAAAGGTTTTCTATTCTTTCTCGGTTAGCTGT
CTTCTTCCTAAAAGAAAAGAAATAAACTTGCTTTTCTTTTACCTAATTCTGAGAATAATT
AACGGCAACCTTAAAAAAGAAAAAGGAAAAAACAAAAAAAAAACAAAAACCACATTAATC
AGTTACAGCTCTCATTGCTCTCTTTAATGGAAGACTTCTACTATTTCATCACATTATTAA

GCAGCATACTAACCACACTCCTTATAATAGGACTATATACTGTAGCTCCTGCCAGGATGA
AAATCCTAATCACATCAACCTTCTTTCTATCAGCCTTCCTTGTAACAGCAATTCACTCCT
ACCTTTAACTCAGCCTGGAAAAAATGATGTCATCTTCCAGAGCACCCCCCTTATCTTTCT
ATTTACTATTTCCCTAACAACACCTCATGATTCCTTTCACACTTCCTGCAATCACTCCTC
CCCATACTCCAGCCCCTAACTCCCACTACAAAACACTCCACTTGACAATGTGCCTCCCCG
GAAAAACAGAACAACCATTCTATAGGAAATTATTGAAGGCTGGGTAGCCCCCTACAAACC
CCCACATATGTTGCCACTCACACTCACATGAAAAACAAATGCTATAACACTACAACTCTC
TGCACTCATAATAAACAAAAACACCTTATAAGACAAAAAAAAAAAATAAAAACTAGCTGAC
CTAAAAAACGGGGAACCAACACTTGTTGGACATACTATACCCATACAGGTATGTCTAACA
AAGGAGAAGTCCAACTTTTACGTTAATAAGACTAAAAAATAACATATAAAACAAGTAATC
AAAAACCTAAACCAAGTACACAAAACACCAAGACCATATAAAAAATTAGACCTCTCAAAA
CTACAAAAAACCCTCAGTTATCATACTCGCCCCTGGAGCCTATTTAACACCACCCTTACA
ATTAAACCCTACAACTTCAAGCCCCAACTGATCATAGTAACTTCTGAGTCACCCAAACAG
CTCCATTCAAATGGTTTGTCTGCTCAGGGCCCCCAGAATACATGAGGTCTCCCCTAATAA
ACCAACAAACTGTTGGATGTGACTCCCCATGCATTTCCAGCCATACATTTCAATCCCTGT
CCCCAAACAGTGGAACAACCACTACTAGTGAATGCCCTCTAAATTTCTCTTTCAGTCACT
CTCTCAAAAAGTACAAAACACATCCAAATTAGTAGGTCCCATAGTTACCAATATAGAAAA
CACAGAGACCTCAAAACTCACATGCATAAACTTAAGCAAGACTATATACAAAAACATCTC
CCAATAAATTTCATAGGAAACAACACAGTCACGAATCGCCTGGCTAACATAAAATAAAAA
TAAAGAAAAATCAACCAAGGAATATTATTCTTCTGTTATAACACAACCTATCGATACCAG
AATAGTATCTTAATAAACAAATGCACTTTCGAAGCTCAAAATGCTTACATTTAAATATAT
AAAGATCACAAGCCTCACTTATATTCTCTTGTCTCTGACACCGAAAATTGATCTGAATAA
CTGACTATGAGAAACACGAATTCTTTACAAAATCAGACTACCAAATATCCACCTAGAAAC
TTGCACAAGATCAAGGCCACCATTGAACCCTTGAAGAGCCCCCATTCTAAAATTTATTGT
TTGAGCGCCAATGATAGATACAATCCAAGGGAATAGAACTGCAGGAATAACAACCTCAAC
ATAGACTAAACACAAACTATCAAAACAATTAAAGGAAACAATAAAAAAAATAACCAAACA
ATAACAGACACTACAAACACACCTTAACTCTGTGGAACATGTACCCTTCCGAAACTGCAG
AGCCATGAACCTTTCCTTGGTGGAACAAAGAAGAACCTTTTCCTTGTCCGATATAATATG
GGGAGATTTGATTTATTAAACTACCAAAACAATAAACATAACACATAAATTTAAAGAAAT
TCAAGAACGCACACTAAGTAGACAAAAGGAGCCTCATAACACAGAATCCTTCAACCTCCT
TAGCAGATGGATACCATCGCTTCTCCCCTTTTTAGGTCCTGTAACAGCAATCATATTGCT
ACTCGCCTTTGGGCCATGTATCTTTAACCTCCTTGTCAAATTTGTTTCCTCCAGAATCGA
GGCCATCAAGCAACAAAGGTTCTCACAAATGGAACCACAAATAAGCTCAACAACCAACTT
CAGAACACCAGACCAAAGAATACCAAGGACCCGTGGACCCACCCCCTAGCCCCTCCACCG
AACTGAATATCCAAAAACGATACCCTCTCGAGGACATAACAAAGGCAGGGCCACTTCTTC
GCCCCTATTCAGCAGGAAGTAGTTAGAAAAAAAATCCGCCAACCTCCCCAACAGCATTTG
GCATTTCCTGTTTAGAAGGGGGACATGAGGAAGAATGGGGATA

>5_cons
AGTGGCGTCCACGTGGGGGCTCTCTGGCGCCCAAAAGGGGTTTCTCACTCACTGAGGGAA
AGTGTGGGAGTCCATTTAGCCTGTCTACCTCATCCTCTCACTGCACTCCGAGTGATGACT
AAGGACGGCAGCTCTTAAGCATCGGGGGGAAGAGTGGCCACATAGGGTATTTCCGAAACTC
CCCTCGTTAAACTTTTACACACCGGTTGTGTTGTAGAAGTATTGTATAAAAAACCTACTG
AAAACACGAGAAATAACAGTTATTTATAACACTACTATATAAAACTAAAGGAAAGACAGG
GTCATATCCACGCGGCTTCCAACAATCTGAACTCTTTATTTAAAAAGATGGAAAAATACT
GTCCTTAGTTTCCTGAAAAAGGAACCATGGAGCTAAAAGTATGGGACCGAGTTGGTGCAA
CATTCCGGCAACGGGTCACAGCAGGTAATTATCTTCCCATCACTATTTGGAGTGAATGGG
CCCTAATACGTGTTGCCTTACTTCCATACCAGTCCAGTGAACCCCTACAACTACCACAAC
TTAACGCACATGGCGACCCGCAGCCTTTACCTCAGATATCCACCCCCACTCCGACTTCAC
TTTCTGATCACCAAATACAATACAATCTACCTCTCCTACCTCACCAAAAGGAGGAATCTA
TGAATAACTCCCGAAACATCCCCTTAACCTCACCACCTGAATATCTTAAATCTTTTCAAA
CAGAGCTGCTACTCCCAGAACCAGCGGAACAGACTCAGCCATCCTGTGAACATCTAAATC
CTCATTCTTCTCACCCCAATCATCAGCACCCCTACTCTAAGCCTACTCCTACTAGCAACG
AGACCAAACAACATATTTCATAAACTTATACTGCCCCTCCCCCAAAGACTACAGCCCCAC
ACCCTCCTAACCTTTCGCTCATTCACCCGGCCACTGTTCAACCCATCCAAACTACTAATC
AGCACGCAACTTAAAACATGAAGACAACTAATCACCAGGAAGTTTAGGCCCCTCCAACAC
CCACAACCCCCAAGTCTCAAACTCCAATACCGGTCCGACCTCCTCAACCTCAGTTTCCCT

TATCTACACATACTTTTCCTGTCACTTCTATGCCGACTCCGTCTCATGTGCCTGCTCTTG
AAACTTCCATGCAATGCTTATTACGCGTAGGTAGGAAACTGTGCGTCAGCCTGGCTTGAA
GTCTCTCAACTTCTATATTCAAAACCCCAATTTCTTTTCTGCTTCTGGTCCAGTCACAAC
TGCTGTTGCTACCCATAAGCAATAGGTTACATACATTCCTGATAATGACACCCCTCTTAT
GAGGGCCATTCTCAGGGCAAGGGAATACGGGGATCCCGAGGCATGGTGTCCTGTTATTCT
ACAATCTCCTATACCTGCTGCCCCCATTCTAGCTGCCCCTGCTCTGGCTGCAATGGATCA
GCCACCACCTGCTGACCAAGTTCAGCAGGCAGCTGACGCCACTGCCTCTCCAGACCCGCA
GCTCAGGGGATCAGGCTCCTCAGCCAGTGCAAGAAGGGCCTGATGTCCCAGCAGAGCCAG
TTCCTGAAATTTGATTTAGAGGCGTGGCAGTACCCCGTCACACTACACCCCCCAGATAAA
CAAGCAAAAGACATGCGACAATATGAACCTTTCCCTTTAAAATTCCTAAAAGAATTTAAA
GATGCTTGAAATCAGTATGGACCAAATTCTCCTTATGTCAAAACAGTACTAAAAACCTTT
GCTACTGAAAAACGATTGGTTCCTATTGACTGGGACATTCTAGCAAAAGCTGTTCTAACT
CCATCTCAATACTTACAATTTAAGACATGGTGGGCAGATGAGGCCCAGATTCAAGCTCCG
CTAAATCAGGAAAATGAAACTCAAATTAATGTGACTACTGACCAGCTTCTGGGAGGGGGC
GATTGGGCGGCTATAAGTAACCAACAAATAGCCTAGGATAAACCCACTTTAGATCAGGTT
ACCAGAACAAGTTAGTTAGGAGGAAAGGAAAAAATCCCCTTTGAAGGTCTTGCCTTTTTG
CAAATAACAGCTATTAAACAGGGTCAAAATGAACCATACCCCTGATTTCATGGCTCAAAT
ACAAGATGCTGCTGAAAAATCTATTCCTGATACGAATGCACAAGATATAGTCCTGCAAAT
GTTAGCTTTTGAAAATGCTAATCCAGAGTGTCAGGCTGCTATACAATCTGTCCAACGTAA
AACCCAACCAGAAAATGATTTGACCACTACCTATATCAAAAATAGAGCAGGTGTTGGTAG
AACATCACAAACTCATAAAGTTAAGTAGAGTCTTCTAAAGCTATCTGTTTCTCTATTTCC
TTTTCTGCCTGCTTTGAATCTGCTGTTATTAAGCTACCGGTGTTGAGATAAAACTCACTG
TTTATGGTACCGCTAGCCTCAGCAAAGAATATAGCCCACTTGCACAAAAGCATTCTTTGA
GCACAGGCAATGAAAGAACCCAAACAAAATAAAGCAAATAATTCCTTTTCCAGATCAACC
TGTTACATCAGGAAACAAGGTCATACTCGACAAGATCAAAAAACTGTAGCCTAAAAGACC
GAAAGAAACAAAGTCCTTTACTTAATGCTAATCCCCAACAACCTGCTCCTCAGAGACGGA
CAAAAACGAATACCTGTGTGGAATGTATGCCCAAGATGGAAAAAAGGAAAACATTGGACA
AATCATTGTCACTCTAAATTCGATATAAATGGTAACCCGTTACCGCAAATTCAGATAAAC
GGAAAGGGCAGGAAGCCCCAGCCCACACAACCAAACAGGAGGAGGCACCAGCCTCAAGCC
CCAATCCCAATCAGGGCTCCGGGGTTTCGTCCACAACTCCCAGCACCTCCCACTAAAACC
AAAACAGCATTCCCGCATCAACCAGTCCAAACAAAGCCACAGACACAACCTCAAATACTC
AAACCACAACCATATGCGTCTCAGCCCCTTCTCTTATCCCAGTACAATGCCCGTCCACCG
CCACAACAGGAGGTGCCGCAGTAGATCTATGCAGTACTATACCTATGACCCTACTACCTG
GGGAACCCCCTAAAATTGTCCCCACAGGAGCCAATGGCCCTTTACCTGGAACTTTAACTA
GATAAATTTTGGCCAACCCCTGCTTAGCAACAAAAGGTGTTAAAGTTCATACCGGACTCA
TTGATTCTGATTACTCTGGGGAAATAAAAATTGTTATTTCTACTAAAGTTCCCTTTAAAA
CTGAAGCAGGAGAATGAATTGCTCAACTTCTGCTTCTCCCGTAACTGAAAATCGGTACAA
ATAAAGGTAAACAAACAAGAGGCCTTGGGAGTACCAATAAACAAGGAAAAGCCGCTTATT
GGGTTAATAAAATTTCTGATAAACGGTCCGTGACCTGAAGGTAGAGACACTATAAAGGGA
AAGAACCTCCATGATTTTCTAGACAGAGGAACTGATTTTTCTATAATTTCTCCTCAGCAA
TGGCCTTCCACCTGGCCAAAACAACCCGCAAAAATCAAATTAGTGGGAGTTGGAAAAGCC
CCGGAAGTTTATCAAAGCTCTTTTATTTTGCATTGTACAGGCCCAGATGACCAAATGGGA
ACAATTCAACCATATATAACTCTTTCCCCATGTAATCCAGGGAACAGTGCTGCACTACAA
CAATGGGGGGCGGACATGACCAAACACAAGAAAATCACAGGCTATGAAAATGGTTAACAG
GAAAATAACTTATAATCCTGACCAGCTTTGTCTAAGGCCTTGGTTCTCAAAATCCACATA
AATTAAATACTAAGCAAAAACAAAATAAGTTAAAAAAAATGAGATATATGCCTGGAAGAG
GACTAGGAGAAAATTGGCAAGGGATAAAAGAACCCCTGCAACTCACCAAAAAACTTGACA
ACTAAGGATTTGGATATCCTTTTTAGTGGCGGCCATTGTCAAGCCTCCAGACCCTATCCC
TTTAAAATGGATATCTGATAAGCCAGTTTGGATAGAGCAGTGGCCGCTTCCTAAAAAAAA
ACTGGAGGCTTTAAATAAATTAGTTAATGAACAATTAGAAGATGGACACATTGAGCCATC
TTTCTCTCCATGGAATTCACCTGTGTTTGTAATACAAAAAAAATCAGCGGAAAATGGAGA
ATGGTAACTGACTTAAGAGCCATTAATGCAGTAATTAAACCTGGACGGTCACCACCCAGG
GGCACGTACAACCCGGCATGCCCTCCCCCGCTATGATCCCTAAAAATTGGCCTCTAATAC
TCATAGATCTTAAAGATTGCTTTTTTAATATTCCTTTAGACAAGCAAGACTGTGAAAAAT
TTGCTTTTACTGTACCTTCAATCAACAATCTGGAGCCTGCAACTCGTTATCAATGGAAAG
TACTACCACAAGGAATGCTAAACAGTCCTACAATTTGCCAGCCTTATGTTGGGCAAGTGC
TTCAACCTGTCCGACATAAATTTCCACAGGGTTACATTCTTCATTATATGGATGATATAC

TTTGTGCTGCCCCCACTGAAGAAGAATTAATTCACTGTTTTGCCTTCTTGAAACAAGCCA
TTTCAGAGGCTGGATTAAACATAGCTCCAGATAAAATTCAAAATACCACTCCTTTTCAAT
ATTTGGGAATGCAGGTAGAAGACAAACCCATTAAGCCACAAAAAGTCCAACTTAGTAGAG
ATAATTTAAAAACCTTAAATGACTTTCAAAAATTACTAGGTGACATTAATTAGATAAGAC
CTACTTTAGGCATCCCTACATATGCGATGTCTAACCTGTTTGCCACACTATGTGGAGATC
CAAATCTAAACAGTCAAAGGCCTCTAACAGAAACCGCAGACTAAAAGAGGCTAAACCAGA
GTTGCAATTGATGGAAAAAAGAGTCCAAAAGGCTCAAGTAACTAGAATAGATCCAAATTA
GCCTTTACATTTTCTAATTTTTCCAACTCAGCACTCTCCTACGGGACTAATAGTTCAACA
GCATGATCTAGTTGAATGGGGTTTTCTTCCTCATTCCACTTCAAAAACTCTAACTATTTA
TCTGGACCAAATCGCCACCATAATTGGGCAAGCAAGATCTCATATTATTAAAATTTACGG
ATATGATCCTAAAAAAATTATAGTCCCTTTAAAACAACAACAAATACAACAAGCCTTTAC
AAATTCTCTTACTTGGCAAATAAATTTGGCTGACTTTATTGGCATTATTGATAATCATTT
GCCTAAAAAAAAATTGTTTCAATTTCTAAAAATAACTTCTTGGATTCTACCTAAAATAAC
CAAAGATAAACCAATTACAGGAGCCGTTACAATGTTCACTGATGGGTCCAGTAATGGAAA
AGCGGTCTACGTCCCACCAAAACACCAAGCAATCCACACAACATCTGCCTCCTTTGAAAT
ATCATATATAATAAACATAGAAGAAAGAAGGGGGGGCTTACTGCCGTTTCTGAGCCATTC
AAGGAGATTAATATACCCCTAAAAATTGTCTCTGATTCTGCATATGTAGTACATGCCACT
AAGAAAATAGAAACAGCTACCATCAAATATATTGCTGATGAAAAACTGATTTCTTTATTT
CCAAGGTTACAAACGGGACCTAGGAACCTTAGTCACCACCCCCTTAAGCCGCCCACTAAA
AACCTGCCCCATACCCATCCGCCCCGAAACCGGTCTGCTGGCAATCATAAAGCTGATGCT
CTAGTCTCTTCCGCAATTAAAGAAGCACGACACTTTCATAATCTCACTCATGTCAATGCC
GCAGGACTCAAACACAAATACCCTCTCACATGGAAAGAAGCTAAACATATTGTACAGCGC
TGTTCACATTGCAAAGAGAGAATGGGAAAAACAACAAAGCAGGCAATGAAGCATACAACC
ACCACCTTCCAGACTCAACACCAAACAACCCCCGCCACGAGTCAATCCCACAAACAATTC
CGTCAAGTGCTAATCCTACCAACTCTGGCTCCAGGAGTTAATCCCAGAGGCTTGGCACCT
AACGCTCTTTGGCAAATGGATGTCACCCATGTTCCATCTTTTGGAAGACTAGCTTATGTA
CATGTATCAGTAGACACCTTTTCACATTTTATCTGGGCTACATGCCAAACAGGAGAAGGC
ACTGCCCGTGTTAAAAGACATATGTCTTCTTGTTTTGCGGTTATGGGCATTCCACCTCAG
ATTAAAACAGACAACGCCCCAGGCTATACCAGCAAAGCTTTTAAAAAATTTATTCAACAA
TGGAATATTAACCGCACTACTGGAATCCCTTATAAGCCCCAAGGACAGGCTCCAGTAGGA
GTGAGCAAATAACACTTCCAAAAAACAACAGTTACAAAAACAGAAAGAAAGAAAAAAGGA
ATTAAGTACCCCCCACAAGCAATTAAATCTGGCACTTCTGACTCTGAATCCTTCCATTTT
GTCAAAGCTCCGTCCTCTAATGGCAGCCGAACAACACTATACAGGCAATAAATTTTTTCG
GCACCTAAACAAGAAATTAAGAAACAAGCACAAAAAAAAAGAAGAAAAACAAATGACGCAC
TGGAAAGAACAAAGAACAAAAAGTTGGCAAATAGCTAAAAAAGAAAGTTCGGGCCCACTG
CCGGTGTTCTCTCCTTGATGTCTGGGAGCAGACCAGATTGGTAGACAAACACGAAAACCA
AAAACTAGGCAAGAGGAACAATCAGAACACGAGAAACAGGGTATGCTACAGGTCCACCAG
AAGAAAATCAATCCCCTGTTTGGGTCCCTACTAGAAATCCCTGAGTCCGTCTGAAGAATG
ACAATGAAAACAACAAGAAAAAAACATCAGGGCCACAAACCACCCGCAAACATAGCCAAA
ACTGGGCAAAAGAACAAAAAACAGACACGACAAACCAAAAACCATATAACCCAACAAGAC
CAGAGCACAAAGAAAACATAGAAATTTAATCCCAAAACCCCAAATTCCGTCACGATCAAC
ACACGCAAACATAAAAAACTAACCTAGGAAAAGACCAAGAAACATACACAGCTCAAAGCA
ATCCACCAAACCAAAGCAAAGAATCTCGCCCCAACGATGATTGTCATGATCACTCTGATA
CGCATAATCAACACTGCAGTAACTCTCCCTTACACCAAGCTGCATACAACAATAAATCTG
TCTCAGTGGACTTCTTTGCCTTTTCCTCCACTTATTCGACCCATCACATGGATGGATGCT
CCTGTAGAAGTCTATACTAACGATAGTGCTTGCATGCCTGGATCTATAGATGACCGTTGT
CCTGCTCAACCAGGAGAAGAAGGAACGCCTTTTAATGTTACCATTGGATATAAATATCCA
CCTTTGTGCCTGGGACATGCACCTGGTTGTATCCCATTAGATAATCAAAATTGGCTGGCG
ACACTACCAGCCGGCAACACTGATACGAAATAGGGACATATGGTCTCAGATCTCACAATT
AAACCTTTAAGATATACTATTACGGGTGTGGCAGACTACACTCAAAAATCTCAATATAAG
CCAATAGGAACCACGCCAGAGCAGACGAACTTGCTCGCAGTGCCAGACCCCTAAAAAGAC
CAAAAAAAGGGAATAAAAAACTAAAAATTTAATATGGAAAGATTGCATTAACGCACAAGC
AGAAGTGCTAAAAAATGATTCCCACAGAATCATTATTGACTGGGCCCCAAAGGGGCATTT
TAGGAATAATTGCTCTGCTCAGCAAACACAATGTCAGGAGGCTACCTATTTTATTGCTTA
TTAAGAGAATAGCGACCACCCTCACATATTAAAGGAAAGGTTGACCACATTCTGTCCCTC
TAATTGGAAAAATAAAGGCATTGCCTGCATGAGACCAGGAGCCAGGGTCGGCCTCCGAGG
GACAGAAAAAATAGAACCAAAAAAGAACCCTCAACAGTTAGAAATATGGAAATTGGCTAT

AGCCATATCCGGAATCAGAGTATGGGAAGGTGATAATAATAAATCTATTATAACAACTAA
AACACAAAAAAAAACAGATTTCTTTACACTATGATAAACACACACCCAAAAATATCACTAT
GGCAAACACCAAAACGCCAATACAAAAATCCGACAGGAAAGACGATAAGGACTAACAACC
CACCGATGCCAATCTACCCCCACAACATCCCAAACAACCTAGGACCCCCACCCCACAGTC
AAAAAAAAGAGGAATTGCCACCGCCGCTCCTCTCCCTCAGTATCAACCCAAAAACAGACA
TACTACCTAGATGAACTCCAACAAAACTATACCAATAAAAAGTTGTGTTAAACCACCATA
TATGTTATTAGTAGGAAACATAAATATTAGCACAAAAAATCAAACTATTAAATGCATTAA
TTGTAAATTGTATACTTGTATTGACTCAACATTTGATCCAAAAAAAAGTGTTATAATGGT
CAGAGCCAGAGAAGGAATATGGATACCAGTAACTTTACACAGACCTTGGGAATCCTCCCC
TTCAATCCATTTTATTAAAAAAATTCTACAGAGAATTCTTAAAAGAACTAAGAGATTTAT
TTTTACTTTAATTGCAGTGATAATGGGCTTAATTGCTGTTACTGCAACAGCCGCTACTGC
TGGAGTAGCATTACATCAATCTATTCAAACCGCTCATTTTGTGGATAAATGGCAAAAAAA
TTCCACCCGAATGTGGAATTCTCAGCCAGGCATTGATCAAAAATTGGCCAATCAAATTAA
TGATCTAAGACAGACTGTTATATGGATGGGAGATAGGATAATGAGTTTAGAACATCGAAT
GCAAATGCAATGTGATTGGAATACTTCTGATTATTGTATAACACCATATAGATATAATGA
GAATCACCACAGTTGGGAAACAGTAAAAAGCCATCTACAAGGAAGCGATGATAATTTATC
CTTAGACATAACAAAACTAAAAGAACAAATTTTTGAAGCCTCCCAAGCTCACTTAACTAC
TATACCCGGAGCTGAAGTGTTTGAAGGAATCGCAGAAAGATTATCTGATCTAAACCCCAT
TAAATGGATAAAATCTCTTGGAGGCTCCATTATTGTAAATATTGTACTGATTTTAATCTG
TTTTATTTGTTTGTTTTTAGTCTGCAGAACTCGACACAGAATCCTACGATAAAATCGTGA
CCAGGACCAAGCCATCATCGCAATTGTTGACTTAGCAAAAAAGAAACGGCGACAGAATAT
AAGAGCTGGGGAACGGGGTGGAT

Supplementary Material S3:

Figure 18: The regulatory motifs of 17bp subsequence of LTR in lincRNA.

LTR

Transcription factor	Seq1				
	Sequence	From	To	Score	Strand
Broad-complex_1	AAAATAAATAAATA	1	14	9.221	+
Broad-complex_4	AAATAAATAAA	2	12	9.487	+
HFH-2	AAATAAATAAAT	2	13	12.485	-
HFH-3	AAATAAATAAAT	2	13	12.439	-
HMG-IY	AAATAAATAAATAAAT	2	17	9.534	+
HNF-3beta	AAATAAATAAAT	2	13	11.759	-
Hunchback	AAATAAATAA	2	11	6.384	+
Athb-1	AATAAATA	3	10	7.226	-
HFH-1	AATAAATAAAT	3	13	7.914	-
SQUA	ATAAATAAATAAAT	4	17	8.950	+
Broad-complex_1	TAAATAAATAAATA	5	18	8.121	+
Broad-complex_4	AAATAAATAAA	6	16	9.487	+
HFH-2	AAATAAATAAAT	6	17	12.485	-
HFH-3	AAATAAATAAAT	6	17	12.439	-
HNF-3beta	AAATAAATAAAT	6	17	11.759	-
Hunchback	AAATAAATAA	6	15	6.384	+
Athb-1	AATAAATA	7	14	7.226	-
HFH-1	AATAAATAAAT	7	17	7.914	-
Athb-1	AATAAATA	11	18	7.226	-

A

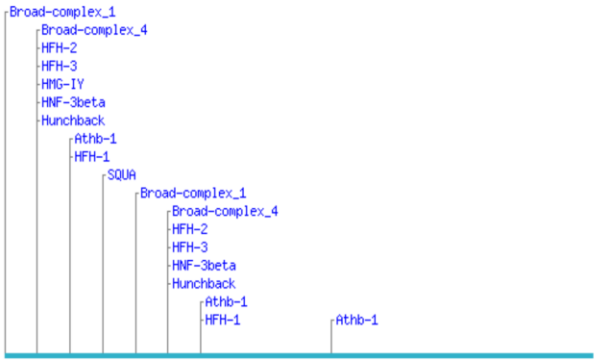


Figure 19: The regulatory motifs from two subsequence of ERV in lincRNA.

B

ERV

Transcription factor	Seq1				
	Sequence	From	To	Score	Strand
Pbx	TCTGATTGAGGA	9	20	10.707	-

Transcription factor	human				
	Sequence	From	To	Score	Strand
Myf	CGGGTGCTGGTG	4	15	7.989	+
Snail	CGGGTG	4	9	6.063	+
Snail	CTGGTG	10	15	6.770	+

Figure 20: The regulatory motifs of 150bp subsequence of MIR in lincRNA.

C

MIR

Transcription factor	Seq1				
	Sequence	From	To	Score	Strand
Snail	CAGGTT	2	7	7.278	+
Spz1	CAGGTTACCCT	2	12	7.659	-
Snail	CAGGAG	41	46	6.028	+
Snail	CAGGAG	50	55	6.028	+
Snail	CACCAG	67	72	6.770	-
Snail	CAGGAG	74	79	6.028	+
CREB	TCTGATGACGAC	80	91	10.371	+
bZIP910	ATGACGA	84	90	7.070	+

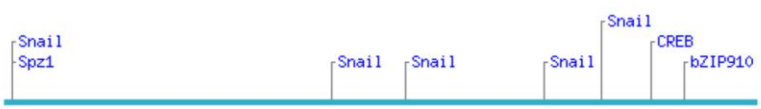


Figure 21: The regulatory motifs of 100bp subsequence of SVA in lincRNA.

SVA

D

Transcription factor	Seq1				
	Sequence	From	To	Score	Strand
COUP-TF	CTGTTCAAGGGGCC	4	17	8.853	-
NF-kappaB	GGCCAGTCCC	14	23	7.528	-
TBP	TCCCTCCATGTATAC	20	34	7.528	-
AGL3	CCATGTATAC	25	34	5.939	+
CF2-II	CCATGTATAC	25	34	8.425	-
CF2-II	ATGTATACAA	27	36	6.185	+
FREAC-4	ATGTATAC	27	34	8.269	-
FREAC-4	GTATACAA	29	36	8.102	+
Snail	AACCTG	35	40	7.278	-
SOX17	ACCATGGTC	44	52	6.411	+

Reference

1. Belancio, V.P., D.J. Hedges, and P. Deininger, *Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health.* Genome Res, 2008. **18**(3): p. 343-58.
2. Mirkin, S.M., *Expandable DNA repeats and human disease.* Nature, 2007. **447**(7147): p. 932-40.
3. Kunarso, G., et al., *Transposable elements have rewired the core regulatory network of human embryonic stem cells.* Nat Genet, 2010. **42**(7): p. 631-4.
4. Lynch, V.J., et al., *Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.* Nat Genet, 2011. **43**(11): p. 1154-9.
5. Cowley, M. and R.J. Oakey, *Transposable elements re-wire and fine-tune the transcriptome.* PLoS Genet, 2013. **9**(1): p. e1003234.
6. Pereira, V., D. Enard, and A. Eyre-Walker, *The effect of transposable element insertions on gene expression evolution in rodents.* PLoS One, 2009. **4**(2): p. e4321.
7. van de Lagemaat, L.N., et al., *Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions.* Trends Genet, 2003. **19**(10): p. 530-6.
8. Rebollo, R., M.T. Romanish, and D.L. Mager, *Transposable elements: an abundant and natural source of regulatory sequences for host genes.* Annu Rev Genet, 2012. **46**: p. 21-42.
9. Jacques, P.E., J. Jeyakani, and G. Bourque, *The majority of primate-specific regulatory sequences are derived from transposable elements.* PLoS Genet, 2013. **9**(5): p. e1003504.
10. Conley, A.B., J. Piriyapongsa, and I.K. Jordan, *Retroviral promoters in the human genome.* Bioinformatics, 2008. **24**(14): p. 1563-7.
11. Medstrand, P., J.R. Landry, and D.L. Mager, *Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans.* J Biol Chem, 2001. **276**(3): p. 1896-903.
12. Franchini, L.F., et al., *Convergent evolution of two mammalian neuronal enhancers by sequential exaptation of unrelated retroposons.* Proc Natl Acad Sci U S A, 2011. **108**(37): p. 15270-5.
13. Sela, N., et al., *Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome.* Genome Biol, 2007. **8**(6): p. R127.
14. Schoeftner, S. and M.A. Blasco, *Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II.* Nat Cell Biol, 2008. **10**(2): p. 228-36.
15. Azzalin, C.M., et al., *Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends.* Science, 2007. **318**(5851): p. 798-801.
16. Nilsson, M.A., et al., *Expansion of CORE-SINEs in the genome of the Tasmanian devil.* BMC Genomics, 2012. **13**: p. 172.
17. Kelley, D. and J. Rinn, *Transposable elements reveal a stem cell-specific class of long noncoding RNAs.* Genome Biol, 2012. **13**(11): p. R107.
18. Mc, C.B., *The origin and behavior of mutable loci in maize.* Proc Natl Acad Sci U S A, 1950. **36**(6): p. 344-55.

19. Kramerov, D.A. and N.S. Vassetzky, *Short retroposons in eukaryotic genomes.* Int Rev Cytol, 2005. **247**: p. 165-221.
20. McCarthy, E.M. and J.F. McDonald, *Long terminal repeat retrotransposons of Mus musculus.* Genome Biol, 2004. **5**(3): p. R14.
21. Singer, M.F., *SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes.* Cell, 1982. **28**(3): p. 433-4.
22. King, R.C. and W.D. Stansfield, *A dictionary of genetics*. 5th ed. 1997, New York: Oxford University press. vii, 439 p.
23. Kajikawa, M. and N. Okada, *LINEs mobilize SINEs in the eel through a shared 3' sequence.* Cell, 2002. **111**(3): p. 433-44.
24. Jurka, J., *Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons.* Proc Natl Acad Sci U S A, 1997. **94**(5): p. 1872-7.
25. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
26. Chen, C., T. Ara, and D. Gautheret, *Using Alu elements as polyadenylation sites: A case of retroposon exaptation.* Mol Biol Evol, 2009. **26**(2): p. 327-34.
27. Roy-Engel, A.M., et al., *Human retroelements may introduce intragenic polyadenylation signals.* Cytogenet Genome Res, 2005. **110**(1-4): p. 365-71.
28. Shen, S., et al., *Widespread establishment and regulatory impact of Alu exons in human genes.* Proc Natl Acad Sci U S A, 2011. **108**(7): p. 2837-42.
29. Sela, N., et al., *Characteristics of transposable element exonization within human and mouse.* PLoS One, 2010. **5**(6): p. e10907.
30. Vorechovsky, I., *Transposable elements in disease-associated cryptic exons.* Hum Genet, 2010. **127**(2): p. 135-54.
31. Chen, L.L., J.N. DeCerbo, and G.G. Carmichael, *Alu element-mediated gene silencing.* EMBO J, 2008. **27**(12): p. 1694-705.
32. Levanon, E.Y., et al., *Systematic identification of abundant A-to-I editing sites in the human transcriptome.* Nat Biotechnol, 2004. **22**(8): p. 1001-5.
33. Hohjoh, H. and M.F. Singer, *Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon.* EMBO J, 1997. **16**(19): p. 6034-43.
34. Kolosha, V.O. and S.L. Martin, *In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition.* Proc Natl Acad Sci U S A, 1997. **94**(19): p. 10155-60.
35. Feng, Q., et al., *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell, 1996. **87**(5): p. 905-16.
36. Mathias, S.L., et al., *Reverse transcriptase encoded by a human transposable element.* Science, 1991. **254**(5039): p. 1808-10.
37. Fanning, T. and M. Singer, *The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins.* Nucleic Acids Res, 1987. **15**(5): p. 2251-60.
38. Weiner, A.M., *SINEs and LINEs: the art of biting the hand that feeds you.* Curr Opin Cell Biol, 2002. **14**(3): p. 343-50.
39. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.
40. Team, R.D.C., *R : A language and environment for statistical computing*. 2010, R Foundation for Statistical Computing.
41. Vance , A., *Data Analysts Captivated by R's Power.* New York Times, 2009.
42. Dreszer, T.R., et al., *The UCSC Genome Browser database: extensions and updates 2011.* Nucleic Acids Res, 2012. **40**(Database issue): p. D918-23.
43. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

44. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.
45. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
46. Meyer, L.R., et al., *The UCSC Genome Browser database: extensions and updates 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D64-9.
47. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
48. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.* Mol Biol Evol, 2009. **26**(7): p. 1641-50.
49. Han, M.V. and C.M. Zmasek, *phyloXML: XML for evolutionary biology and comparative genomics.* BMC Bioinformatics, 2009. **10**: p. 356.
50. Talavera, G. and J. Castresana, *Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.* Syst Biol, 2007. **56**(4): p. 564-77.
51. Huang da, W., et al., *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.* Genome Biol, 2007. **8**(9): p. R183.
52. McCue, A.D. and R.K. Slotkin, *Transposable element small RNAs as regulators of gene expression.* Trends Genet, 2012. **28**(12): p. 616-23.
53. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
54. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
55. Belancio, V.P., D.J. Hedges, and P. Deininger, *LINE-1 RNA splicing and influences on mammalian gene expression.* Nucleic Acids Res, 2006. **34**(5): p. 1512-21.
56. Belancio, V.P., A.M. Roy-Engel, and P. Deininger, *The impact of multiple splice sites in human L1 elements.* Gene, 2008. **411**(1-2): p. 38-45.
57. Cenik, C., et al., *Genome analysis reveals interplay between 5'UTR introns and nuclear mRNA export for secretory and mitochondrial genes.* PLoS Genet, 2011. **7**(4): p. e1001366.
58. Cenik, C., et al., *Genome-wide functional analysis of human 5' untranslated region introns.* Genome Biol, 2010. **11**(3): p. R29.
59. Barrett, L.W., S. Fletcher, and S.D. Wilton, *Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements.* Cell Mol Life Sci, 2012. **69**(21): p. 3613-34.
60. Faulkner, G.J., et al., *The regulated retrotransposon transcriptome of mammalian cells.* Nat Genet, 2009. **41**(5): p. 563-71.
61. Kim, D.D., et al., *Widespread RNA editing of embedded alu elements in the human transcriptome.* Genome Res, 2004. **14**(9): p. 1719-25.
62. Cohen, C.J., W.M. Lock, and D.L. Mager, *Endogenous retroviral LTRs as promoters for human genes: a critical assessment.* Gene, 2009. **448**(2): p. 105-14.
63. Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr., *Exon shuffling by L1 retrotransposition.* Science, 1999. **283**(5407): p. 1530-4.
64. Costa, F.F., *Non-coding RNAs, epigenetics and complexity.* Gene, 2008. **410**(1): p. 9-17.

65. Martianov, I., et al., *Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript.* Nature, 2007. **445**(7128): p. 666-70.
66. Caretti, G., et al., *The RNA helicases p68/p72 and the noncoding RNA SRA are coregulators of MyoD and skeletal muscle differentiation.* Dev Cell, 2006. **11**(4): p. 547-60.
67. Siomi, M.C., et al., *PIWI-interacting small RNAs: the vanguard of genome defence.* Nat Rev Mol Cell Biol, 2011. **12**(4): p. 246-58.