STRUCTURAL, FUNCTIONAL AND EVOLUTIONARY CHARACTERIZATION OF SENSE-ANTISENSE TRANSCRIPTS IN MAMMALS

A Dissertation

by

CHARLES MICHAEL DICKENS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2009

Major Subject: Genetics

STRUCTURAL, FUNCTIONAL AND EVOLUTIONARY CHARACTERIZATION OF SENSE-ANTISENSE TRANSCRIPTS IN MAMMALS

A Dissertation

by

CHARLES MICHAEL DICKENS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Christine G. Elsik
	Loren C. Skow
Committee Members,	James E. Womack
	David L. Adelson
	Penny K. Riggs
Chair of Genetics Faculty,	Craig J. Coates

May 2009

Major Subject: Genetics

ABSTRACT

Structural, Functional and Evolutionary Characterization

of Sense-Antisense Transcripts in Mammals.

(May 2009)

Charles Michael Dickens, B.S., Texas A&M University;

M.S., University of Arkansas

Co-Chairs of Advisory Committee: Dr. Christine G. Elsik
Dr. Loren C. Skow

Sense-antisense transcripts (SATs) are messenger RNA (mRNA) transcripts that have regions that are complementary to regions of other mRNA transcripts. SATs may play an influential role in the regulation of gene expression. One evolutionary event that has had a dramatic impact on many genomes is the widespread dispersal of repetitive sequences which includes transposable elements (TEs) as well as simple and tandem repeats. Approximately 45% of the human and 37.5% of the mouse genomes are composed of repeats derived from transposable elements. A group of SATs was identified as resulting from transposable elements integrating into the coding strand of some genes and into the template strand of the coding region of other genes. These SATs may add to the complexity of an organism's regulatory network or they may be the result of rather recent TE activities yet to succumb to sequence divergence.

The human, mouse and bovine genomes were analyzed for SATs using publicly available datasets and bioinformatics analysis tools. Each sense-antisense binding region (SABR) was aligned to transposable elements from the RepBase repeat database revealing many SABRs containing TE sequence in a large portion of the sequence. A Gene Ontology analysis on subsets of the data showed enrichments for the functional category of ‘DNA repair’ and the component category ‘cytoplasm’.

An analysis of the substitution rates in human and mouse across the 3' UTRs of transcripts containing SABRs at the 5' end of their 3' UTRs showed that the substitution rate in the region of the SABR was lower than compared to the beginning of the 3' UTR. The lower percent GC composition found at the 3' end of the 3' UTRs could be attributed to conserved poly-A signals in this region.

DEDICATION

I dedicate this dissertation to Raymond D. Broussard, TAMU class of ‘54, my BSA Scoutmaster (Troop 155), my mentor and my friend.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

Sense-antisense transcripts (SATs) are messenger RNA (mRNA) transcripts that have regions complementary to other mRNA transcripts. SATs have been demonstrated to function in gene regulation by two types of mechanisms depending on their genomic arrangement. SATs located at the same genomic locus may regulate each other's expression through transcriptional interference as was demonstrated by using atomic force microscopy showing RNA polymerases stalled during simultaneous transcription of *cis* genes (Crampton et al. 2006). SAT regulation in *trans* by non-coding transcripts has also been demonstrated with the binding of separate mRNA transcripts at the post-transcriptional level of gene expression resulting in the reduction of the polygalactouronase enzyme (Smith et al. 1988). These two forms of gene regulation demonstrate the potential for SATs to regulate expression levels in *cis* or *trans* at either the transcriptional or post-transcriptional phases of gene expression. Understanding gene regulation and how SATs may play a role is important in designing antisense treatments for diseases such as cancer in which gene expression has often been significantly altered by changes in biological mechanisms or regulatory elements (Hanahan and Weinberg 2000; Wacheck and Zangemeister-Wittke 2006).

Sense-Antisense Transcripts

Although antisense transcripts have been a well-known phenomenon in prokaryotes for many years (Simons 1988), they were discovered more recently in

---

This dissertation follows the style and format of *Genome Research.*

eukaryotes (Knee and Murphy 1997). The first evidence of antisense RNA transcription was reported by Williams and Fried in 1986 using mouse as a model organism (Terryn and Rouze 2000). Sense-antisense transcripts may be transcribed from the same genomic locus (*cis*) or from different loci (*trans*). The *cis* SATs can be grouped into three main categories: convergent, divergent and containment (Figure 1).

Convergently overlapping pairs were the most abundant class of *cis* SATs found in some organisms (Jen et al. 2005; Zhang et al. 2006). However, a different *cis* SAT identification approach found the divergent class to be more abundant than convergent (Zhang et al. 2006).

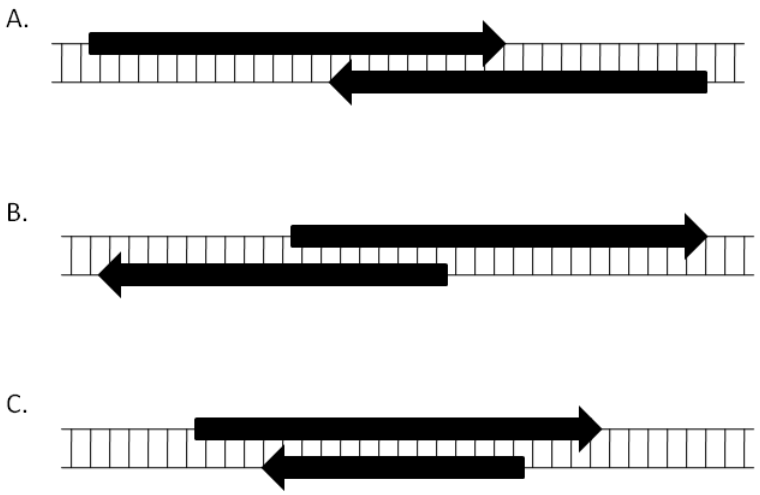


Figure 1.
Genomic arrangement of three types of *cis* SATs. Convergent (A), divergent (B) and containment (C) arrangements of genes at the same locus that form SATs.

Although the roles of SATs are not fully known, specific SATs have been demonstrated to play active roles in genomic imprinting (Sleutels et al. 2002), skeletal development (Blin-Wakkach et al. 2001), eye development (Alfano et al. 2005), cell

proliferation inhibition (Chao and Spicer 2005) and circadian cellular process regulation (Kramer et al. 2003). SATs have been found to be co-expressed in some tissues and differentially expressed in others (Vanhee-Brossollet and Vaquero 1998; RIKEN 2005). Tissue specific gene expression of *trans* SATs in *Arabidopsis thaliana* indicated that most *trans* SAT pairs were co-expressed in the same tissues and cells suggesting that *trans* SATs have the potential for interaction in the cell (Wang et al. 2006). The X inactivation gene Xist is negatively regulated by a *cis* antisense transcript called Tsix which is implicated in Xist chromatin modification (Ohhata et al. 2008).

Naturally occurring non-coding antisense RNA transcripts target specific RNAs for regulatory purposes (Wagner and Simons 1994). Therefore, the sense-antisense binding regions (SABRs) of coding transcripts forming SATs may also play an influential role in gene regulation. Some researchers have suggested that SATs play a role in gene regulation by forming the long duplexes essential for post-transcriptional regulatory mechanisms, mRNA transport and RNA stability (Lipman 1997; Kumar and Carmichael 1998). Antisense transcripts have been demonstrated to regulate gene expression *in vivo* such as the *trans* regulation of polygalactouronase (PG) levels within a transformed plant containing a 730 nucleotide (nt) fragment from the PG clone DNA (cDNA) (Smith et al. 1988). Also antisense post-transcriptional gene silencing was accomplished using an antisense gene in Virginia pine (Tang et al. 2005). *cis* SATs may also have regulatory roles at the transcription level such as the colliding transcriptional mechanisms called 'transcriptional interference' proposed by Prescott and Proudfoot 2002, and later validated using atomic force microscopy (Crampton et al. 2006). Additionally, an expression analysis of *cis* SATs was observed to be consistent with the transcriptional interference model (Osato et al. 2007). Other researchers have found gene regulation by

SATs to be a result of RNA-dependent epigenetic modifications and not transcriptional interference (Camblong et al. 2007).

Antisense transcripts that bind to their sense complement in the nucleus may prevent an mRNA from being transported out of the nucleus and into the cytoplasm. This along with the tendency of SATs to be poly-A negative and nuclear localized, suggest that the primary regulatory role of SATs as regulators of gene expression may occur in the nucleus (Kiyosawa et al. 2005). The expression correlation of sense and antisense genes may provide some support for expression regulation of one transcript by the expression levels of another. A genome-wide analysis of *cis* SATs found that co-expressed and inversely expressed patterns of SAT pairs are from a group of evolutionary conserved pairs (Chen et al. 2005).

Transposable Elements

The insertion of a transposable element (TEs) in two different genes can create a SAT pair. TE activity has played a role in the formation of *trans* SATs such as the mouse B1 SINE copying itself into the coding region of one gene and into the template strand of an exon of another gene (Figure 2). This same transposon copied into various exons and template strands of exons of various genes can create a group of transcripts that have similar SABRs. The SABR of a single transcript matching various antisense transcripts supports the Class I transposon 'copy and paste' mechanisms for *trans* SAT formation.

Researchers found that TE insertions in exonic regions were less frequent in the coding and more frequent in the template strand at 47.3% and 52.7% respectively in the cow genome (Almeida et al. 2007). Other researcher also found TEs more likely to be inserted into the template strand of genes (Makalowski et al. 1994; Lorenc and Makalowski 2003).

TE insertions can be advantageous resulting in new regulatory functions (Britten 1996; Brosius 1999b; Jordan et al. 2003; van de Lagematt et al. 2003; Thornburg et al. 2006) or they may be deleterious resulting in genetic diseases (Wallace et al. 1991; Tufarelli et al. 2003). Alignments of human and mouse sequences to protein domains from the PROSITE (Hulo et al. 2006) database revealed a small number of exonized TEs with the potential to either contribute to or not affect protein functionality (Sela et al. 2007). Alignments of transposable element sequences to PDB entries found a small number of proteins with fragments encoded by TE cassettes (Gotea and Makalowski 2006).

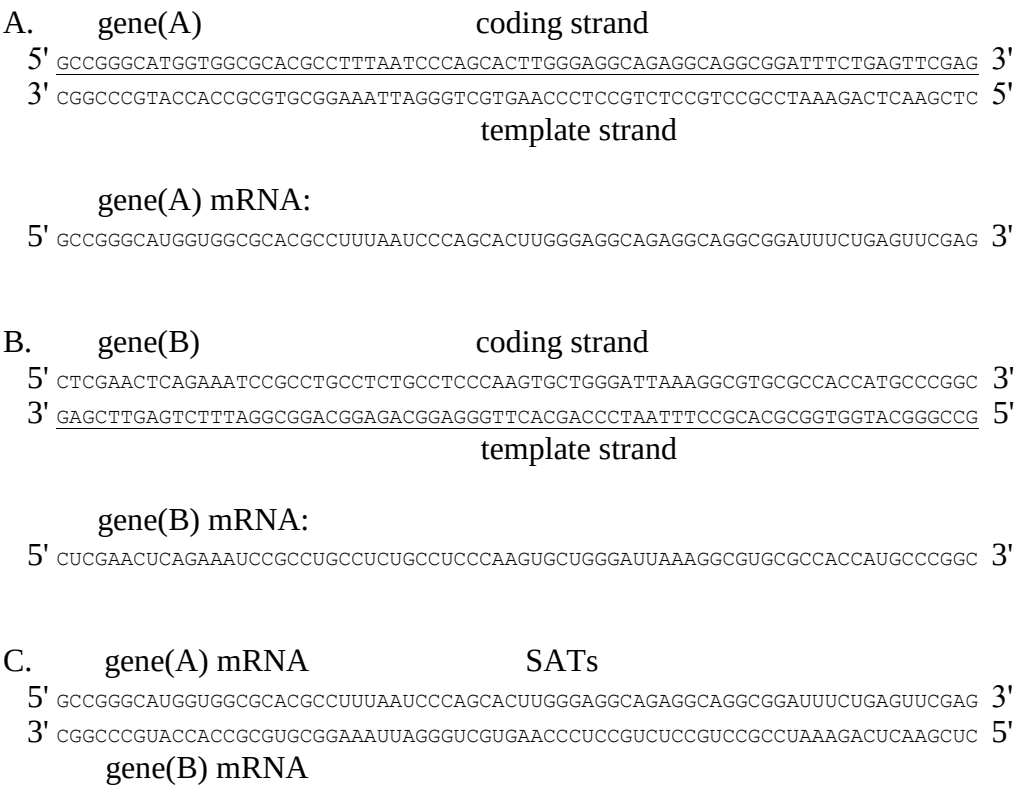


Figure 2.
Example of a *trans* SAT pair created by a transposable element inserted into two genes. Example of a portion of the mouse B1 SINE (underlined) integrated into an exon of one gene (A) and into the template strand of an exon of another gene at a different locus (B) creating a *trans* SAT pair of mRNAs (C).

Evidence suggests that TE activity has shaped various genomes (reviewed in Cordaux et al. 2006); however, TE insertion is usually detrimental when inserted into the coding or regulatory regions of genes. Some TE sequences in genes have withstood many years of purifying selection. Understanding past transposon activities may help us understand why some genes with TE insertions remain functional. Additionally SAT research can help the growing research interest in using artificial antisense oligonucleotides transcripts for therapeutic purposes (Dai et al. 2007).

Transposable elements make up a large percentage of many genomes and have contributed to a number of protein coding sequences (Brosius 1999a; Nekrutenko and Li 2001). TEs have been commonly labeled as 'junk' DNA with the assumption that they served on purpose. Today, many researchers regard TEs as parasitic (Rouzic et al. 2007) while others argue that the relationship between TEs and their host are more symbiotic (Brosius 1999a) since there are some examples of TEs that have become functional non-coding elements (Bejerano et al. 2006).

Transposable elements are relatively short sequences in the genome that can move around to different loci on the same or different chromosome within an organism. First discovered by Barbara McClintock in maize around 1950, TEs are categorized as Class I and Class II. Class I TEs use an mRNA intermediate for transposition. This results in two genomic loci containing the transposon sequence. Class II TEs have a DNA intermediate which is excised from its genomic location and moved to another region of the genome.

Class I TEs are termed retroposons or retrotransposons because of their retrotransposition capabilities using an RNA intermediate. Most mammalian retrotransposons fall into three classes, short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs) and retrovirus-like elements (IHGSC 2001).

SINEs and LINEs make up the largest proportion of interspersed repeats in human.

LTR Retrotransposons

Retrotransposons that contain long terminal repeats (LTRs) at both ends are called LTR-transposons. Many LTRs have contributed to gene evolution (DeBarry et al. 2006) and the formation of various SATs.

LINEs

Retrotransposons that do not contain LTRs are called non-LTR retrotransposons and contain a polyadenylation sequence at the 3' terminus. These non-LTR retrotransposons include LINEs and SINEs. L1 is an example of a family of LINE elements found in mammalian genomes. A full length Line element can be anywhere from 4-6 kb in length and contain two coding regions, one for the nucleic acid binding protein and another for an endonuclease with reverse transcriptase. LINEs can be grouped into families based on their sequence divergence in their 3' end, which can be used as a temporal estimate of their insertion relative to other LINEs (Smit et al. 1995).

SINEs

SINEs range from 100-400 base pairs (bp) and usually average around 300 bp and are the most abundant TE in human gene promoter regions (Thornburg et al. 2006). Unlike LINEs, SINEs do not code for any proteins, and rely on reverse transcriptase produced by LINEs for replication. By far the most abundant SINE in the human genome is the Alu element with over a million copies dispersed throughout the genome. Although Alu sequences have created many SATs, only a small percentage (< 0.05) is still active

(Mills et al. 2007).

The origin of the primate specific Alu element dates back about 50 million years ago to the precursor sequence 7SL RNA. The abundance of Alu sequences in the human genome and not the chimp indicate that Alu proliferation was specific to the human genome (Hedges et al. 2004). The Alu elements do not code for their own reverse transcriptase but instead are mobilized in *trans* by utilizing reverse transcriptase produced by the L1 elements. TE sequences have been identified in thousands of human proteins and in some cases constitute greater than 80% of a gene sequence (Nekrutenko and Li 2001; Britten 2004). B1 is an abundant SINE in the mouse genome and originated from an ancestral Alu sequence (Kriegs et al. 2007). SINEs have been used as phylogenetic markers showing the relationship between extant species (Nikaido et al. 1999).

Retrovirus-like Elements

Retrovirus-like elements (RLEs) code for proteins similar to those in vertebrate retroviruses. They have long terminal repeats which can be used to estimate when the RLE insertion occurred and can be used as taxonomic markers (Cantrell et al. 2001). Vertebrate-specific endogenous retroviruses (ERVs) are the only RLEs that have been active in mammalian genomes (IHGSC 2001). RLEs have become such an integrated part of our genome that the presence or absence of specific ERV terminal repeats can be used as markers for disease risk in humans (Donner et al. 1999). RLEs can create SABRs similar to other transposable elements.

DNA Transposons

DNA transposons are characterized by short inverted terminal repeats which are

important for transposase recognition and can be either autonomous, containing an encoded transposase, or non-autonomous containing no transposase coding region. DNA transposons that do not code their own transposase rely on transposase produced by other transposons. DNA transposons move throughout the genome in a "cut and paste" manner and are suggested as playing a role in chromosomal rearrangement (Smit 1996).

Effect of TEs on Alternative Splicing

TEs have generated many SATs and they can also affect gene regulation in other ways such as restructuring a gene through alternative splicing. Around 98.71% of splice sites in mammals contain the canonical dinucleotides GT as donor sites and AG as acceptor sites. An additional 0.56% of splice junctions contain the non-canonical dinucleotides GC – AG and the remaining 0.73% are small groups of different donor and acceptor dinucleotides (Burset et al. 2000). TEs have been implicated in gene evolution by exonization in which an exon is created due to a TE sequence containing canonical splice sites either in the sense or complimentary antisense sequence (Sorek et al. 2002; Lev-Maor et al. 2003; Kazazian 2004; Piriyapongsa et al. 2007; Wu and Sun 2007) and by intronization in which an Alu insertion into an exon creates an intron within that exon (Sela et al. 2007). These introduced splice sites can lead to transcript isoforms which may have premature stop codons and may be transcribed and translated into non-functional proteins. TE activity has also been found to increase the frequency of double stranded breaks, which can affect genome integrity (Gasior et al. 2006; Hedges and Deininger 2007). They also provide potential genomic regions for homologous recombination that can enhance or disrupt chromosome integrity (Sen et al. 2006). Some TEs can integrate into a gene disrupting its structure and causing non-functional transcripts which can lead

to diseases such as neurofibromatosis type 1 (NF1). This condition has been shown to be a result of an *Alu* element inserted into the intronic region of the NF1 gene (Wallace et al. 1991).

TEs have been found in human promoter regions (Jordan et al. 2003) and have been shown to play important functional roles (Britten 1997; Jordan et al. 2003). Some genes are derived entirely from TE sequences (Britten 2004). Often a fraction of a TE sequence, called a TE cassette, is found in an exon due to introduced splice sites or a deletion after homologous recombination. TE cassettes tend to be alternatively spliced in the human genome (Wu et al. 2007).

Gene Regulation

SATs may play a role in gene regulation. Genes are regulated both at the transcript and protein levels. Greater than 95% of mammalian RNA transcripts remain in the nucleus where they are broken down and eventually recycled (Jackson et al. 2000). Upon transcription, RNA transcripts form secondary structures such as stem-loops leading to a double-stranded RNA (dsRNA). The dsRNA molecules can lead to a number of RNA altering pathways such as adenosine deaminase that acts on RNA (ADAR) and RNA interference (RNAi).

RNA Editing

Double-stranded RNAs can be edited by a class of enzymes known as ADAR (Bass 2002; DeCerbo and Carmichael 2005). These ADAR enzymes can act on dsRNA, such as the SABRs created by SATs, by transforming or ‘editing’ adenosines to inosines (A-to-I) by hydrolytic deamination leading to the alteration of codons (Bass 2002). The

alteration occurs because inosine is translated as if it were guanosine (Basilio et al. 1962). This RNA editing phenomenon was first discovered in the mitochondria of trypanosomes (Benne et al. 1986). ADAR2 is involved in the creation of the canonical splice junction AG by editing intronic AA to AI which mimics the conserved AG acceptor site (Reuter et al. 1999).

A-I editing plays a role in gene expression (Häsler and Strub 2007) and is necessary for an organism's survival (Higuchi et al. 2000; Hartner et al. 2004). The Alu transposable element has been identified to be prime sites for RNA editing (Kim et al. 2004).

Premature stop codons introduced by Alu sequences have been found to be edited out by an RNA-editing event (Lev-Maor et al. 2007). Editing of repetitive elements found in transcripts may be the predominant role for the RNA editing mechanism (Athanasiadis et al. 2004).

ADAR enzymes work on dsRNAs with a minimal length of 25-30 bp and are most efficient with duplexes of greater than 100 bp (DeCerbo and Carmichael 2005). Past studies in mice have shown that hyperedited RNAs, those with a double stranded region of greater than 50 bp and at least 20% A-to-I editing, tend to remain in the nucleus (DeCerbo and Carmichael 2005). Long SAT SABRs could provide double stranded RNA precursors necessary for the ADAR enzyme to carry out RNA editing processes, consequently changing the mRNA sequence and affecting the ability of a protein to fold properly.

RNAi

Duplexes of RNAs formed by two separate transcripts such as SATs may form dsRNA regions leading to the RNA interference (RNAi) pathway. The RNAi mechanism uses members of the RNase III enzyme family such as Dicer and Drosha to cleave dsRNA at specific positions on a transcript. Drosha generates pre-miRNA intermediates upon recognition and cleavage of the ~80 nucleotide RNA hairpin structures (Xuezhong et al. 2004). However, a study set of *cis* SATs in *Arabidopsis thaliana* did not find evidence for a role of small RNAs in the disproportionate regulation of gene expression (Henz et al. 2007).

mRNA Masking

Nearly all eukaryotic mRNAs have poly-A tails which assist the transcript in transport, stability and translation (Hu et al. 2005). Translation generally occurs after poly-A elongation and in some cases in conjunction with other poly-A independent mechanisms (Vardy and Orr-Weaver 2007). There are specific hexamer poly-A signals near the end of mRNA sequences such as AAUAAA or a close variant. Many proteins and mRNAs are deposited maternally into the oocyte during oogenesis. These maternal mRNAs are subject to translation regulation by mRNA binding proteins that bind to specific sequences in the 3' UTR (Fu et al. 1999, Loning et al, 1999). This form of translational repression, also called mRNA masking, can occur in an organism by utilizing repressor proteins that recognize a short sequence in the 3' UTR of mRNA transcripts. However, it has been found that when a 3' UTR repressor site forms a dsRNA duplex with an antisense transcript or the repressor site is removed, the transcript can become translationally active (Standart et al. 1990). In this manner, SATs may function

in gene activation upon binding over the repressor site of other transcripts.

Nonsense-Mediated mRNA Decay (NMD)

NMD is a mechanism by which alternative transcripts with premature stop codons, such as those created by the activity of TEs, can be degraded preventing energy being expended in the synthesis of truncated proteins (Frischmeyer and Dietz 1999). Studies have shown that mouse embryos cannot develop if they lack the NMD protein Upf1; however, the loss of Upf1 is tolerated in lower eukaryotes (Medghalchi et al. 2001). In NMD, a premature stop codon must be located more than 50 nucleotides upstream of the final intron position (Hillman et al. 2004). When introns are removed from a transcript, exon-junction complexes are deposited at the splice sites. The recognition of these complexes assists in transport from the nucleus (Hillman et al. 2004). As the ribosome moves along the transcript during translation, the exon-junction complexes are displaced. If one or more of the splicing junction complexes is located sufficiently downstream of the stop codon, the transcript is degraded. NMD may also serve to regulate the levels of mRNA production and degrade transcripts that can produce harmful proteins (Maquat 2005).

Estimating Sequence Evolution

Some SAT SABRs have a high percent complementation indicating that they may be conserved for a biological purpose. Further investigation into this group can help understand the evolutionary history of these SABRs and provide clues as to why they have maintained a high level of conservation.

Sequence evolution can be estimated by identifying the number and types of

nucleotide substitutions in orthologous sequences. Nucleotide substitutions in coding regions can be classified as one of two types of substitutions; transitions or tranversions. Transitions are mutation of a purine to a purine such as A to G or pyrimidine to pyrimidine such as T to C. Transversions are changes of a nucleotide from a purine to a pyrimidine or pyrimidine to a purine. Transitions occur more often than transversions since transitions do not result in a change in the number of chemical ring structures. Transversions result in the addition or loss of a ring in the chemical structure of the resulting substitution at the mutation site. The different rates of transition and transversions are important when estimating selection based on the type of replaced nucleotide. Nucleotide substitutions in coding regions can also be classified as synonymous or nonsynonymous.

The genetic code is degenerate such that substitutions at certain codon position, especially at the third 'wobble' position, result in a codon for the same amino acid. These mutations or substitutions are called silent or synonymous substitutions (K_S) since the protein sequence is unchanged. However, if a mutation results in a codon change resulting in a change in the amino acid sequence it is called a nonsynonymous substitution (K_A). Therefore, substitutions at the third position of a codon are thus more common than at the first two positions since there is a greater chance of a third codon position substitution resulting in the same amino acid. Substitutions in the first two codon positions are more likely to result in changes in the amino acid sequence which can be detrimental to the organism.

Determining synonymous and nonsynonymous rates is not trivial, because some amino acid changes have resulted from substitutions in multiple codon positions at different times. There can be different pathways to reach the resulting codon since it is

difficult to determine which codon position substitution occurred first. Thus models of evolution are used to predict the rates of nonsynonymous and synonymous substitution.

The Nei-Gojobori codon based method (Nei and Gojobori, 1986) determines the number of synonymous and nonsynonymous substitutions based on the number of potential synonymous and nonsynonymous sites respectively. The two rates are normalized by dividing by the number of each potential site respectively determining the p-distance. Additionally, the p-distance, which is defined as the proportion of nucleotide sites differences of two compared sequences, can be adjusted using the Jukes-Cantor correction (Jukes and Cantor, 1969) to adjust for multiple substitutions at the same site for synonymous and nonsynonymous sites and are often referred to as d_S and d_N respectively. This correction is useful when the time of divergence between two species is high since the probability of a second substitution at the same site as a previous substitution increases over time.

The proportion of nonsynonymous to synonymous substitutions (K_A / K_S) can give an estimate of the relative type of selection that is occurring. Selection is considered neutral or no selection occurring when ($K_A = K_S$). Purifying or negative selection is inferred when ($K_A < K_S$) and positive or Darwinian selection is inferred when ($K_A > K_S$).

Estimating substitution rates in non-coding sequence such as UTRs is different than for coding sequences since the measurement of synonymous and nonsynonymous changes is reserved for protein coding sequences. Simply determining if there was a substitution is not sufficient to estimate the rate of sequence change. The probabilities of a transition or a transversion occurring are different and must be incorporated in the calculation methods. The Kimura two-parameter model was developed to account for the differences in transition and transversion substitution rates (Kimura, 1980).

Determining percent GC content for regions of the 3' UTR can also supplement conservation studies of the SABR and the rest of the 3' UTR. Previous studies have identified a negative correlation between percent GC content and the level of conservation in mammalian 3' UTRs (Shabalina et al. 2003). If SABRs lie in regions of relatively high GC content, then conserved SABRs in regions of increased conservation may be due to GC content alone.

Conservation of SATs

If SATs play a role in gene conservation, then the conservation of SAT pairs can give insight into which genes may be regulated by these mechanisms. The evaluation of the conservation of SATs depends on the definition of a SAT. Researchers usually set a high percent complementarity within the overlapping region which limits the number of overall SATs (Li et al. 2006). Lowering the percent complementarity in the SABR will increase the number of overall SATs, but SATs of low percent complementarity are unlikely to have a biological function due to reduced stability of dsRNA complexes. The number of conserved SAT pairs was generally low for a study of SATs in a population of three yeast, one fungi and one microsporidian species (Steigele and Nieselt 2005), and conservation was only about 20% for all human and mouse antisense genes in a different study (Veeramachaneni et al. 2004). The number of conserved *cis* and *trans* SATs using a HomoloGene search approach was 736 for human and mouse (Galante et al. 2007). Only a small percentage of identified human and mouse orthologous antisense transcripts are conserved across the two species (Veeramachaneni et al. 2004).

The antisense transcripts of two particular genes, thymidylate synthase (TS) and hypoxia inducible factor-1alpha (HIF), were found at 1,000-fold higher concentration in

the nucleus than in the cytoplasm; however, the two gene sense strands had similar expression levels in the cytoplasm and nucleus, suggesting that the antisense transcripts for these two genes may not block transport into the cytoplasm and that antisense transcript binding activity may be most pronounced in the nucleus (Faghihi and Wahlestedt 2006). Additional support for this theory comes from the findings that only about 5% of mRNAs are exported to the cytoplasm for translation while the remaining 95% are broken down in the nucleus (Jackson et al. 2000).

Structure and Functional Annotation

Genome sequencing is just an initial step in one approach to the identification of genes and the biological function of their products. Protein structure determination is done on a protein-by-protein basis. Solved structures are submitted to data repositories such as the Protein Data Bank (PDB) (http://www.rcsb.org/pdb). The PDB provides confirmed structure sequences in text format as well as 3-D images. With the growing number of databases storing functional data for various species, it is beneficial to have common terms or controlled vocabularies to describe functions across species. The Gene Ontology (GO) project is dedicated to using a controlled vocabulary set to describe functions, processes and components of genes and their products.

Objectives of this Study

The goals of this project were to structurally characterize SATs that were identified in three different organisms; human, mouse and cow and investigate why the percent complementation of some sense-antisense binding regions (SABRs) are more conserved than others. The Gene Ontology was used to give insight to possible regulatory

mechanisms shared among organisms or mechanisms that are unique to each species. Structures in the PDB were used to determine potential effects of TE insertions on protein folding.

Sequence evolutionary rates were estimated to determine whether SABRs in the 3' UTR evolve at a different rate than the rest of the 3' UTR and the coding sequence. Finding significantly lower substitution rates in a 3' UTR SABR compared to the rest of the 3' UTR would suggest SABR functionality. However, a genome wide orthologous gene substitution rate could show that the 3' UTR localized SABR is no different than the rest of the 3' UTR

CHAPTER II

STRUCTURAL CHARACTERIZATION OF SENSE-ANTISENSE TRANSCRIPTS

Overview

The integration of transposable elements (TEs) into the coding and template strands of the coding and UTRs may be an important contributor to the formation of sense-antisense transcripts (SATs). The goals of this project were to characterize the structure of SATs that were identified in three different organisms; human, mouse and cow and to investigate why the percent complementation of some sense-antisense binding regions (SABRs) are more conserved than others. The human, mouse and cow RefSeq datasets were individually analyzed for SATs. Unlike previous studies that filtered out sequences that aligned to TEs, the SAT datasets in this project were aligned to transposable elements sequences from their respective organisms. The data revealed a group of SATs that was created as the result of TEs integrating into opposing strands of coding regions of various genes creating a one-to-many relationship of sense to antisense transcripts within a particular genome. Although TEs have been often dismissed as 'junk' DNA, the generation of SATs by TE integration could potentially have added to an organism's complex regulatory networks. However, the results of this project demonstrate that many SATs created by TEs have not remained conserved and others are the result of currently active TEs. It is difficult to distinguish whether recent TE insertions have contributed to species specific gene regulation or whether there has not been enough evolutionary time in order for these regions to accumulate mutations.

Introduction

There is a growing interest in investigating potential gene regulatory mechanisms initiated by SATs as is demonstrated by many recent publications as well as public SAT databases on the Internet (Galante et al. 2007; Yin et al. 2007). The large percentage (>20%) of transcripts in the human genome that have the potential to form SATs (Chen et al. 2004) demonstrates the possibility of complex regulatory networks composed of many mRNA transcripts, including coding transcripts, that potentially bind to other mRNA transcripts and regulate their expression.

The public release of numerous sequenced genomes has created opportunities to apply a number of *in silico* techniques to identify SATs. Various methods and datasets have been used to search for SATs using expressed sequence tags (ESTs) and cDNAs, from repositories such as UniGene (Wheeler et al. 2003) and RefSeq, as well as clone datasets (Kiyosawa et al. 2003). Many researchers have concentrated specifically on identifying and classifying *cis* SATs within species (Chen et al. 2004; Chen et al. 2005; Knee and Murphy 1997; Shendure and Church 2002; Yelin et al. 2003; Wang et al. 2005; Zhang et al. 2006) while other researchers have studied the conservation of *cis* SATs across various species (Zhang et al. 2006). Each dataset and technique has identified many SATs both *cis* and *trans*, however, no common features have been identified which would allow scientists to categorize SATs with a general regulatory function. Some SATs may have gained a regulatory function while others may just be insignificant consequences of transposable element activity, segmental duplications, cross-over events or some other evolutionary event. Unlike previous studies, this project focuses on the roles of TEs in the creation of SATs.

Objectives

The objectives of this project were to structurally characterize SATs using the locus of each SABR in transcripts of human, mouse and cow to determine whether SATs are more abundant in specific regions of a transcript.

Materials and Methods

A goal of the RefSeq dataset is to provide accurate and full-length sequence data (Pruitt et al. 2005). It is composed of genomic, transcript and protein sequences retrieved from the public sequence repository GenBank (Benson et al. 2007). The RefSeq dataset was chosen for this project because it is a non-redundant, comprehensive set of gene transcript sequences for various model organisms.

The human and mouse RefSeq datasets were downloaded from NCBI on April 2, 2007. The cow RefSeq dataset was downloaded from NCBI on October 17, 2007. Only the NM_ prefixed RefSeq accessions were used because the XM_ prefixed RefSeq data set consists of computed gene models with little experimental evidence, whereas the NM_ prefixed accessions are either curated from full-length cDNA clones or created by aligning ESTs or cDNAs to a genome sequence. Alternative transcripts were maintained in the dataset. Each dataset was aligned in an all vs. all approach using the FASTA (Pearson 1988) alignment tool with an e-value of 1e-10 and the -i option to reverse complement one copy of each organism's dataset. This technique can also be accomplished using BLASTN (Li and Su 2006). A Perl script was used to remove reciprocal hits and parse the FASTA output to retrieve the alignment coordinates of SAT sequences containing ≥95% SABR complementation and an E-value of 1e-10. A 95% cutoff was set since these SATs have a greater chance of forming stable dsRNA

duplexes. The E-value was maintained at 1e-10 but the percent identity for the SABR was reduced to ≥80% for a second dataset of SATs. This would identify any SABRs that have accumulated mutations since their initiation. The genomic coordinates for the sequences involved in SATs were obtained from NCBI's GeneID coordinates.

A Perl script was used to parse the sequences of the SABR regions. These SABR sequences were aligned to transposable elements using CENSOR and the repeat libraries included with the CENSOR package. A Perl script was used to remove non-TE sequences such as satellite, simple and tandem repeat sequences, tRNA and rRNA sequences from the CENSOR repeat libraries. The libraries used were human (hum), primate (pri), mammal (mam), rodent (rod) and vertebrate (vrt). The human library was used first for human, the rodent library was used first for mouse and the mammal library was used first for cow. The CENSOR libraries were searched in the order as given on the command line: censor input_file.fasta -bprm '-filter=none' -s -nofilter -mode {sens} -lib hum -lib pri -lib mam -lib rod -lib vrt. For example, in the command above, each sequence was first aligned to the human repetitive sequence dataset. If there were no human repetitive sequences aligned, then the next repetitive element datasets in the order mammalian, vertebrate and rodent were used. If a part of a sequence aligned to a human repeat, it would be masked and no longer aligned to the sequential libraries; however, the rest of the unaligned sequence would be aligned to the libraries in the order given at the command line.

All individual SAT sequences were merged into groups based on the GeneIDs (unique identifiers for each gene locus) as provided by Entrez Gene (Maglott et al. 2007). This allowed the clustering of alternative transcripts to represent one gene locus when defining the number of *cis* and *trans* SATs. Various Perl scripts were used to get statistics

such as average SABR length and the percentage of the SABR consisting of a TE sequence.

The BioPerl Perl module was used in a Perl script to retrieve all protein sequences and gene symbols for each RefSeq accession of human and mouse that had ≥95% SABR complementation. FASTA alignments were conducted aligning all human to mouse, mouse to cow and human to cow protein sequences using an E-value of less than 0.001. The aligned sequences were scored as homologs if the alignment had greater than 30% identity across greater than 95% of both sequence lengths.

The data from Table 6 was used as a population for the components involved in the *trans* SATs for the ≥95% complementation group. The *cis* SATs were not used since they generally overlap at the same genomic locus and thus have a high percent complementation due to being on opposite complementary DNA strands. The hypergeometric distribution (Figure 3) was used to test for any overrepresented components existing in the ≥95% SABR complementation dataset which is a subset of the 80% complementary dataset.

To investigate whether older Alu subfamilies were preserved in SATs the FASTA alignment tool was used for comparing all SABRs with the younger AluY subfamily sequences and the more ancient AluJo and AluSp subfamily sequences.

$$\Pr(r \mid n, p, k) = \frac{\binom{pn}{r}\binom{(1-p)n}{k-r}}{\binom{n}{k}}$$

Figure 3.
The hypergeometric distribution. (Castillo-Davis and Hartl, 2003).

The hypergeometric distribution can be used to identify overrepresented terms from a study sample within a population. The population is represented with *n* and the sample is represented by *k*. The number of SAT genes that have a particular component is represented by *r*. The proportion of genes in the population with a particular identity is *p* (Castillo-Davis and Hartl 2003).

Results

A SAT pair consists of two distinct sequences a sense transcript and an antisense transcript. Therefore, it might be expected that 100 SAT pairs consist of 200 unique sequences if there is a one-to-one relationship among SAT sequences. However, of the 1,030 sequences comprising the 515 human SAT pairs, there were only 745 unique sequences. This resulted because specific sense sequences can form a SAT pair with several different antisense transcripts in a one-to-many relationship. This is the case for all organisms in both levels of percent complementation (Table 1). As expected, there is a one-to-very-many relationship of sense to antisense transcripts in the ≥80% SABR complementation dataset as there are relatively few unique sequences making up the thousands of SATs for each organism (Table 1).

Table 1. Number of SATs and unique sequences identified in each organism for two different percent identity datasets. These include all alternative transcripts.

	≥95% SABR complementation			≥80% SABR complementation		
	SATs	**unique sequences**	**GeneID groups**	**SATs**	**unique sequences**	**GeneID groups**
human	515	745	585	589,479	3,640	2,874
mouse	446	725	685	42,777	1,787	1,688
cow	67	128	128	5,431	517	516

A Perl script was used with the GeneID coordinates obtained from NCBI to determine the genomic locations of SATs and to cluster alternative transcripts of the same gene in order to differentiate between *cis* and *trans* SATs. The *cis* SATs were categorized into the three positional categories as defined in Chapter I (Figure 1) using the GeneID coordinates. All three organisms had a larger number of convergent *cis* SATs than the other two *cis* categories for both the ≥95% complementation and ≥80% complementation datasets (Table 2). The divergent category was the second most abundant category for all three organisms and both datasets. There was a larger number of *cis* than *trans* SATs in the ≥95% SABR complementation dataset for all organisms; however, there were significantly more *trans* SATs than *cis* SATs in the ≥80% SABR complementation datasets for each organism (Table 2). When a transposable element copies itself to another locus in the genome, it is at initially identical with its sequence of origin. If an important new function emerges from the transposition event, then is expected that the sequence would maintain a high level of conservation over evolutionary time. The lack of sequence conservation in the SABRs for *trans* SATs suggests that there these pairs do not confer a critical function such as gene regulation. However, *trans* SATs that have maintained high percent complementation and are the result of Alu sequences that proliferated more than 60 million year ago (MYA) would suggest a strongly conserved biological mechanism. The *cis* SATs, however, are the predominant category in the ≥95% SABR complementation dataset. The reason for a high level of conservation of *cis* SATs is not conservation, but the fact that these transcripts consist of opposing strands at the chromosomal DNA level.

Table 2. Number of *cis* and *trans* SATs in human, mouse and cow for the ≥95% complementation and ≥80% complementation datasets for GeneID groups. Counts for all three organisms exclude alternative transcripts but demonstrate the one-to-many relationship between SATs for two levels of SABR complementation (comp.).

	Human	Mouse	Cow
≥95% SABR comp.			
Convergent (*cis*)	191	202	15
Divergent (*cis*)	50	51	0
Containment (*cis*)	9	12	0
Total *cis* SATs	250	265	15
Total *trans* SATs	84	130	52
Total SATs	334	395	67
Total SABRs with TE	74	137	19
≥80% SABR comp.			
Convergent *(cis*)	205	222	15
Divergent (*cis*)	60	57	2
Containment (*cis*)	12	14	0
Total *cis* SATs	277	293	17
Total *trans* SATs	390,386	37,232	5,391
Total SATs	390,639	37,525	5,408
Total SABRs with TE	380,452	37,167	5,393

The average SABR lengths for the ≥80% SABR complementation dataset were shorter than those in the ≥95% SABR complementation dataset for all three organisms (Table 3) perhaps due to the increased number of short SINE sequences in the ≥80% SABR complementation dataset.

Table 3. Statistics of SABR lengths for human, mouse and cow.

	Average	Min	Max	Median
human ≥95% complementation	344	60	3,561	213
human ≥80% complementation	274	60	3,561	299
mouse ≥95% complementation	330	62	2,304	192
mouse ≥80% complementation	167	62	2,304	163
cow ≥95% complementation	238	63	1,111	158
cow ≥80% complementation	172	63	1,111	135

To access the extent of TEs in all sequences, the RefSeq sequences from each organism were aligned to the transposable element sequences provided in the CENSOR (Jurka 2005) repeat libraries. A Perl script was used to parse the results and calculate the percentage of TEs that aligned to the coding strand and the percentage of TEs that aligned to the template strand. These alignment results were also later used in the GO analysis section in Chapter II. There was a consistently higher percentage of TEs incorporated into the template strand than the coding strand for all three organisms (Table 4). This agrees with the results from previous researchers (Makalowski et al. 1994; Lorenc and Makalowski 2003; Almeida et al. 2007) although the biological significance is not quite clear. However, it has been suggested that certain TE sequences are rich in canonical splice junctions that can alter the structure of the gene if it integrates into the coding strand (Sela et al. 2007). If the TE incorporates into the template strand, the number of introduced canonical splice junctions is reduced.

Table 4. Percentage of transposable elements aligning to RefSeq sequences in the coding and template strands.

	% Coding	% Template
human	46.0	54.0
mouse	46.9	53.1
cow	45.6	54.4

The SABRs were distributed across relatively the same regions of the transcripts in human and mouse, with the 3' UTR having the most localized SABRs (Table 5). Duplicate arrangements for each specific gene SABR were counted as a single arrangement (Table 5). The coding sequence (CDS) contained few localized SABRs and

the 3' UTR contained the most for the ≥95% SABR complementation dataset (Table 5). This is expected since the CDS displays more conservation across human and mouse than the UTRs (Makalowski et al. 1996). Additionally, the conservation of the coding sequence across species may indicate that it is less likely to contain a SABR due to a duplication or transposition event that would disrupt the coding portion of the transcript.

Table 5. Localization of SABRs for human and mouse *trans* SATs. Total for all *trans* SATs having ≥80% SABR complementation (≥80% comp) and ≥95% SABR complementation (≥95% comp). Duplicate arrangements for each specific gene SABR were counted as a single arrangement.

	Human		Mouse	
Arrangement	**≥80% comp**	**≥95% comp**	**≥80% comp**	**≥95% comp**
CDS – CDS	23	0	2	0
5' UTR - 5' UTR	132	14	19	4
5' UTR – START	41	10*	4	0
5' UTR – CDS	133	4*	20	2
5' UTR – STOP	417	3	42	0
5' UTR - 3' UTR	1,915	29	610	28
START – START	6	2	2	2
START - CDS	19	0	5	4*
START - STOP	20	0	0	0
START - 3' UTR	860	0	44	3
CDS - STOP	54	8*	11	0
CDS - 3' UTR	1,713	4	348	6
STOP – STOP	47	5*	8	2
STOP - 3' UTR	1,702	17	441	14
3' UTR - 3' UTR	2,123	62*	1,023	100*
TOTAL	**9,205**	**158**	**2,579**	**165**

* $p < 0.001$

Conserved SABR Components: Conserved Function or Recent TE Event?

The data reveal that there are a number of conserved 3' UTR SAT SABRs that may serve a biological function. However, since TE activity has played a role in the formation of many SATs, these conserved SABRs could be the result of not only TE activity but recent transposition events from active TEs that have not had much time for sequence divergence. Currently we have an understanding of the periods of transposition

events of certain Alu elements. The AluJ family is known to have proliferated 60 million years ago; AluS proliferated between 60 and 20 million years ago; AluY began transposition activity approximately 20 million years ago. Some AluY subfamilies are even younger than 20 million years and include actively retrotransposed elements in the human population (Jurka et al. 2002). Other active elements in the human genome include L1 and SVA (Mills et al. 2007). The SABRs of *trans* SATs with a high percent complementation in the SABR may indicate either little evolutionary time for sequence divergence after a recent TE event, or a TE event that has resulted in a critical biological role.

The hypergeometric distribution (Castillo-Davis and Hartl, 2003) was used to test for significant enrichment in the *trans* SATs for each component. The 3' UTR – 3' UTR SAT arrangement was overrepresented in both human and mouse datasets. Further investigation found that in the human dataset, the AluY transposable element which is currently an actively transposing element was significantly enriched in the ≥95% SABR complementation group using the hypergeometric distribution analysis ($p < 0.01$). There were no significant enrichments for a particular mouse transposable element in the ≥95% SABR complementation group when compared to the ≥80% SABR complementation dataset.

Human SATs with ≥95% SABR Complementation

There were 515 SAT pairs in the dataset with ≥95% SABR complementation. There were only 745 unique sequences of the 1,028 sequences involved in the 514 SAT pairs indicating a one-to-many relationship between sense and antisense transcripts.

The most common TE in the human dataset was the Alu sequence which totaled

45 of the 70 SABRs containing TEs from the 515 SAT SABRs. This is expected since Alu is the most abundant transposable element in the human genome, with over a million copies. The Alu sequence was also responsible for an average of 99% of the SABR alignment lengths with an average of only 45% of its sequence (Table 6). This suggests the potential of the Alu sequence to introduce canonical splice sites which would leave some of the Alu sequence in the intron regions. The Alu sequence in the CENSOR dataset was 334 nucleotides long, but generated an average SABR length of 141 nucleotides. This is most likely due to alternative splice sites introduced by the Alu sequence or new exons created with only a portion of the Alu sequence. TEs such as MER85 averaged 100% of the TE sequence found in a fraction of the SABR sequence while other TEs such as THE1A covered an average of 100% of the SABR sequence with only a fraction of the TE sequence.

Mouse SATs with ≥95% SABR Complementation

There were a total of 446 SAT pairs consisting of 725 unique sequences for the dataset with ≥95% SABR complementation. This resulted in a one-to-many relationship between sense and antisense transcripts similar to the human SAT results. As an example, the mouse *trans* SAT sense sequence NM_001081387.1 has five antisense transcripts each binding to the sense transcript at relatively the same coordinates (Table7). Although the B1_Mus2 SINE has a length of 147 nucleotides, the SABR length is only 105 nucleotides. The B1_Mus2 SINE sequence spans the entire length of the SABR sequence, but only part of the B1 Mus2 SINE sequence is present in the SABR either due to alternative splice sites introduced by the B1_Mus2 sequence or due to mutations, insertions and deletions at the ends of the SABR that have accumulated since the transposition event occurred.

Table 6. Average percent coverage of human SABRs (≥95% SABR complementation) by transposable elements. The transposable elements (TEs) are classified as SINEs (S), LINEs (L), DNA transposons (D) and endogenous retroviruses (E).

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
ALU (S)	141	99	45	45	0	45
ERV1-2-I_XT (E)	1,981	3	1	1	1	0
ERV2_MD_I (E)	204	42	1	1	1	0
HERV70_I (E)	834	23	3	1	1	0
L1 (L)	96	100	2	1	0	1
L1A_OC (L)	1,136	7	1	1	1	0
L1HS (L)	244	99	27	1	0	1
L1PA3 (L)	122	100	14	1	0	1
L2A (L)	788	47	3	2	2	0
MER2 (D)	619	28	60	1	0	1
MER2B (D)	438	21	23	4	0	4
MER30 (D)	593	6	15	1	1	0
MER34B (E)	1,545	23	66	1	1	0
MER5A1 (E)	3,561	2	62	1	1	0
MER85 (D)	1,461	9	100	1	1	0
MIR (S)	1,721	34	54	3	3	0
MIRb (S)	360	22	29	1	1	0
RNLTR17_I (E)	238	71	2	1	1	0
SVA (S)	97	99	6	1	0	1
THE1A (E)	68	100	19	1	0	1
Total	**392**	**78**	**39**	**70**	**15**	**55**
Total for each TE type						
SINE	234	94	44	50	4	46
LINE	400	67	8	6	3	3
DNA transposon	631	19	41	7	2	5
Endo. Retrovirus	1,204	38	22	7	6	1

Table 7. One-to-many relationship of *trans* mouse SATs. One sense transcript with five antisense transcripts having the sense-antisense binding region (SABR) region relatively in the same locus on the sense transcript aligning to the B1_Mus2 SINE (strand: c=coding, t=template, Chr=chromosome).

Sense transcript	SABR start	SABR end	Chr	SINE strand	Antisense transcript	Chr	SINE strand
NM_001081387.1	2521	2627	2	t	NM_026282.2	9	c
NM_001081387.1	2521	2626	2	t	NM_018804.3	3	c
NM_001081387.1	2521	2626	2	t	NM_198107.1	10	c
NM_001081387.1	2521	2626	2	t	NM_021534.2	2	c
NM_001081387.1	2521	2628	2	t	NM_198417.1	7	c

In many cases with the dataset having ≥95% SABR complementation in the SABRs, TEs spanned 95-100% of the entire length of the SABR sequence especially for

the most common TEs aligning to SABRs such as B1Mus1, B1_Mm and B2_Mm1. The TEs that cover 98 – 99% of the SABR could be due to mutations at the sequence ends after insertion or perhaps due to the difficulties sequence alignment programs have in aligning the ends of matches. Additionally, TEs covering a smaller percentage of the SABR suggest an introduction of a canonical splice sites by the TE sequence. The B1_Mus1 element is found in the highest number of SABRs and covers an average of 99% of the entire SABR length (Table 8).

Mouse SATs with ≥80% SABR Complementation

Although SABRs with lower than 95% complementarity are not likely to form RNA duplexes or confer a biological function, a dataset with a relaxed complementarity criterion would represent a set in which TE sequence divergence has occurred since the time of TE insertion. Differences between a set with a relaxed criterion and a 95% complementarity set could reveal biologically significant properties in the 95% complementarity set. To create a set for comparison, the percent complementation was relaxed to ≥80%. This resulted in 42,777 SAT pairs formed by 1,787 unique sequences. Of these SAT pairs, 42,545 had TE sequence in their SABR demonstrating the influence of TEs on the creation of SATs at this particular threshold. For the ≥80% SABR complementation dataset, the B1_Mus1 element was the most abundant TE in the SABRs with an average percentage of 89% of its sequence aligning in SABRs (Appendix C). This may indicate that mutations at the ends of the TE sequence caused misalignments that were not reported in the FASTA results or it may be due to the introduction of new splice sites.

Table 8. Average percent coverage of mouse SABRs (≥95% SABR complementation) by transposable elements. The transposable elements (TEs) are classified as SINEs (S), LINEs (L), DNA transposons (D) and endogenous retroviruses (E).

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
B1 (S)	243	81	61	5	1	4
B1F (S)	727	15	71	3	3	0
B1_Mm (S)	114	98	75	13	0	13
B1_Mur1 (S)	602	26	94	1	1	0
B1_Mur2 (S)	1,002	12	79	1	1	0
B1_Mur4 (S)	367	38	94	1	1	0
B1_Mus1 (S)	116	99	77	35	0	35
B1_Mus2 (S)	144	93	80	12	1	11
B1_Rn (S)	371	69	82	3	1	2
B2 (S)	67	100	32	1	0	1
B2_Mm1a (S)	145	99	74	14	0	14
B2_Mm1t (S)	280	82	99	3	1	2
B2_Rat2 (S)	631	26	89	1	0	1
B2_Rat4 (S)	1,534	12	99	1	1	0
B2_Rn2 (S)	77	99	41	4	0	4
B3 (S)	766	22	61	2	2	0
B3A (S)	717	31	77	2	2	0
BC1_Ma (S)	2,304	2	29	1	1	0
Harbinger-2_XT (D)	249	14	1	1	1	0
IAPEZI (E)	136	100	2	1	0	1
ID_Rn1 (S)	938	6	55	1	1	0
L1MC4 (L)	270	21	2	1	1	0
L1_MM (L)	250	99	4	4	0	4
L2A (L)	105	80	3	1	1	0
LTRX_ME (E)	388	12	8	1	1	0
LX (L)	94	100	9	2	0	2
MEN (S)	441	35	55	1	1	0
MER20 (D)	1,864	5	51	1	1	0
MER21C (E)	687	7	5	1	1	0
MTA (E)	141	100	36	2	0	2
MTAI (E)	106	100	10	1	0	1
MTA_Mm_LTR (E)	140	100	36	1	0	1
MTEb_LTR (E)	346	25	24	1	1	0
MYS1_PL (E)	126	25	1	1	1	0
ORR1D1_LTR (E)	976	35	94	1	1	0
PB1D10 (S)	2,038	5	85	1	1	0
PB1D9 (S)	1,066	10	80	2	2	0
RLTR10C (E)	114	100	26	1	0	1
RLTR16 (E)	2,304	4	18	1	1	0
RSINE1 (S)	919	13	81	1	1	0
Total	**309**	**79**	**66**	**131**	**32**	**99**
Total for each TE type						
SINE	288	81	75	109	22	87
LINE	195	87	5	8	2	6
DNA transposon	1,057	10	26	2	2	0
Endo. Retrovirus	467	59	25	12	6	6

Cow SATs with ≥95% SABR Complementation

There were fewer SATs revealed in the analysis of cow SATs probably due to the small data set available. There were 67 SAT pairs comprised of 128 unique sequences.

Although the cow genome is not near the level of annotation as the human and mouse genomes, this study detected similar patterns in the number of SATs and unique sequences at different percent complementation in the SABRs. Cow SAT SABRs also contained more SINEs than any other transposable element for both complementation datasets (Table 9 and Appendix D).

Table 9. Average percent coverage of cow SABRs (≥95% complementation) by transposable elements. The transposable elements (TEs) are classified as SINEs (S), LINEs (L) and endogenous retroviruses (E).

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
ART2A (S)	218	100	40	2	0	2
BOVA2 (S)	134	95	48	9	0	9
BOVB (S)	131	98	4	1	0	1
L1_BT (L)	165	99	14	3	0	3
L1P_MA2 (L)	124	99	2	1	0	1
RNLTR17_I (E)	242	33	1	1	0	1
Total	**153**	**93**	**35**	**17**	**0**	**17**
Total for each TE type						
SINE	146	96	44	12	0	12
LINE	155	99	11	4	0	4
Endo. Retrovirus	242	33	1	1	0	1

Conserved SATs in Human, Mouse and Cow Orthologs

Conservation of SATs can provide a clue to understanding SAT function and whether conservation of specific SAT loci reflects functional importance of SATs. SATs that are maintained over evolutionary time demonstrate that they may be essential for the organism's survival and maintenance.

Only *cis* oriented SATs were found to be conserved across human, mouse and cow genomes (Table 10). This may demonstrate that if *trans* SATs have some function within a particular organism, it may not be shared across species. This concept is further investigated in Chapter III. The fewer conserved cow sequences reflect the lower level of annotation of the cow genome versus the human and mouse.

Table 10. Conservation and arrangement of *cis* SATs with ≥95% SABR complementation for human, mouse and cow. Mouse gene in parenthesis when the name differs from human ortholog.

Species	Sense	Antisense	*cis* orientation
Hs. Mm.	B3GAT2	SMAP1	convergent
Hs. Mm.	HTRA2	LOXL3	convergent
Hs. Mm.	DNAJC14	ORMDL2	convergent
Hs. Mm.	PRR3	GNL1	divergent
Hs. Mm.	PHYHIPL	FAM13C1	convergent
Hs. Mm.	SLC3A1	PREPL	convergent
Hs. Mm.	KRR1	GLIPR1	convergent
Hs. Mm.	C20orf132	RPN2	divergent
Hs. Mm.	DNAH11	CDCA7L	convergent
Hs. Mm.	TMEM140	C7orf49	convergent
Hs. Mm.	MGC4655	C16orf70	convergent
Hs. Mm.	C12orf48	PMCH	convergent
Hs. Mm.	DKFZp667G2110 (BC043118)	MINA	convergent
Hs. Mm.	ATP5F1	WDR77	divergent
Hs. Mm.	COG1	FAM104A (D11Wsu99e)	convergent
Hs. Mm.	VRK2	FANCL	convergent
Hs. Mm. Bt.	POLR3H	ACO2	convergent
Hs. Mm.	DMTF1	C7orf23	convergent
Hs. Mm.	C20orf132 (4922505G16Rik)	RPN2	divergent
Hs. Mm.	YPEL1	PPIL2	convergent
Hs. Mm.	RPE	FLJ23861 (LOC273425)	convergent
Hs. Mm.	LRRC59	EME1	convergent
Hs. Mm.	MORG1	DHPS	convergent
Hs. Mm.	CRBN	TRNT1	convergent
Hs. Mm.	PKMYT1	PAQR4	convergent
Hs. Mm.	THOC6	HCFC1R1	divergent
Hs. Mm.	RBM13	C8orf41 (BC019943)	convergent
Hs. Mm.	DTX3L	PARP9	divergent
Hs. Mm.	KRI1	ATG4D	convergent
Hs. Mm.	ANGPTL6	FLJ11286 (A230050P20Rik)	convergent
Hs. Mm.	FAM151A (BC026682)	ACOT11	containment
Hs. Mm.	CNTD1	BECN1	convergent
Hs. Mm.	TMEM85	SLC12A6	convergent
Hs. Mm.	CCDC111	MLF1IP	convergent
Hs. Mm.	LRRC39	CCDC76	convergent
Hs. Mm.	C16orf58	SLC5A2	convergent
Hs. Mm.	PIGF	RHOQ	convergent
Hs. Mm.	RAF1	MKRN2	convergent
Hs. Mm.	NOL3	LOC653319 (4931428F04Rik)	convergent
Hs. Mm.	WDR6	DALRD3	convergent
Hs. Mm.	POLR2B	IGFBP7	convergent
Hs. Mm.	GBA2	CREB3	convergent
Hs. Mm.	CTSA	NEURL2	divergent
Hs. Mm.	FANCI	POLG	convergent
Hs. Mm.	ABI1	PDSS1	convergent
Hs. Mm.	PPP1CA	RAD9A	convergent
Hs. Mm.	LRIG1	SLC25A26	convergent
Hs. Mm.	TSR1	SRR	convergent
Hs. Mm.	FCHSD1	RELL2	convergent
Hs. Mm. Bt.	RAD9A	PPP1CA	convergent
Hs. Mm.	C14orf45 (2900006K08Rik)	ALDH6A1	convergent
Hs. Mm.	ACAT2	TCP1	convergent

Discussion

SATs

The dataset of ≥80% SABR complementation in all three genomes having thousands of SATs composed of relatively few unique sequences shows the one-to-many relationship of sense to antisense transcripts. However, the detection of so many SATs at this threshold may show that the reduction in complementation over time may be due to the SABRs of these SATs having no conserved function. The relatively similar coordinates of the SABR in each transcript involved in multiple SAT relationships points to a similar mechanism of origin. One of these events is the process of TEs integrating into the exons on the coding strand of some genes and the template strand of the exons of other genes creating a one-to-many relationship of sense to antisense transcripts. There are more SABRs localized in the 3' UTR of the transcripts for all three organisms. This suggests the ability of the 3' UTR to extend by addition of TE sequence while still maintaining transcript coding functionality. The CDS has fewer SABRs since transposon insertions in the CDS often disrupt coding sequence, which leads to non-functioning proteins. The 5' UTR has more SABRs localized than the CDS but not as many as the 3' UTR. Perhaps this is because the 5' UTR is generally much shorter than the 3' UTR and the possibility of 5' UTR regulatory region disruption.

If there were a conserved regulatory relationship between *trans* SATs, the sequence conservation in the SABR would most likely have been higher and many *trans* SATs would have been identified in the dataset of ≥95% SABR complementation. While the relatively few *trans* SATs in the ≥95% SABR complementation group may have a biological function in individual species, the search for homologous *trans* SATs across species does not support conserved function, as no homologous *trans* SATs were

identified across species.

TEs

Using the two datasets with ≥80 and ≥95% SABR complementation thresholds allowed a better assessment of the evolutionary fate of SATs created through transposition or other duplication events. Transposable elements were found in various sequences of all three genomes. The fact that the majority of *trans* SATs were found in the ≥80% SABR complementation dataset, but not the ≥95% dataset suggests that the majority of *trans* SATs do not have a biological function. For a transposable element involved in SAT SABRs to create or become part of a regulatory process, the transposition events in each gene would have had to occurred at a relatively close temporal interval. Conservation in the SABRs of a few *trans* SATs may be due to either the necessity to form a RNA duplex as part of a critical biological function or due to recently created SATs that have not had enough time to diverge. The high numbers of *trans* SATs that have TE sequence in the SABRs (Table 2) and the biased localization of the SABR to the 3' UTR (Appendix A) demonstrate the influence TEs have on the origin of the majority of the SATs in the ≥80% SABR complementation dataset and also suggest that TEs may contribute to the observation that 3' UTRs are on average longer than 5' UTRs.

RNA Editing: An Overlooked Gene Regulation Mechanisms of SATs?

SATs have been demonstrated as having gene regulatory properties both for *trans* (Smith et al. 1988) and *cis* (Prescott and Proudfoot 2002; Crampton et al. 2006) arrangements. A possible mechanism of SAT regulation involves RNA editing. SATs

that produce a RNA duplex in the coding region of a transcript are targets for RNA editing. SAT duplexes of the *4f-rnp* and a transcriptional read-through of *sas-10* genes in *Drosophila melanogaster* were suggested in providing the dsRNA region for RNA editing enzymes that edit A-I which lead to the observed down-regulation of *4f-rnp* by *sas-10* (Peters et al. 2003).

This research project demonstrated that transposable elements have played a role in the origin of SATs. The average SABR length exceeded 100 bp which can lead to an increased occurrence of A-I editing events called hyper-editing which occurs more frequently in duplexes longer than 100 bp (DeCerbo and Carmichael 2005). Hyper-editing can edit up to 50% of adenosines to inosines (Nishikura et al. 1991) as was found in a sense-antisense region of two RNA transcripts of *Xenopus* oocytes (Kimelman and Kirschner 1989).

Another result of A-I editing is the removal of premature stop codons as was demonstrated *in vitro* (Woolf et al. 1995). If premature stop codons are not removed through A-I editing, the transcripts are prime candidates for degradation by the NMD regulatory mechanism since the splice junction complexes remain on the transcript after the pioneering round of translation.

The creation of stop codons is not possible through an A-I editing process since the only way to edit a codon to a stop codon is to begin with a stop codon. However, premature stop codons could be created after a less common C-U editing process by altering codons such as the codon for glutamine (CAA) but this process was found to be dependent on specific genes and is not as widespread as A-I editing. An example of C-U editing is apolipoprotein B (apoB) in which a glutamine (CAA) is changed to a stop codon (UAA) by the RNA editing enzymes resulting in apoB48 a smaller isoform of

apoB (Bostrom et al. 1989). Additional editing events of T-C were observed in cDNA clones; however, the A-I and T-C conversions were not observed in the same clone.

Component Analysis of trans-SATs

The *cis* SATs are expected to be conserved since they overlap at the same locus. There is little difference between the numbers of conserved *cis* SATs in the ≥95% SABR complementation group compared to the ≥80% SABR complementation group. The few *cis* SATs in the ≥80% SABR complementation group but not in the ≥95% group may be due to sequencing errors or polymorphisms.

The fact that 13/15 of the TEs found in the 3' UTR of the conserved human *trans* SATs were the younger AluY group indicates that these were the result of recent transposition events (<20 million years) that have not had time to diverge below the ≥95% SABR complementation threshold. There were no significant alignments between *trans* SATs and two older Alu families (AluSp and AluJo) indicating that AluSp and AluJo proliferation in the human genome did not have a significant impact on conserved *trans* SAT formation.

CHAPTER III

IN SILICO FUNCTIONAL ANALYSIS OF SENSE-ANTISENSE TRANSCRIPTS

Overview

The general function of sense-antisense transcripts (SATs) is still under investigation. A functional Gene Ontology (GO) analysis of the gene products of SATs may provide insights into the general function of SATs or the effects of sense-antisense binding regions (SABRs) on protein folding properties. The goals of this project were to perform a computational functional analysis of groups of SATs in human, mouse and cow and to use confirmed structural sequences to determine whether the SABRs of SATs created by transposable elements disrupt the ability of a protein to properly fold. RNA editing in SABRs was also investigated to determine if these double stranded RNA complexes have a significant effect on protein folding capacities. The GO analysis showed enrichments for the functional category of 'DNA repair' and the component category 'cytoplasm'. A GO Slim analysis showed enrichments for 'catalytic activity' which is the parent ontology of 'DNA repair'. A GO analysis can give insights to possible enrichments of molecular functions, biological processes or cellular components for a specific set of genes; however, the results can often be broadly interpreted even if there are no highly significant enrichments.

Introduction

GO analysis is an *in silico* approach to investigating functional significance of groups of genes, and will be more informative as total genome annotations improve. Previous researchers used GO analysis of *trans* SATs in *Arabidopsis thaliana* and found

over-represented groups for a number of GO attributes such as transferase activity, protein binding, and protein modification (Wang et al. 2006). GO Slim terms are terms that are mapped to upper level terms of specific lower level Gene Ontology categories using maps such as the one provided by the European Bioinformatics Institute (EBI; see Appendix E). A GO Slim functional analysis of five fungal species, *Saccharomyces cerevisiae*, *Ashbya gossypii*, *Saccharomyces pombe*, *Neurospora crassa* and *Neurospora cuniculi* SATs using the GO Slim terms provided by the *Saccharomyces* Genome Database found the common GO Slim term 'nucleobase, nucleoside, nucleotide and nucleic acid metabolism' in *S. cerevisiae*, *S. pombe and N. crassa* (Steigele and Nieselt 2005). Additionally they found that a high number of SAT transcripts have nuclear localization in four of the five fungal species studied.

Objectives

A goal of this project was to investigate functional categories of protein products encoded by the sense strand of SATs using the GO. Tests were performed to identify significant enrichments of specific GO terms within SATs and between SAT and non-SAT sequences. Another goal was to investigate the effect SABRs and TEs on the production of folded proteins. The protein-coding portions of SAT sequences were compared to nucleotide sequences of proteins with known structures to determine whether the SABRs of SATs potentially affect the folding properties of proteins as a result of processes that change the nucleotide sequence such as TE insertions and RNA editing. If the SABR or TE portion of the protein is found in a homolog in a protein structure database, one can infer that the SABR or TE does not disrupt protein folding.

Materials and Methods

Gene Ontology Analysis

The Gene Ontology is organized into three principle ontologies: molecular function (F), biological process (P) and cellular component (C). The molecular function ontology describes activities occurring at the molecular level such as catalytic and binding (The Gene Ontology Consortium 2000). The biological process describes the contribution of the gene product to a biological objective, and the cellular component describes where a gene product is found in the cell (The Gene Ontology Consortium 2000). A gene product may be annotated with one or more terms from one or more of the three principle ontologies.

GO terms available from Ensembl Biomart (http://www.biomart.org/biomart/martview/) were used in conjunction with GeneMerge to identify overrepresented GO Slim terms among the ≥95% SABR complementation group of SATs. A BioPerl Perl module was used to retrieve protein sequences for the list of RefSeq nucleotide accessions from the RefSeq repository. Since the Ensembl GO terms are more current than RefSeq, the RefSeq protein sequences were aligned using FASTA to the Ensembl protein. GO terms were assigned to each RefSeq sequence that aligned with 100% identity to an Ensembl protein sequences annotated with at least one GO term. After the GO terms were assigned to the RefSeq sequences, the July 24, 2007 version of the GO Slim map file (Appendix E) provided by EBI was used to map each full GO term to its corresponding GO Slim term. This step is necessary because the full GO terms do not report all upper level GO terms for each accession and some over-represented upper-level GO term may be identified using this approach. Some full GO terms had multiple GO Slim terms since the Gene Ontology is not a hierarchical tree.

Perl scripts were used to divide GO Slim terms into the appropriate GO category (F, P or C) using the August 8, 2007 version of the Gene Ontology (Appendix F) and to remove redundancy of GO Slim terms for individual sequences. Each accession could have multiple GO Slim terms for each category as well as terms from any combination of the three categories. The final GO Slim enrichment analysis was done using GeneMerge on each of the three GO category files which was more sensitive than performing the analysis on all three ontology groups together. The RefSeq sequences that had no alignments to the Ensembl dataset or aligned but had no GO term assigned were not used in this analysis.

In order to determine which SABRs had TE sequence, the full set of RefSeq sequences and the specific SABR sequences were aligned to a repetitive element database provided by RepBase using the CENSOR repeat screening software.

The Gene Ontology in conjunction with GeneMerge was used to identify any overrepresented GO terms within the ≥95% SABR complementation dataset in various analyses using the RefSeq sequences for each organism as a population and various study datasets as explained in Table 11.

Table 11. Population and study datasets for the Full GO and GO Slim analyses of the ≥95% SABR complementation dataset.

Population	Study Dataset	Results
Full GO		
All RefSeq sequences	a) All SATs	Table 14
	b) All RefSeq aligning to TE	Table 15
GO Slim		
All RefSeq sequences	a) All SATs	Table 16
	b) All SATs with TE in SABR	Table 17
	c) All SATs without TE in SABR	Table 18
	d) All RefSeq aligning to TE	Table 19

The population used in all tests was all the NM prefixed RefSeq mRNAs that aligned to GO defined Ensembl transcripts. Study samples were always a subset of the population. Tests were performed to determine whether each study set has a different distribution of GO terms compared to the population set. The first analysis compared all SATs to RefSeq sequences to determine if SATs in general were enriched for specific Full GO terms. Another Full GO analysis was to investigate if RefSeq sequences containing TEs were enriched for specific GO terms compared to all RefSeq sequences.

GO Slim tests were done to increase the level of statistical power. The first GO slim test compared all SATs to all RefSeq sequences. The second GO Slim analysis compared just SATs with TEs in the SABRs to all SATs which would demonstrate that any enrichment in this group could be contributed to SABRs formed by TEs. A third GO Slim study compared SATs without TEs in the SABR to all SATs. The final GO Slim analysis compared RefSeq sequences that aligned to TEs to all RefSeq sequences in order to understand if TEs were found in genes with particular functions.

GeneMerge uses the hypergeometric distribution to identify any overrepresented GO terms in the population and incorporates Bonferroni correction for testing more than one hypothesis with a single dataset. For *n* independent hypotheses, the significance level that should be used is 1/*n* times as compared to a single hypothesis test. The Bonferroni values of less than 0.5 were recorded since the Bonferroni adjustment may not maintain statistical power for small populations (Leon 2004).

Effects of SABRs on the Production of a Folded Protein

Often sequences that align to TEs are assumed non functional and dismissed from gene predictions and subsequently gene annotations and functional analysis (Curwen et al. 2004; Li et al. 2006). To test if SABRs or transposable elements are less likely to

produce functional proteins, all the structurally defined proteins sequences that also have a defined function were downloaded from the PDB website on November 28, 2007. The PDB structure sequence dataset contains both protein sequences of proteins with solved structures as well as nucleotide sequences of ribosomal and other RNAs. Only the 103,538 confirmed protein structure sequences from the PDB dataset were used, although the sequences were somewhat redundant. The PDB sequences were compared to the RefSeq nucleotide dataset using BLASTX with an e-value of 1e-100. The alignments were reviewed best significant hits were used since the e-value cutoff was stringent and the PDB sequence consistently was annotated with the same function as the RefSeq accession even though some of the alignments were as low as 70% identity. The functional descriptions for the RefSeq accessions and PDB accessions for all significant alignments were manually reviewed to confirm the match. A Perl script was used to evaluate sequence alignment coordinates to determine if any portion of the PDB structure aligned to any part of the RefSeq sequence that either formed the SABR or aligned to a transposable element using coordinates from the SAT FASTA alignment search and the CENSOR TE alignment. A SABR or TE region aligning to a PDB protein structure sequence will provide support that the SABR or TE did not adversely affect the protein structure or function at least for that particular gene transcript. However, this analysis method cannot provide support that a SABR or TE sequence has adversely affected the functionality of a folded protein.

The population dataset for these tests were all RefSeq sequences aligning to a PDB confirmed protein structure sequence and the study datasets were 1) the SATs from the population that have a SABR overlapping a CDS that aligns to a PDB structure sequence and 2) all RefSeq sequences that have a TE region that aligns to a PDB

structure sequence.

Results

Gene Ontology Analysis Results

An initial GO term enrichment analysis was conducted using GeneMerge on the full GO terms as provided by the Ensembl BioMart project. There were 9,285 and 10,288 human and mouse sequences respectively that contained GO terms. Both human and mouse analyses indicated enrichments for 'DNA repair' in the full GO term analyses. Additionally in the full GO analyses, the human study set was enriched for 'nuclease activity' and mouse study set was enriched for 'response to DNA damage stimulus' and 'citrate metabolic process' (Table 12). Three of the 11 study fraction sequences with the 'DNA repair' GO identifier in human and the 14 study fraction sequences in mouse are homologs of the two species. A GO analysis conducted by previous researchers for *cis* SATs in human also found the GO term 'DNA repair' enriched in the sample dataset (Lehner et al. 2002). GO analysis results of SATs such as the significant e-score for 'citrate metabolic process' found in mouse could be misrepresented since an estimated >20% of all human transcripts have the potential to form SATs and all three genes that are GO annotated as 'citrate metabolic process' could be contained within this 20%.

Table 12. Full GO term enrichment analysis for human and mouse SATs with ≥95% SABR complementation.

Species	GO identifier	Population fraction	Study fraction	Bonferroni e-score	GO term
human	GO:0004518	11/9285	5/261	0	nuclease activity
human	GO:0006281	91/9285	11/261	0.03	DNA repair
mouse	GO:0006974	103/10288	14/333	0	response to DNA damage stimulus
mouse	GO:0006281	115/10288	14/333	0.01	DNA repair
mouse	GO:0006101	3/10288	3/333	0.02	citrate metabolic process

The Full GO term enrichment for all RefSeq sequences aligning to some portion of a transposable element sequence revealed various GO enrichments (Table 13). The gene products of these transcripts were localized in the nucleus which also increased the significance level of 'intracellular' of which 'nucleus' is a lower level component. Also there was an overrepresented set of gene products that functioned in 'metal ion binding' and its lower level molecular function 'zinc ion binding'. There was also a significant enrichment for gene products involved in transcription and its lower level process 'regulation of transcription, DNA-dependent'.

Table 13. Full GO term enrichment analysis for human and mouse RefSeq sequences containing transposable element sequence in any part of the RefSeq sequence.

Species	GO identifier	Population fraction	Study fraction	Bonferroni e-score	GO term
human	GO:0003676	235/7165	125/2895	0.078	nucleic acid binding
human	GO:0005634	1712/7165	785/2895	1.852e-04	nucleus
human	GO:0008270	937/7165	476/2895	6.819e-09	zinc ion binding
human	GO:0005622	793/7165	404/2895	2.618e-07	intracellular
human	GO:0046872	907/7165	444/2895	3.120e-05	metal ion binding
human	GO:0006350	581/7165	291/2895	0.001	transcription
mouse	GO:0003676	367/6750	157/2263	0.160	nucleic acid binding
mouse	GO:0006355	620/6750	267/2263	2.314e-04	regulation of transcription, DNA-dependent
mouse	GO:0008270	635/6750	287/2263	1.617e-07	zinc ion binding
mouse	GO:0005634	1221/6750	494/2263	2.187e-05	nucleus
mouse	GO:0005622	567/6750	237/2263	0.019	intracellular
mouse	GO:0046872	742/6750	310/2263	7.058e-04	metal ion binding

GO Slim

From the original 23,663, 19,724 and 9,538 RefSeq sequences in the human, mouse and cow datasets respectively, there were 7,165, 6,735 and 1,757 sequences respectively that matched an Ensembl sequence with assigned GO terms that had GO

Slim terms assigned in the GO Slim map file (Appendix E). These were used as the overall population for the analysis from which the sample datasets were derived after dividing into the three GO categories. Although there are fewer *Bos taurus* sequences with GO terms as compared to the more annotated human and mouse genomes, the significant enrichment categories in *Bos taurus* were similar to mouse and human.

GO Slim: All SAT Sequences: F: Molecular Function

The most significant enriched molecular function (F) GO Slim terms shared by human and mouse SATs were 'catalytic activity' and 'hydrolase activity' (Table 14). Cow did not have sufficient data to complete a GeneMerge analysis for the molecular function and cellular component GO categories. Hydrolase is defined as "the catalysis of the hydrolysis of various bonds" (http://geneontology.org). Although reverse transcriptase which is encoded by an open reading frame in autonomous LINEs is considered a catalytic process, it is unlikely that the catalytic activity enrichment is a result of transposons since the majority of TEs in SAT SABRs are SINEs which lack coding potential for transposition enzymes.

P: Biological Process

The most significant biological process was identified as 'metabolic process'. Other significant biological processes such as 'macromolecule molecular process' and 'nucleobase, nucleoside, nucleotide and nucleic acid metabolic process' are lower level terms of 'metabolic process'. 'DNA repair' which was a significantly enriched category found in the full GO term analysis is one of the lower level components of the 'metabolic process' term found in the GO Slim analysis. Metabolic process involves small molecule

transformation but also macromolecular processes such as DNA repair and replication as well as protein synthesis and degradation ultimately resulting in cell growth. However, metabolic process does not involve protein-protein interaction or protein-nucleic acids interactions (http://www.geneontology.org).

C: Cellular Component

The GO Slim analysis showed enrichment of SATs for localization in the 'cytoplasm' and 'intracellular' (Table 14). The GO Slim term 'intracellular' is an upper level cellular component term of 'cytoplasm' and 'nucleus' GO terms so enrichment in either of these two should show enrichment for 'intracellular'. Other researchers have identified SATs as being nuclear localized using Northern hybridization (Kiyosawa et al. 2005). This could be possible when describing the localization of the mRNA transcripts since the majority of transcripts remain in the nucleus where they are eventually broken down into their components. However, the GO Slim terms 'intracellular' and 'cytoplasm' describe the functional location of the gene products not the transcripts. These 'intracellular' gene products are located in the cytoplasm as opposed to other locations such as 'membrane' or 'cell surface'.

GO Slim: All SATs with TE in SABR and all SATs without TE in SABR

There were no specific GO Slim term differences between the group of SATs that had TE sequence in the SABR (Table 15) and the group of SATs that did not have TE sequence in the SABR (Table 16). Both are beginning to reflect enrichments for the GO Slim terms that are enriched for the dataset containing all SATs but neither dataset showed any gain of a specific specialized function or process.

Table 14. GO Slim analysis for all SAT sequences. Three species (Sp.), human (Hs), mouse (Mm) and cow (Bt), are divided into the three principle Gene Ontology categories using the population dataset of all RefSeq sequences and study set of all unique sequences involved in a SAT pair with ≥95% SABR complementation.

Sp.	GO identifier	Population fraction	Study fraction	Bonferroni e-score	GO term
		F: molecular function			
Hs	GO:0003824	2273/6356	77/175	0.318	catalytic activity
Hs	GO:0005215	498/6356	22/175	0.407	transporter activity
Hs	GO:0016787	959/6356	37/175	0.422	hydrolase activity
Mm	GO:0003824	2043/6018	99/194	1.098e-05	catalytic activity
Mm	GO:0016787	866/6018	46/194	0.007	hydrolase activity
Mm	GO:0005488	3854/6018	143/194	0.053	binding
Mm	GO:0016740	698/6018	34/194	0.194	transferase activity
		P: biological process			
Hs	GO:0006810	1014/5817	41/161	0.113	transport
Mm	GO:0008152	2734/5397	112/166	0.001	metabolic process
Mm	GO:0043170	2145/5397	90/166	0.002	macromolecule metabolic process
Mm	GO:0009058	474/5397	26/166	0.046	biosynthetic process
Mm	GO:0006810	938/5397	42/166	0.109	transport
Mm	GO:0006139	1099/5397	46/166	0.248	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
Bt	GO:0008152	765/1231	10/11	0.405	metabolic process
		C: cellular component			
Hs	GO:0005737	1767/5750	63/167	0.301	cytoplasm
Mm	GO:0005622	2873/5539	132/178	5.099e-09	intracellular
Mm	GO:0005737	1474/5539	81/178	2.804e-07	cytoplasm

Table 15. GO Slim analysis for sequences with SABR containing TE sequence. Three species (Sp.), human (Hs), mouse (Mm) and cow (Bt), are divided into the three principle Gene Ontology categories using the population dataset of all RefSeq sequences and study set of all unique sequences involved in a SAT pair with ≥95% SABR complementation and aligning to a TE sequence.

Sp.	GO identifier	Population fraction	Study fraction	Bonferroni e-score	GO term
		F: molecular function			
none					
		P: biological process			
Hs	none				
Mm	GO:0008152	2734/5397	30/41	0.0353	metabolic process
Bt	none				
		C: cellular component			
Hs	none				
Mm	GO:0005622	2873/5539	38/50	0.0031	intracellular
Mm	GO:0005737	1474/5539	25/50	0.0026	cytoplasm
Bt	none				

Table 16. GO Slim analysis for sequences with no TE sequence in the SABR. Three species (Sp.), human (Hs), mouse (Mm) and cow (Bt), are divided into the three principle Gene Ontology categories using the population dataset of all RefSeq sequences and study set of all unique sequences involved in a SAT pair with ≥95% SABR complementation and not aligning to a transposable element sequence.

Sp.	GO identifier	Population fraction	Study fraction	Bonferroni e-score	GO term
		F: molecular function			
Hs	GO:0016787	959/6356	33/145	0.1976	hydrolase activity
Hs	GO:0003824	2273/6356	67/145	0.1304	catalytic activity
Mm	GO:0016787	866/6018	37/149	0.0103	hydrolase activity
Mm	GO:0003824	2043/6018	79/149	2.2776e-05	catalytic activity
Mm	GO:0005488	3854/6018	111/149	0.0890	binding
Bt	none				
		P: biological process			
Hs	none				
Mm	GO:0009058	474/5397	20/126	0.1205	biosynthetic process
Mm	GO:0008152	2734/5397	83/126	0.0066	metabolic process
Mm	GO:0043170	2145/5397	66/126	0.0468	macromolecule metabolic process
Bt	none				
		C: cellular component			
Hs	GO:0005737	1767/5750	55/138	0.1324	cytoplasm
Mm	GO:0005622	2873/5539	95/130	4.2190e-06	intracellular
Bt	none				

GO Slim: All RefSeq Sequences Aligning to a TE Sequence

The GO Slim analysis showed enrichment of human and mouse RefSeq transcripts that aligned to any transposable element sequence for localization in the nucleus and its upper level component ‘intracellular’ (Table 17). The GO Slim and full GO term mouse analyses found significant enrichments for ‘nucleic acid binding’ in addition to ‘catalytic activity’. The principle category ‘biological process' showed enrichment for the GO Slim terms ‘metabolic process’ which is an upper level process of ‘transcription’ and ‘regulation of transcription, DNA-dependent’ terms which were found significant in the full GO term analysis for all RefSeq sequences aligning to a TE sequence. Additional GO Slim enrichments in this dataset included ‘nucleobase, nucleoside, nucleotide and nucleic acid metabolic process’ and ‘macromolecule

metabolic process' both of which are lower level processes of 'metabolic process' which was also found as significant in the GO Slim analysis. The GO Slim term 'regulation of biological process' also had a high level of significance in mouse and a significant but higher e-score in human and was the only significant GO Slim term found in cow for this dataset (Table 17). A summary of all of these GO analyses is listed in Table 18.

Table 17. GO Slim analysis for all RefSeq sequences aligning to a TE sequence. Three species (Sp.), human (Hs), mouse (Mm) and cow (Bt), are divided into the three principle Gene Ontology categories using the population dataset of all RefSeq sequences and study set of all RefSeq sequences containing an alignment to a known transposable element sequence.

Sp.	GO identifier	Population fraction	Study fraction	Bonferroni e-score	GO term
		F: molecular function			
Hs	GO:0030528	642/6356	284/2541	0.277	transcription regulator activity
Hs	GO:0004386	76/6356	42/2541	0.115	helicase activity
Mm	GO:0030528	465/6018	177/1988	0.238	transcription regulator activity
Mm	GO:0004386	54/6018	27/1988	0.169	helicase activity
Mm	GO:0016829	53/6018	26/1988	0.263	lyase activity
Mm	GO:0016787	866/6018	322/1988	0.072	hydrolase activity
Mm	GO:0003824	2043/6018	776/1988	8.244e-08	catalytic activity
Mm	GO:0003676	1026/6018	425/1988	9.247e-09	nucleic acid binding
Mm	GO:0016740	698/6018	268/1988	0.021	transferase activity
Bt	none				
		P: biological process			
Hs	GO:0050789	1875/5817	806/2293	0.002	regulation of biological process
Hs	GO:0006139	1509/5817	709/2293	4.732e-11	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
Mm	GO:0009058	474/5397	185/1741	0.014	biosynthetic process
Mm	GO:0006810	938/5397	340/1741	0.051	transport
Mm	GO:0007610	131/5397	56/1741	0.146	behavior
Mm	GO:0008152	2734/5397	1067/1741	4.284e-26	metabolic process
Mm	GO:0030154	561/5397	222/1741	0.001	cell differentiation
Mm	GO:0050789	1381/5397	545/1741	6.526e-10	regulation of biological process
Mm	GO:0006139	1099/5397	464/1741	9.152e-14	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
Mm	GO:0008219	223/5397	99/1741	0.002	cell death
Mm	GO:0043170	2145/5397	842/1741	8.325e-18	macromolecule metabolic process
Bt	GO:0050789	281/1231	98/346	0.055	regulation of biological process
		C: cellular component			
Hs	GO:0005634	1786/5750	818/2295	7.063e-09	nucleus
Hs	GO:0005622	3717/5750	1572/2295	3.704e-06	intracellular
Mm	GO:0005634	1283/5539	523/1815	4.162e-11	nucleus
Mm	GO:0005622	2873/5539	1105/1815	3.856e-20	intracellular
Mm	GO:0005694	99/5539	49/1815	0.004	chromosome
Mm	GO:0005737	1474/5539	545/1815	0.001	cytoplasm
Bt	none				

Table 18. Summary of significant GO enrichments for the ≥95% SABR complementation dataset analyses. All results for human (Hs) and mouse (Mm) species (Sp) are significant to $p < 0.001$.

GO analysis	Study Set	Sp	Results Table	GO term
Full GO	All SATs	Hs	Table 14	nuclease activity
Full GO	All SATs	Mm	Table 14	response to DNA damage stimulus
Full GO	RefSeq w/ TE	Hs	Table 15	nucleus
Full GO	RefSeq w/ TE	Hs	Table 15	zinc ion binding
Full GO	RefSeq w/ TE	Hs	Table 15	intracellular
Full GO	RefSeq w/ TE	Hs	Table 15	metal ion binding
Full GO	RefSeq w/ TE	Hs	Table 15	transcription
Full GO	RefSeq w/ TE	Mm	Table 15	regulation of transcription, DNA-dependent
Full GO	RefSeq w/ TE	Mm	Table 15	zinc ion binding
Full GO	RefSeq w/ TE	Mm	Table 15	nucleus
Full GO	RefSeq w/ TE	Mm	Table 15	metal ion binding
GO Slim	All SATs	Mm	Table 16	catalytic activity
GO Slim	All SATs	Mm	Table 16	metabolic process
GO Slim	All SATs	Mm	Table 16	intracellular
GO Slim	All SATs	Mm	Table 16	cytoplasm
GO Slim	SABRs w/o TE	Mm	Table 18	catalytic activity
GO Slim	SABRs w/o TE	Mm	Table 18	intracellular
GO Slim	RefSeq w/ TE	Mm	Table 19	catalytic activity
GO Slim	RefSeq w/ TE	Mm	Table 19	nucleic acid binding
GO Slim	RefSeq w/ TE	Hs	Table 19	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO Slim	RefSeq w/ TE	Mm	Table 19	metabolic process
GO Slim	RefSeq w/ TE	Mm	Table 19	cell differentiation
GO Slim	RefSeq w/ TE	Mm	Table 19	regulation of biological process
GO Slim	RefSeq w/ TE	Mm	Table 19	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO Slim	RefSeq w/ TE	Mm	Table 19	macromolecule metabolic process
GO Slim	RefSeq w/ TE	Hs	Table 19	nucleus
GO Slim	RefSeq w/ TE	Hs	Table 19	intracellular
GO Slim	RefSeq w/ TE	Mm	Table 19	nucleus
GO Slim	RefSeq w/ TE	Mm	Table 19	intracellular
GO Slim	RefSeq w/ TE	Mm	Table 19	cytoplasm

w/ = with
w/o = without

The Effects of SABRs on Protein Folding Properties

The alignment of PDB structures sequences can give an insight to the affect of SABRs and TEs on the functionality of gene products. As with the GO analysis, the analysis is only as accurate as the current state of annotations for the gene products of an organism. There were some alignments of the PDB structure sequences to the regions of transcripts that aligned to a transposable element sequence for all three organisms. Additionally, there were alignments of the PDB structure sequences to the SABRs for the ≥80% SABR complementation dataset for each organism. There may be some level of the dsRNA regions of SABRs affecting the protein folding potential, however, the results show that there are some SABRs that align to PDB confirmed structure sequences for human and mouse (Table 19). Additionally, there were some alignments to the regions of genes that are transposable element sequence either in the sense or antisense orientation.

Table 19. **Population and study dataset numbers for the functional inhibition of SABRs study.** Population includes all RefSeq sequences that align to a PDB structure sequence. Number of sequences in the ≥95% SABR complementation dataset that have a PDB structure sequence aligning to a SABR that overlaps a CDS (PDB in SABR) and number of sequences in the RefSeq dataset where a transposable element overlap a PDB structure sequence (TE in PDB).

	Population	PDB in SABR	TE in PDB
Human	3,268/23,663	4/37	79/9950
Mouse	2,406/19,724	21/21	47/7044

Discussion

Using GO Slim terms to perform a GO analysis provides insight to possible enrichments for particular GO categories for selected study samples. However, a GO analysis relies on annotators to provide the GO terms relating to the gene or gene

products. A GO analysis will become more useful and precise as future research provides additional genes that are annotated with GO terms. The human and mouse had more significant results than cow due to the higher number of annotations that include GO terms. The most significant e-values were found in mouse for most of the enriched GO Slim terms. Patterns of GO Slim term enrichment similar to human and mouse began to emerge in *Bos taurus* perhaps since the most studied and conserved sequences are prime candidates for gene annotation using comparative genomics.

The GO Slim analysis gives insights into general upper level GO categories enrichments while the full GO term analysis gave more specific processes. The full GO term analysis didn't find the enrichments for cellular component terms ‘nucleus’ and ‘cytoplasm’ which were found with the GO Slim analysis. However, the GO Slim did identify enrichments in upper level terms for ‘nuclease activity’, ‘DNA repair’, ‘Response to DNA damage stimulus’ and ‘citrate metabolic process’ which were identified by the full GO term analysis. The full GO term analysis of SATs found ‘DNA repair’ and ‘citrate metabolic process’ enrichments which are both lower level terms of the GO Slim enriched term ‘metabolic process’. The full GO term analysis coupled with the GO Slim analysis better defines enrichments in the study sets.

There was no observable difference in the GO Slim analysis for all SATs containing a transposable element in the SABR and SATs that did not contain a transposable element in the SABR. There was also no significantly enriched category in either of these two subsamples when compared to the study sample of all RefSeq sequences involved in a SAT pair. Based on these findings, there is no support showing that SATs containing TE sequence have any specific specialized function when compared to SATs that do not contain TE sequence

Although other researchers found SAT gene transcripts to be nuclear localized, the results of this GO Slim study suggest that the gene products of SATs are mostly localized within the cytoplasm. RefSeq sequences containing transposable element sequence were enriched for the cellular component GO Slim terms ‘nucleus’ and ‘intracellular’.

The alignments of the PDB structure sequences to SABRs and TEs demonstrated that there are some transcripts in which these two characteristics do not alter the properties of a folding protein. This analysis was challenging since many SABRs are located within UTRs which will not align to PDB structure sequences. Moreover, aligning PDB structure sequences to the entire transcript whose length can be thousands of nucleotides long is somewhat biased when attempting to align PDB structures to small SABR or TE sequences which average around 300 nucleotides.

CHAPTER IV

DETECTING EVOLUTIONARY RATE VARIATION IN SENSE-ANTISENSE TRANSCRIPTS

Overview

Substitution rates across orthologous genes can give insights to specific conserved regions of gene sequences. Lower substitution rates may be interpreted as increased conservation but other characteristics of these genomic regions should be added to the analysis. The goals of this project were to investigate the conservation level across the 3' UTRs of a set of sense-antisense transcripts (SATs) with sense-antisense binding regions (SABRs) residing primarily in the 3' UTR sequence and to possibly identify reasons for increased conservation levels in different areas of the 3' UTRs. The substitution rate was used as a measurement of conservation in the 3' UTRs of a set of 8,384 orthologous human and mouse genes. A subset of human and mouse *cis*-oriented SATs was compared to the conservation of the 3' UTRs of the 8,384 orthologous human and mouse genes to determine if there were significant differences in specific regions in the 3' UTRs of the subset of SAT genes. The data indicated that the SABRs of a set of SATs lie in a region of increased conservation when compared to the rest of the 3' UTR indicating that the SABRs may be conserved and serve a biological purpose. However, these SABRs lie in regions of low %GC which may account for increased conservation levels at the 3' end of the 3' UTRs in human and mouse genes involved in the *cis*-oriented SATs.

Introduction

Natural selection is the driving force of evolution which can occur due to changes

in either coding or non-coding DNA sequences. Selective pressure can be estimated by comparing the number of substitutions or nucleotide mutations that have occurred in the sequences of orthologous genes between two or more species. A set of conserved human and mouse SATs was evaluated in this study to determine substitution rates across the orthologous transcripts. Differences in substitution rates across the 3' UTR can give insight to the possibilities of conserved functionality in conserved orthologous 3' UTR SABRs. The substitution rate for the 3' UTR preSABR (the region from the stop codon to the beginning of the SABR) was compared with the substitution rate of the SABR for each of a set of 13 SAT pairs to identify the substitution rates of these two regions of the UTR which may indicate possible function for a conserved SABR. The Wilcoxon test was used to test for significant differences between the substitution ratios between K_A/K_S, K_{3UPS}/K_S and K_{3USABR}/K_S.

Objectives

The objective of this study was to search for significant differences in substitution rates of the 3' UTRs of human and mouse orthologous genes and identify possible factors contributing to sequence divergence in different regions of the UTRs. This study can provide a better understanding of sequence divergence in the 3' UTRs of human and mouse genes containing SABRs in their 3' UTRs. Conserved 3' UTR SABRs may indicate a conserved function of the SABR sequence; however, other factors contributing to sequence conservation such as %GC are explored.

Materials and Methods

Thirteen SAT pairs (26 genes) were selected based on conservation between

human and mouse for the comparisons of the CDS and SABR, and further selected based on sufficient lengths of preSABR (>25 nucleotides) for the comparison of preSABRs (the 3' UTR sequence located upstream of the SABR) to SABRs. Additionally, these SATs were selected because their overlapping regions are purely UTR sequence. The conserved SAT set was derived from the conserved genes calculated earlier using the best FASTA alignments finding genes with conserved SATs. The SATs were all *cis*-oriented with SABRs overlapping only in the 3' UTRs. This is a good representative set of SATs since a majority of the SABRs are in the 3' UTR convergently arranged genes. The coding sequence, preSABR and SABR sequences were parsed using a Perl script and the CDS and SABR coordinates. The human and mouse corresponding gene sequences were aligned using MUSCLE (Edgar, 2004). The MUSCLE alignment was then translated using EBI translation tool (http://www.expasy.ch/tools/dna.html) and then aligned using NCBI's blast2seq application (Tatusova and Madden, 1999) to verify that the MUSCLE alignment conserved the amino acid sequence for each gene. Once verified, the MUSCLE alignment was used as input for the SNAP web program (Korber, 2000) provided by the HIV database website (www.hiv.lanl.gov) which calculated the synonymous and nonsynonymous rates. The SNAP program uses the Nei-Gojobori codon based method incorporating the Jukes-Cantor correction for multiple substitutions at the same site for protein coding sequences.

The human and mouse orthologous SABRs were aligned using MUSCLE. The chi-square test was used to test for significant divergence from the null hypothesis that the substitution rate in the preSABR was equal to that of the SABR using the Kimura two-parameter model (Figure 4) for both regions. The values used were the number of substitutions and the number if identical nucleotides in the alignments. Gaps in the

alignment could be the addition of sequence in one organism or the deletion of that sequence region in the other showing lack of conservation for the gapped region. Therefore, gaps were removed from the alignment and not used in the computations. Poly-A tails were also removed since they may be present in the sequence submitted to public databases but not present in the genomic sequence and they can alter the calculations of substitution rates.

$$K = - (½) \ln\{(1 - 2P - Q) \sqrt{(1 - 2Q)}\}$$

Figure 4.
Kimura two-parameter model equation. P = the number of transition sites divided by the total number of compared sites and Q = the number of transversion sites divided by the total number of compared sites (Kimura, 1980).

The Wilcoxon test was used to test for significant differences between the substitution ratios K_A/K_S, K_{3UPS} and K_{3USABR}. The Wilcoxon rank-sum test is used when the two populations tested are not normal and returns a p-value to test that the two population distributions are the same. Since different regions of a coding sequence can have different K_A/K_S ratios (Liang et al. 2006), the substitution rate of the 3' UTRs was further investigated by using a window of 100 bases sliding every 25 across the length of the 3' UTR. A local installed version of the sequence alignment tool MUSCLE (Edgar, 2004) was used to make the alignments of the regions of interest. The Kimura two-parameter model test was applied to each window and the results were graphed and analyzed.

The 3' UTRs from 8,384 human and mouse orthologous genes were retrieved from the BioMart database (http://www.biomart.org/biomart/martview) in order to calculate the substitution rate across the 3' UTRs. BioMart uses a reciprocal best hit

approach for determining orthologous sequences. Each 3' UTR of each orthologous gene were aligned using the MUSCLE application. A Perl script was used to remove gaps and poly-A tails from both sequences. Only sequences longer than 300 nucleotides were used. The aligned sequence was then divided into 10 equal sized windows of varying length depending on the length of the 3' UTR. This 10-window approach allowed comparisons of all the 3' UTRs for all the orthologous pairs. The Kimura two-parameter test was then used to calculate the substitution rate within each of the 10 windows for all the alignments and the results were graphed and analyzed. Additionally, the percent GC content for each window for the alignments were calculated. The Tukey-HSD test is used to test all pairwise comparisons among means. In this study, the Tukey-HSD test was performed to test for significant differences for each pair of means for the substitution rates for each of the 10 windows. The Tukey-HSD was also used to test for significant differences in the mean percent GC for each of the 10 windows across the 3' UTRs of human and mouse orthologous genes.

Results

CDS vs. UTR Substitution Rates

Deviations from neutral evolution among coding sequences can be estimated using the K_A/K_S ratio. The ratio of nonsynonymous to synonymous (K_A/K_S = 0.12) substitutions in the human and mouse orthologous gene set was similar to rates found in a study of 52 human genes with mouse orthologs (K_A/K_S = 0.16) (Li and Su, 2006). Neutral evolution can be estimated using similar logic when comparing coding sequence to non-coding sequence substitution rates. When the ratio $K_{3UPS} / K_S = 1$ then neutral evolution can be inferred between the two regions of interest. When $K_{3UPS} / K_S < 1$,

selective constraint can be inferred. Additionally, positive, or diversifying, selection is inferred when $K_{3UPS} / K_S > 1$. This logic has been used in previous sequence comparisons of coding and non-coding regions (Li and Su, 2006). The average ratio of K_{3UPS} / K_S for 12 of the 13 SAT gene pairs (one was removed because the preSABR was < 10 nt) was 0.811 indicating neutral evolution of the preSABR with the corresponding coding sequence. However, the ratio of K_{3USABR}/Ks for the 13 SAT pairs was 0.448 indicating selective constraint. The null hypotheses of $K_A/K_S = K_{3UPS}/K_S$ and $K_A/K_S = K_{3USABR}/K_S$ were rejected, $p < 0.001$ and $p < 0.001$ respectively, using the two-tailed Wilcoxon test. In both cases, the substitution ratio for the CDS was lower than the UTR ratios.

Substitution Rates: preSABR vs. SABR

The estimation of the average sequence substitution rates in the preSABR (K_{3UPS}) for 12 of the 13 SAT pairs (one was dropped because the preSABR sequence was < 10 nt) was 0.325 which was similar to the previous study of the substitution rate of the full 3' UTR of 52 human/mouse genes ($K_{3U} = 0.318$) (Li and Su, 2006). The average SABR length for the 13 SAT pairs was 253 and the average preSABR length was 600. The average substitution rate for the SABRs ($K_{3USABR} = .202$) was lower than the preSABR UTR sequences ($K_{3UPS} = 0.325$) indicating conservation of sequence in this region. A chi-square test of the observed and expected rates in the preSABR vs. the SABR resulted in 17 of the 26 tests with significant p-values ($p < 0.05$) allowing rejection of the null hypothesis of $K_{3UPS} = K_{3USABR}$. All significant preSABR substitution rates were higher than their corresponding SABR substitution rates (Table 20). The null hypothesis of $K_{3UPS}/K_S = K_{3USABR}/K_S$ was rejected ($p < 0.001$) and the two-tailed Wilcoxon test provided evidence for significantly lower substitution rates in the SABRs (K_{3USABR}) than

their corresponding pre-SABRs (K_{U3PS}).

Table 20. Substitution rates for coding sequences (K_S, K_A and K_A/K_S), 3' UTR preSABR (K_{3UPS}) and the SABR (K_{3US}) sequences for 26 human and mouse genes forming *cis* SATs. Chi-square p-values are shown for substitution rate comparisons between preSABR and SABR sequences.

Human	Mouse	KS	KA	KA/KS	K3UPS	K3US	p-value
PHYHIPL	Phyhipl	0.495	0.011	0.023	0.161	0.195	0.166
FAM13C1	1200015N20Rik	0.668	0.080	0.119	0.293	0.195	0.006
ABI1	Abi1	0.374	0.009	0.024	0.183	0.098	0.013
PDSS1	Pdss1	0.478	0.076	0.158	0.169	0.098	7.48e-11
DMTF1	Dmtf1	0.361	0.024	0.065	0.263	0.126	0.010
C7orf23	4930420K17Rik	0.265	0.011	0.041	0.330	0.126	0.002
EME1	Eme1	0.533	0.177	0.332	0.378	0.171	0.001
LRRC59	Lrrc59	0.510	0.021	0.041	0.263	0.171	0.033
GBA2	Gba2	0.479	0.065	0.136	0.259	0.347	0.307
CREB3	Creb3	0.658	0.191	0.290	0.414	0.347	0.551
PPP1CA	Ppp1ca	0.647	0.003	0.004	0.230	0.159	0.222
RAD9A	Rad9a	0.444	0.101	0.229	0.475	0.159	1.74e-08
SLC12A6	Slc12a6	0.061	0.133	2.171	0.311	0.135	0.001
TMEM85	Tmem85	0/350	0.002	0.007	0.401	0.135	3.13e-04
LRIG1	Lrig1	0.609	0.083	0.136	0.282	0.040	3.04e-07
SLC25A26	Slc25a26	0.645	0.062	0.097	0.399	0.040	2.91e-10
BECN1	Becn1	0.470	0.008	0.016	0.146	0.225	0.143
CNTD1	Cntd1	0.461	0.084	0.183	0.389	0.225	1.72e-04
YPEL1	Ypel1	0.687	0.004	0.005	0.351	0.438	0.736
PPIL2	Ppil2	0.763	0.049	0.064	0.510	0.438	0.996
CACNA1H	Cacna1h	0.661	0.076	0.115	0.506	0.191	4.40e-05
TPSG1	Tpsg1	0.676	0.164	0.243	0.297	0.191	0.456
DHPS	Dhps	0.588	0.047	0.080	+++	0.366	+++
MORG1	1500041N16Rik	0.533	0.035	0.065	0.976	0.366	0.004
MGC4655	C76566	0.897	0.086	0.096	0.533	0.130	7.79e-11
C16orf70	D230025D16Rik	0.386	0.015	0.040	0.245	0.130	2.13e-04
AVERAGES:		**0.532**	**0.006**	**0.012**	**0.325**	**0.202**	

+++ Omitted because length is less than 10 bases.

A sliding window of 100 bases moved in 25 base intervals across the length of the 3' UTR revealed that the substitution rate was higher in the initial portion of the 3' UTR and was lowest at the end for the RAD9A and Rad9a orthologous human and mouse genes (Figure 5) presenting evidence that $K_{3UPS} \neq K_{3USABR}$ for these genes. This indicates that in these transcripts, there is some conservation of sequence between human and mouse towards the end of the 3' UTR where the SABR resides.

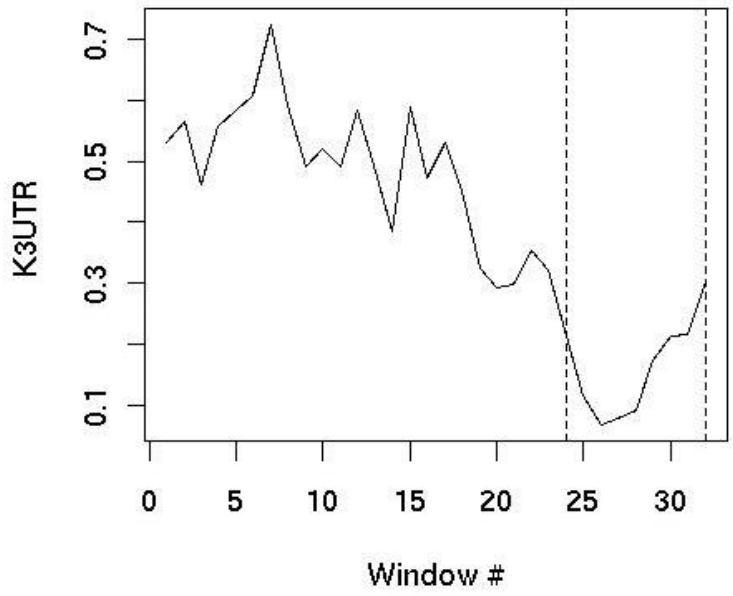


Figure 5.
Substitution rate change across the 3' UTR for the human gene RAD9A and homologous mouse gene Rad9A. A sliding window of 100 bases was moved across the 3' UTR every 25 bases and the Kimura two-parameter test was performed on each window. The SABR is shown as the region between the dashed vertical lines. Gaps and poly-A tails were removed from the alignment.

The gene PPP1CA, which is the other member of the SAT pair with RAD9A and an antisense transcript to RAD9A, also demonstrated significantly lower substitution rates in the SABR when compared to the preSABR sequence (Figure 6).

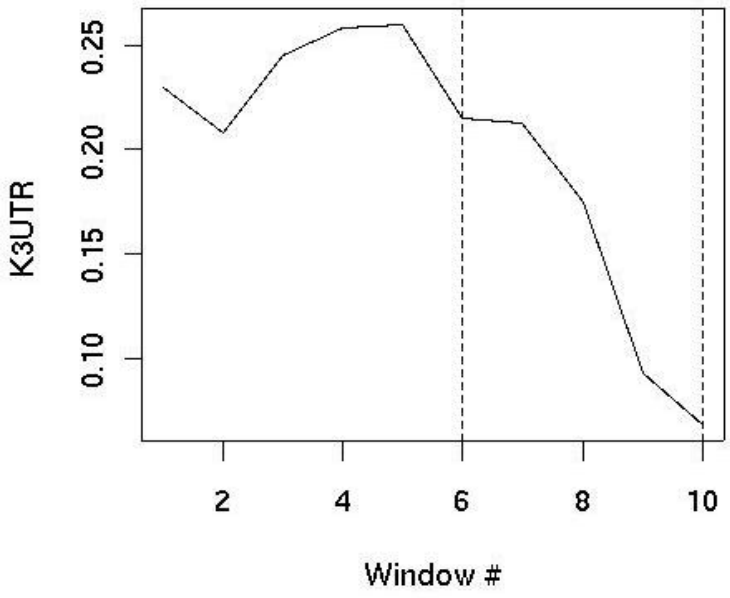


Figure 6.
Substitution rate change across the 3' UTR for the human gene PPP1CA and the homologous mouse gene Ppp1ca. A sliding window of 100 bases was moved across the UTR every 25 bases and the Kimura two-parameter test was performed on each window. The SABR is shown as the region between the dashed vertical lines. Gaps and poly-A tails were removed from the alignments prior to calculating K3UTR.

Using the 10-window approach, the substitution rates for each 3' UTR showed a general trend of lower substitution rates indicating sequence conservation towards the 3' end of the 3' UTR where the SABRs are located (Figure 7). The 5' end of the 3' UTR which is represented by the first window comprising the first 10% of the 3' UTR was significantly higher than the last window of 10% for the 26 *cis*-SATs gene dataset ($P < 0.001$).

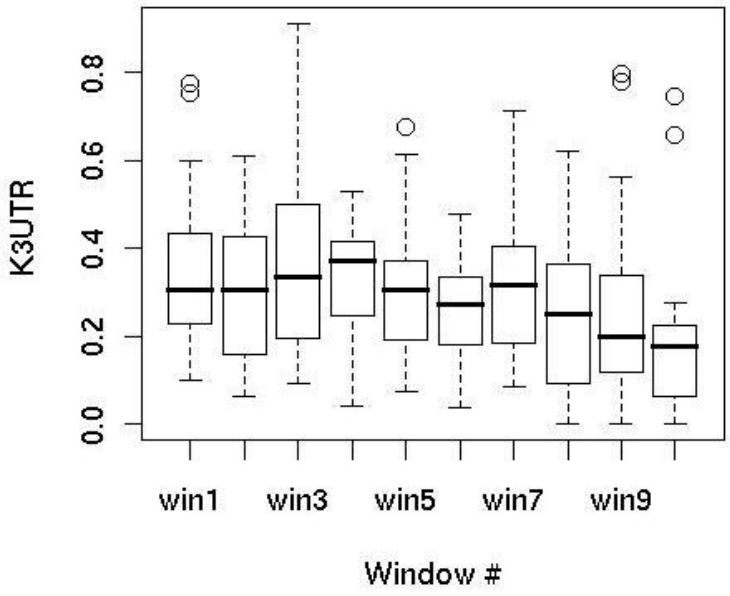


Figure 7.
Substitution rate change across the 3' UTR for the 26 human and mouse genes comprising 13 SAT pairs. The 3' UTR for the alignments of each orthologous human-mouse pair was divided into ten equal sized windows and the substitution rate was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating K3UTR.

Using the 10-window approach, the percent GC was calculated for each of the 10 windows across the 3' UTRs for the 26 human genes (Figure 8). The first window which is the 5' end of the 3' UTR was significantly higher than the last window located at the 3' end of the 3' UTR ($p < 0.001$). Additionally, the percent GC was significantly higher in the first window than the last ($p < 0.001$). The same results were observed for the 26 mouse genes (Figure 9).

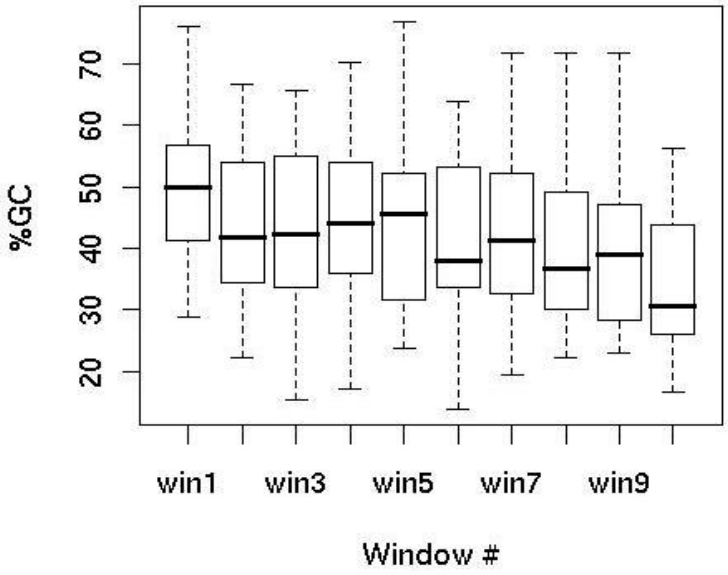


Figure 8.
Percent GC content across the 3' UTRs for the 26 human genes comprising 13 SAT pairs. The 3' UTR for each human gene was divided into ten equal sized windows and the percent GC was calculated for each window. The same windows were used as the K3UTR alignments in which gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

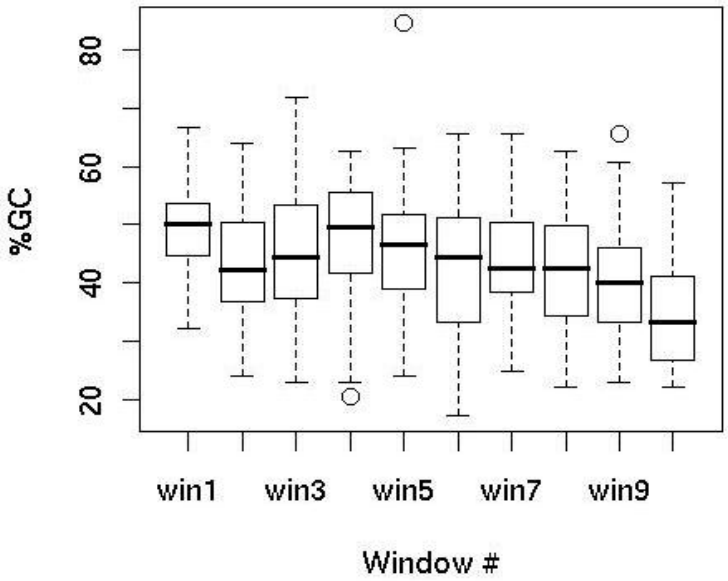


Figure 9.
Percent GC content across the 3' UTRs for the 26 mouse genes comprising 13 SAT pairs. The 3' UTR of each mouse gene was divided into ten equal sized windows and the percent GC was calculated for each window. The same windows were used as the K3UTR alignments in which gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

The substitution rate and percent GC was calculated for 8,384 3' UTRs of orthologous human and mouse genes. Each 3' UTR was divided into the windows to facilitate comparisons of UTRs of varying lengths. The average length of the 3' UTR after removing gaps and poly-A tails was 1,041 nucleotides. The distribution of lengths, rounded down to the nearest hundred, for these alignments is shown in Figure 10. Decreasing substitution rates were found from the 5' to the 3' regions of the 3' UTRs for the set of 8,384 orthologous human and mouse genes (Figure 11). The last window (win10) which comprised the last 10% of sequence for each alignment had a significantly lower mean substitution than the first window (win1) which comprises the first 10% of the sequence alignments ($p < 0.001$). Win10 is the region of the 3' UTR where conserved poly-A signal(s) generally reside and may be a significant factor in the resulting lower substitution rate for win10.

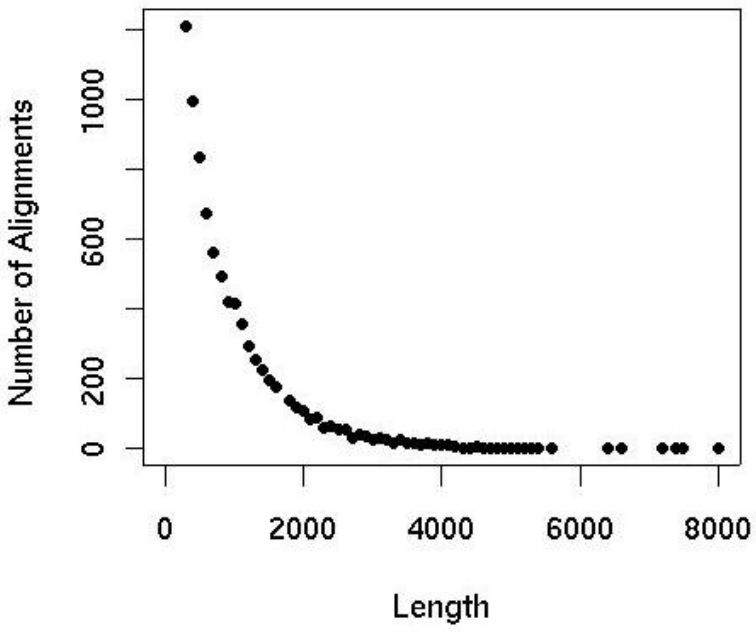


Figure 10.
Distribution of alignment lengths for the 3' UTRs of 8,384 human and mouse orthologous genes.
The alignment lengths were each rounded down to the nearest hundred.

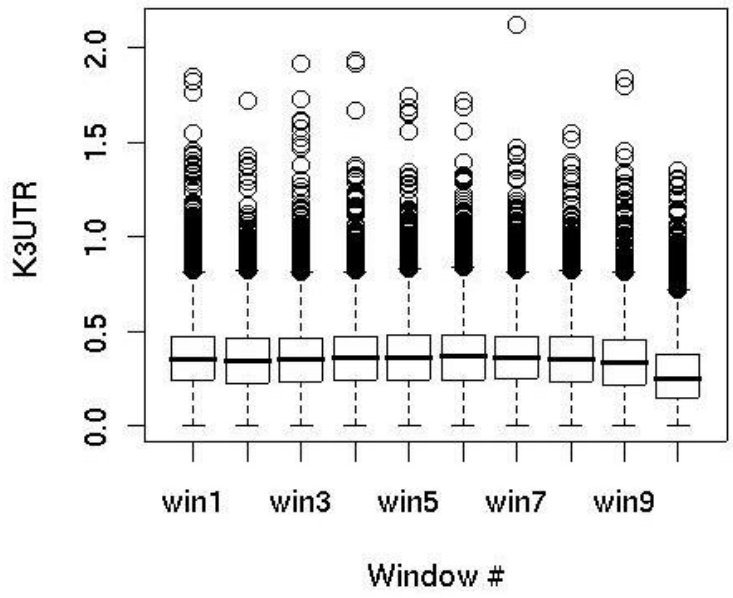


Figure 11.
Substitution rate change (K3UTR) across the 3' UTR for 8,384 human and mouse orthologous genes. The 3' UTRs for the alignments of each orthologous human-mouse pair was divided into ten windows and the substitution rate was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating K3UTR.

Additionally, the percent GC for human and mouse was calculated separately for each of the ten windows used in the substitution rate analysis. Both human and mouse had significantly higher means, $p < 0.001$ and $p < 0.001$ respectively, in the first window (win1) than the last (win10) (Figures 12 and 13 respectively).

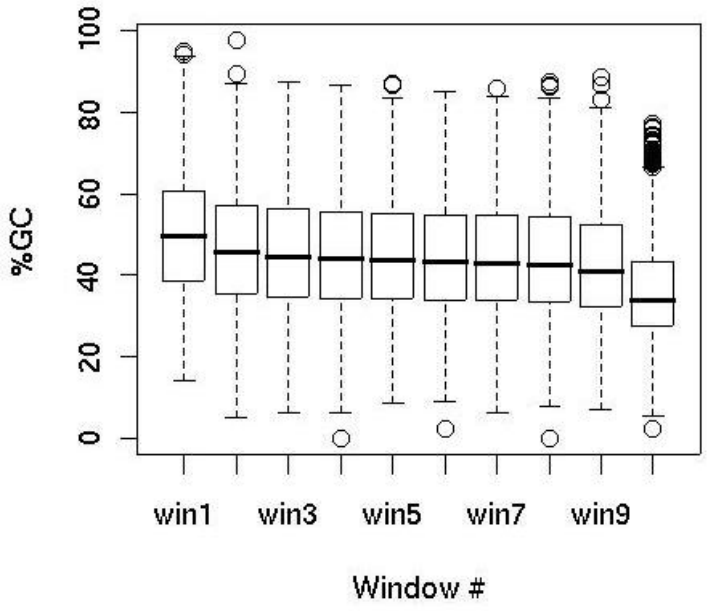


Figure 12.
Percent GC content across the 3' UTRs for 8,384 human genes. The 3' UTR for each human gene was divided into ten equal sized windows and the percent GC was calculated for each window. The same windows as the K3UTR alignments were used in which gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

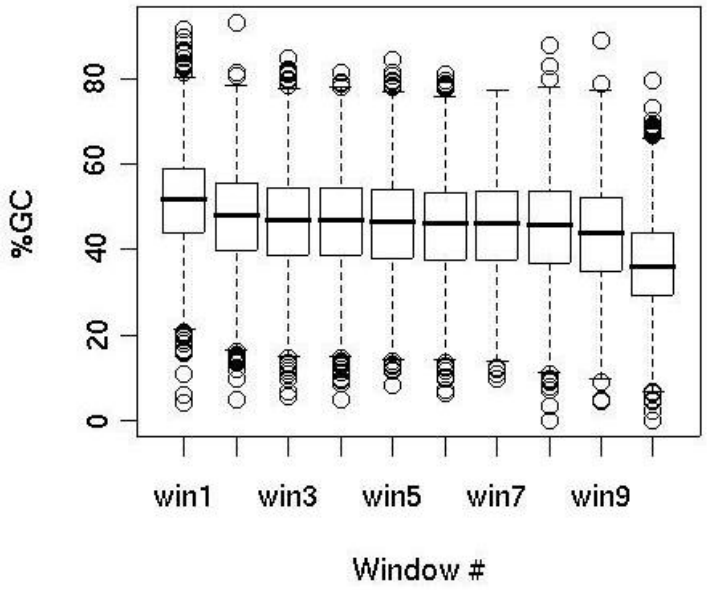


Figure 13.
Percent GC content across the 3' UTRs for 8,384 mouse genes. The 3' UTR for each mouse gene was divided into ten equal sized windows and the percent GC was calculated for each window. The same windows as the K3UTR alignments were used in which gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

The same ten-window algorithm was applied to the region 2,000 bases upstream of the 3' UTR. There were 13,035 alignments after removing gaps. The distribution of the substitution rates, rounded down to the nearest hundred, is shown in Figure 14.

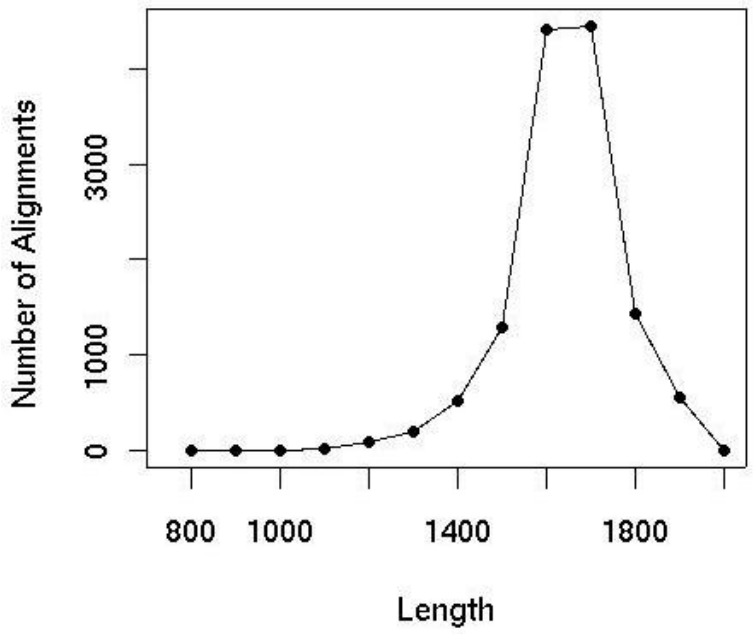


Figure 14.
Distribution of alignment lengths for the 2,000 bases upstream of the 3' UTR of 13,035 human and mouse orthologous genes. The alignment lengths were each rounded down to the nearest hundred.

The substitution rate (Ks) was lowest at window ten where the final exon of a gene would typically reside (Figure 15). Windows one through nine may have higher Ks values due to being located in intronic regions. The percent GC for each of the ten windows was also calculated on the same 13,035 alignment dataset. Window ten (win10) had a significantly higher ($p < 0.001$) GC content than the previous nine windows for both human (Figure 16) and mouse (Figure 17) datasets.

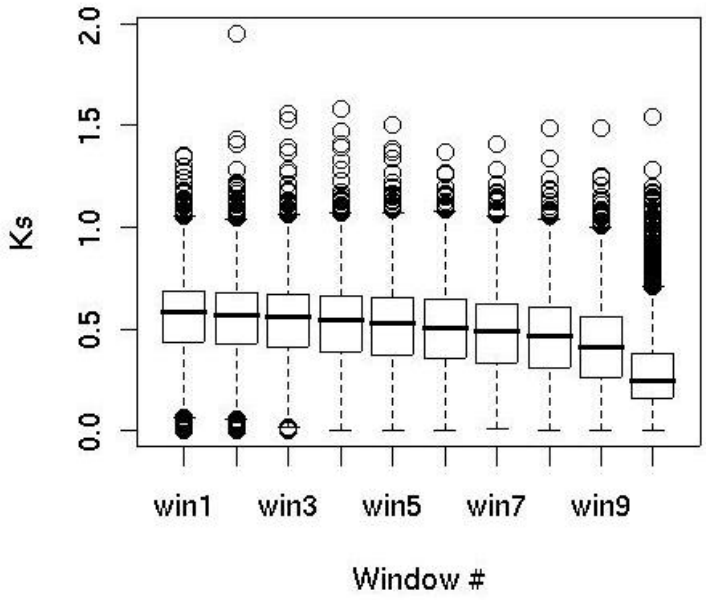


Figure 15.
Substitution rate (Ks) across the 2,000 bases upstream of the 3' UTR for 13,035 human and mouse orthologous genes. The region 2,000 bases upstream of the 3' UTR for the alignments of each orthologous human-mouse pair was divided into ten windows and the substitution rate was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating the substitution rate.

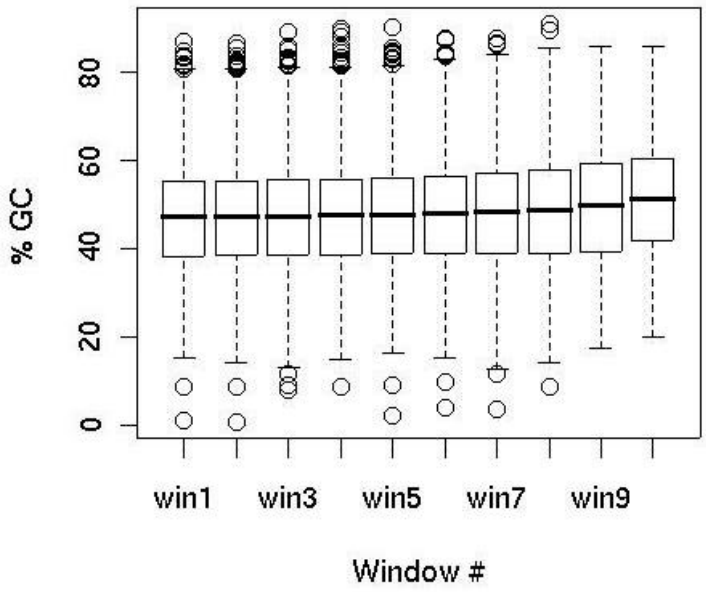


Figure 16.
Percent GC content across the region 2,000 bases upstream of the 3' UTR for 13,035 human genes. The region 2,000 bases upstream of the 3' UTR for each mouse gene was divided into ten equal sized windows and the percent GC was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

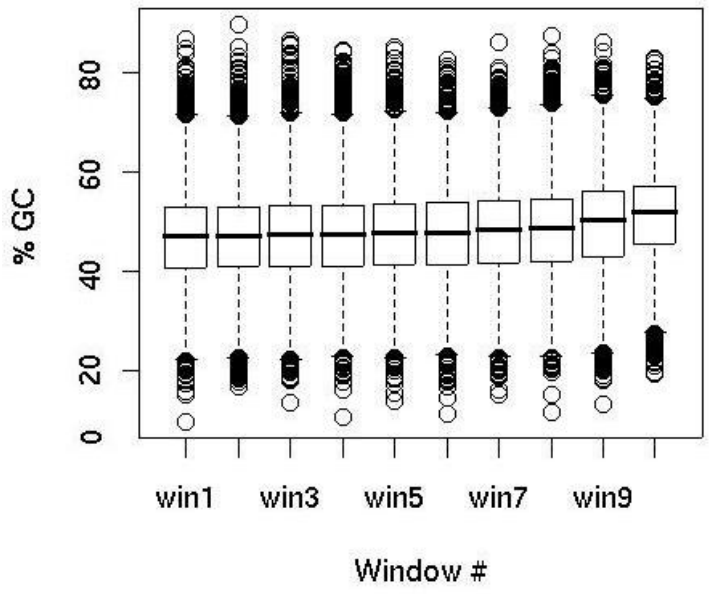


Figure 17.
Percent GC content across the region 2,000 bases upstream of the 3' UTR for 13,035 mouse genes. The region 2,000 bases upstream of the 3' UTR for each mouse gene was divided into ten equal sized windows and the percent GC was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

The 2,000 bases downstream of the 3' UTR were also analyzed using the ten-window algorithm. There were 13,024 alignments after gaps were removed. The distribution of the alignment lengths, rounded down to the nearest hundred, of this dataset is given in Figure 18. The substitution rates are lowest near the end of the 3' UTR (win1) and increase slightly and then plateau traveling farther downstream from the 3' UTR (Figure 19). This suggests that the intragenic region closest to the 3' UTR of a gene maintains a higher level of conservation than the region more downstream.

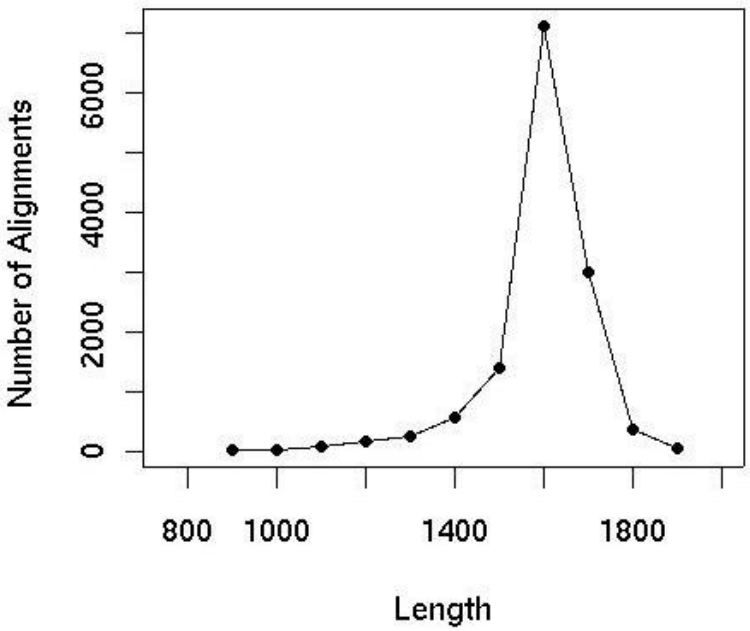


Figure 18.
Distribution of alignment lengths for the 2,000 bases downstream of the 3' UTR of 13,024 human and mouse orthologous genes. The alignment lengths were each rounded down to the nearest hundred.

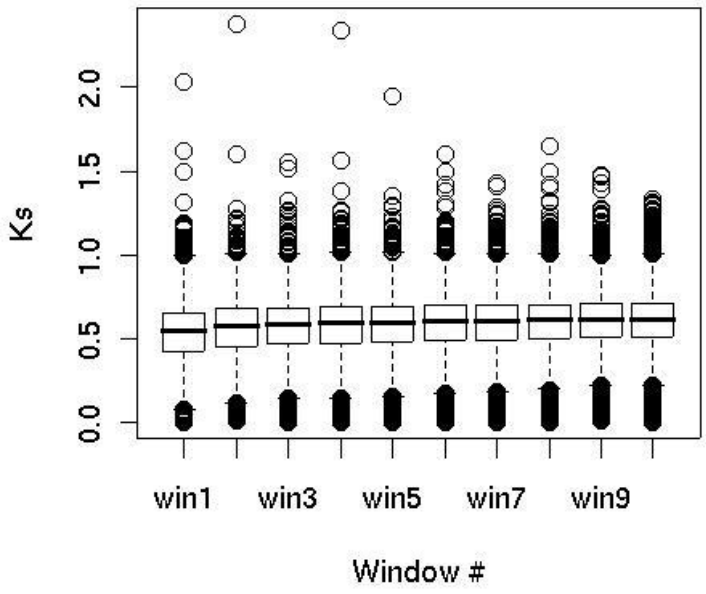


Figure 19.
Substitution rate (Ks) across the 2,000 bases downstream of the 3' UTR for 13,024 human and mouse orthologous genes. The region 2,000 bases downstream of the 3' UTR for the alignments of each orthologous human-mouse pair was divided into ten windows and the substitution rate was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating the substitution rate.

The percent GC for each of the ten windows was also calculated on the same 13,024 alignment dataset downstream of the 3' UTR. Window one (win1) had a significantly lower GC content than the previous nine windows for both human (Figure 20) and mouse (Figure 21) datasets. This is the window closest to the 3' end of the 3' UTR where conserved poly-A signals are generally located.

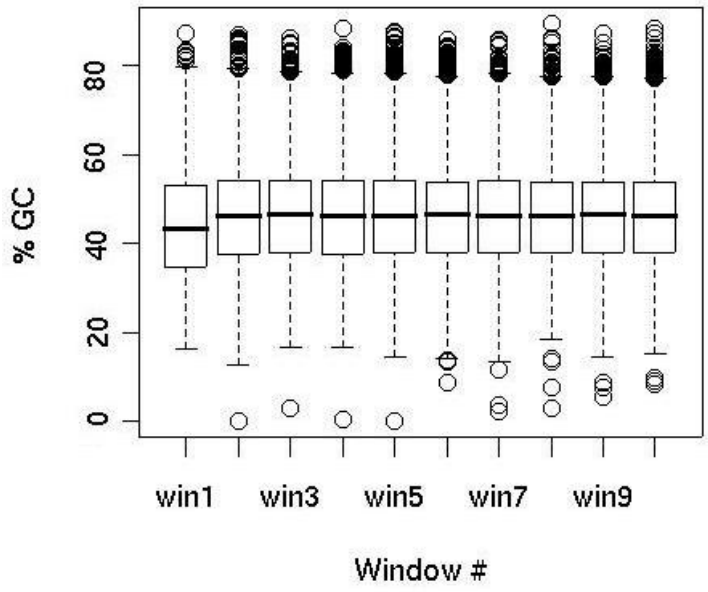


Figure 20.
Percent GC content across the region 2,000 bases downstream of the 3' UTR for 13,024 human genes. The region 2,000 bases downstream of the 3' UTR for each human gene was divided into ten equal sized windows and the percent GC was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

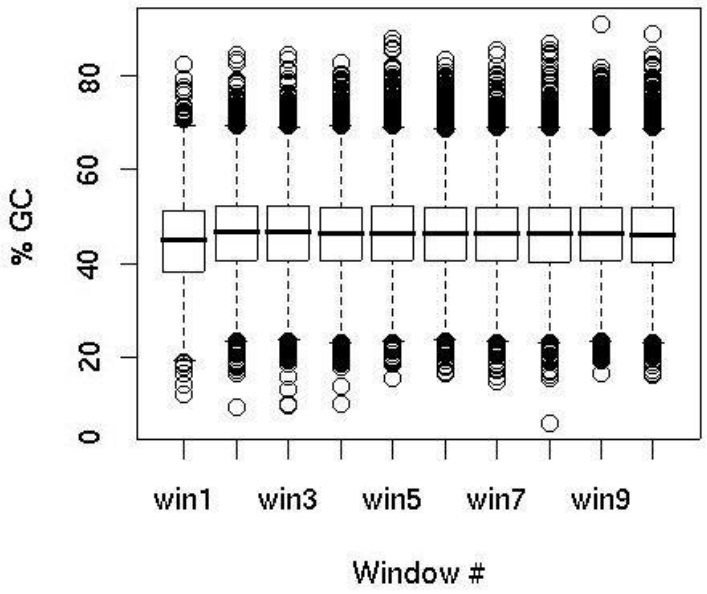


Figure 21.
Percent GC content across the region 2,000 bases downstream of the 3' UTR for 13,024 mouse genes. The region 2,000 bases downstream of the 3' UTR for each human gene was divided into ten equal sized windows and the percent GC was calculated for each window. Gaps and poly-A tails were removed from the alignments prior to calculating percent GC content.

Discussion

The chi-square test results indicated that $K_{3UPS} \neq K_{3USABR}$ in 17 of the 26 tests performed suggesting that some SABRs may be under selective constraint while others are not. However, there may be other conserved motifs across the preSABR that lead to the insignificant differences between the two UTR regions. The two *cis*-oriented SAT genes CREB3 and GBA2 have an overlapping region of 122 nucleotides. There are two poly-A signals at each end of the SABR on opposite strands that conserve the SABR boundaries, one is the conserved antisense hexamer of AAUAAA at the beginning of the SABR and the other contains the hexamer AAUAAA within a 16 nucleotide sequence AAGAAAUAAAUAAGUG at the end of the SABR. The 16 nucleotide sequence may be a translational suppressor site since it contains the poly-A signal and extension of the

poly-A tail is essential for translation initiation. A repressor protein recognizing and binding to this site may prevent poly-A polymerase from binding the poly-A site hindering poly-A extension and translation; this would need to be tested using a wet lab approach. These two regions make up 67% of the 33 identical conserved nucleotide sites located in the alignment between the human and mouse SABRs.

The substitution rate (K_{3U}) across the entire 3' UTR for the human gene, RAD9A, and mouse gene, Rad9a, was lower at the end of the 3' UTR which is the location of the SABR, indicating that the SABR may be under greater selective constraint than the rest of the UTR. This is supported with a significant chi-square test comparing the preSABR sequence substitution rate with the SABR substitution rate and the significant substitution ratio differences using the Wilcoxon test. However, the 3' UTR substitution rates for a dataset of 8,384 human and mouse orthologous gene pairs showed that there were significantly lower substitution rates and percent GC in the final window comprising the last 10% of the UTRs when compared to the other nine windows comprising the first 90% of the UTRs. Therefore, the conservation in the SABRs may be due to the coincidental presence of one or more conserved poly-A signals in the region rather than the need for binding of complementary SATs. Previous studies finding a negative correlation between 3' UTR overall sequence conservation and percent GC in mammalian 3' UTRs (Shabalina et al. 2003) may be a direct result of the poly-A conserved region at the 3' end of the 3' UTR.

CHAPTER V

SUMMARY AND CONCLUSIONS

Sense-antisense transcripts have been shown to be abundant in many prokaryotic as well as eukaryotic organisms. The characterization of SATs in human, mouse and cow provide a detailed look into past transposable element activities leading to the formation of SATs and especially *trans* localized SATs. Each organism evaluated has its own unique set of transposable elements involved in the formation of SATs. The human genome has been shaped by the Alu repetitive sequence, which has played a significant role in the formation of *trans* SATs resulting from TE activity. The mouse *trans* SATs are mostly derived from B1_Mus transposable elements. Although the annotation of the *Bos taurus* genome is an ongoing effort, patterns similar to the human and mouse characterized SATs emerged in the SAT analysis such as various SABRs with a large percentage of the SABR sequence being composed of transposable element sequence. Additionally, this demonstrates the influence transposable element activity on the lengths of the 3' UTRs, 5' UTRs and coding regions. Overall there was a many-to-many relationship among SATs for all three organisms.

Although the proportion of all SABRs that contained TE sequence in the ≥95% SABR complementation dataset was not very high, the percent in the ≥80% SABR complementation group was significant. A significant proportion of *trans* SATs found in the ≥80% SABR complementation group originated from transposable elements or had transposable element cassettes in their sequence. A one-to-many relationship of sense to antisense transcripts in the ≥80% SABR group is expected due to the high level of activity of TEs in the evolution of mammalian genomes, and is a result of the insertion of sequences from particular TE families into the exon of one gene and the template strand

of the exons of many other genes.

The goals of this project were structural, functional and evolutionary characterization of SATs in human, mouse and cow. While previous research has focused on the *cis* regulatory mechanisms of *cis* positioned SATs, there may be some regulation of gene expression by *trans* SATs, but this is expected to be limited to the *trans* SATs that have maintained a high percent conservation in the SABR shared with other SATs allowing binding of complementary transcripts *in vivo*. However, this group of *trans* SATs may just be the result of young currently active transposable elements such as the human AluY subfamily that may have resulted in a biological function or created a SABR that has not had enough time for sequence divergence. There were not any conserved *trans* SATs identified across species suggesting that a prominent *trans* regulatory process for SATs with ≥95% SABR complementation may not be conserved across species and if there are some *trans* regulation of SATs, it may be unique to each species. However, this could be a result of the parameters set for detecting conserved SATs across organisms such as a high percent complementation in the SABR. The *cis* SATs seem to have greater potential for transcript regulation via transcriptional mechanisms colliding and stalling as explained by other researchers since some conserved *cis* SATs were identified. While *trans* SATs still have the potential to serve as a regulatory element, as is evidenced by the commercial use of antisense polygalactouronase gene expression regulation in the FlavrSavr tomato, SAT gene regulation created by TEs would have to be verified using expression data on a case by case basis.

A GO and PDB structure analysis provided insight into the functional classification of SATs containing transposable elements in the SABR as well as all

RefSeq sequences with transposable element sequences. The GO Slim analysis revealed significant enrichment for GO Slim terms in select sets of SATs and all RefSeq sequences containing a TE anywhere in the sequence. A full GO term analysis facilitates a better understanding of the functions and processes of these study sets, however, the full GO term analysis did not find enrichments such as localization of SATs gene products in the cytoplasm which was found using the GO Slim analysis. The enrichment for 'DNA repair' found in human and mouse in this research is supported by other researchers finding the same enrichment from a full GO term analysis of SATs. An additional full GO term enrichment for 'nuclease activity' was found in human SATs which describes genes that break the ester bond linkages within nucleic acids which may be a supporting process for 'DNA repair'. Additionally the full GO term 'response to DNA damage stimulus' was also found enriched in all SATs which may also be a process occurring in conjunction with 'DNA repair'.

The GO Slim analysis of all transcripts involved in SAT pairs supported and enhanced the full GO term analysis. The enrichment of broader GO terms reflected the full GO term analysis results of enrichment for 'DNA repair'. The enrichment of the GO Slim term 'metabolic process' is an upper level category of 'DNA repair' and 'citrate metabolic process'. Additionally, the GO Slim analysis revealed enrichment of terms in the molecular function category such as 'catalytic activity'. The full GO term analysis of SATs identified enrichment for 'catalytic activity' which is an upper level term for the GO Slim enrichment of SATs for 'nuclease activity' which is defined as "catalysis of the hydrolysis of ester linkages within nucleic acids" (http://geneontology.org) which is a lower level process of 'DNA repair'. Although reverse transcriptase which is encoded by an open reading frame in autonomous LINEs is considered a catalytic process, it is

unlikely that this catalytic activity enrichment is a result of transposons since the majority of TEs in SAT SABRs are SINEs which lack the coding potential for transposition enzymes. The GO Slim analysis did however find enrichments of terms for SATs not identified by the full GO term analysis. In the cellular component GO category. The GO Slim term ‘cytoplasm’ was identified as an enriched term for all SATs. Although previous researchers identified SATs as being nuclear localized, their analysis may have been based on a different percent complementarity criterion for SATs. Our GO Slim analysis finding SATs are localized in the cytoplasm does not agree with the enrichment ‘DNA repair’ since this process would be limited to nuclear localized DNA. This may be a result of incomplete GO terms for the current annotations which will improve as annotators continue to update genomes.

The effects of SATs on the production of a folded protein using PDB structure sequences showed that there are few sequences aligning to PDB structure sequences that also align to TE sequences or are part of SABRs for that same sequence. Although this analysis cannot confirm or deny that TE sequence or SABRs inhibit the folding properties of the protein of any individual transcript, it does show that there are few transcript regions occupied by TE sequence or SABRs that do overlap to a defined PDB structure sequence. Whether SABRs affect the folding properties of a protein is difficult to infer since most of the SABRs are located in the UTRs where the PDB structure sequences will not be found.

Transposable elements have been demonstrated to disrupt splicing patterns and create premature stop codons as well as introducing introns into the 3' end of a transcript. Premature stop codons leave some downstream splice junction complexes after the pioneer round of transcription. These transcripts are recognized by the NMD mechanism

as a target to breakdown the transcript into its components. SATs contribute to transcript regulation by providing the long mRNA duplex which is susceptible to hyper-editing of RNA by the ADAR enzymes. This long duplex region recruits the ADAR enzymes that have the potential to transform premature stop codons to tryptophan. More studies may reveal that the SABRs provide the dsRNA regions necessary to recruit the ADAR enzymes and edit premature stop codons into tryptophan in the SAT SABRs, however, there are not many opportunities for this since there are very few SAT pairs that have their SABRs in the coding sequence (CDS) region of the gene. Additionally, a read-through translation of a transcript could result if the SABR overlaps the transcripts stop codon and RNA editing alters it to a tryptophan if SATs do indeed play a role in creating dsRNAs precursors leading to RNA editing.

The estimation of evolutionary conservation in the 3' UTRs of human and mouse orthologous genes found increased conservation in the SABR when compared to the rest of the 3' UTR, which may suggest a functional constraint on the SABR. Additionally, the substitution rate in the coding sequence was lower than the rate found in the UTR. The last 10% of the 3' UTRs of human and mouse orthologous genes also demonstrated decreased substitution rates along the 5' end of the 3' UTRs where the SABRs of a small dataset of SATs reside. However, the percent GC was lower at the 5' ends of the 3' UTRs suggesting that the low substitution rates may be due to the low percent GC content rather than the actual conservation of SABR functionality since percent GC was found to be negatively correlated with sequence conservation in mammalian 3' UTRs.

The regions 2,000 bases upstream and 2,000 bases downstream of the 3' UTR suggest that the regions closest to the 3' UTR are more conserved that the regions father away from the 3' UTR. Additionally, the percent GC was highest in the upstream region

closest to the 3' UTR where the final exon for a gene generally resides. This suggests that GC percentage is highest where substitution rate is lowest in coding regions. However, in the 3' UTR, the region with the lowest substitution rate was also the region of lowest GC percentage. This may be due to the 3' end of the 3' UTR contains one or more conserved regulatory poly-A signals or protein suppressor regions which make this particular conserved region lower in percent GC.

REFERENCES

Alfano, G., Vitiello, C., Caccioppoli, C., Caramico, T., Carola, A., Szego, M., McInnes, R., Auricchio, A., and Banfi, S. 2005. Natural antisense transcripts associated with genes involved in eye development. *Hum. Mol. Genet.* **14:** 913-923.

Almeida, L., Silva, I., Silva Jr., W., Castro, J., Riggs, P., Carareto, C., and Amaral, M. 2007. The contribution of transposable elements to *Bos taurus* gene structure. *Gene* **390:** 180-189.

Athanasiadis, A., Rich, A., and Maas, S. 2004. Widespread A-to-I RNA editing of *Alu*-containing mRNAs in the human transcriptome. *PLoS Biology* **2:** e391.

Basilio, C., Wahba, A.J., Lengyel, P., Speyer, J.F., and Ochoa, S. 1962. Synthetic polynucleotides and the amino acid code. *Proc. Natl. Acad. Sci.* **48:** 613-616.

Bass, B. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71:** 817-846.

Bejerano, G. Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, J.K., and Haussler, D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441:** 87-90.

Benne, R., Van den Burg, J., Brakenhoff, J., Sloof, P., Van Boom, J., and Tromp, M. 1986. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46:** 819-826.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. 2007. GenBank. *Nucleic Acids Res.* **35:** D21-D25.

Blin-Wakkach, C., Lezon, F., Ghoul-Mazgar, S., Hotton, D., Montiero, S., Teilluaud, C., Pibouin, L., Orestes-Cardoso, S., Papagerakis, P., Macdougall, M., Robert, B., and Berdal, A. 2001. Endogenous *Msx1* antisense transcript: *In vivo* and *in vitro* evidences, structure, and potential involvement in skeleton development in mammals. *Proc. Natl. Acad. Sci.* **98:** 7336-7341.

Bostrom, K., Lauer, S.J., Poksay, K.S., Garcia, Z., Taylor, J.M., and Innerarity, T.L. 1989. Apolipoprotein B48 RNA editing in chimeric apolipoprotein EB mRNA. *J. Biol. Chem.* **264:** 15701-15708.

Britten, R. 1996. Cases of ancient mobile element DNA insertions that now affect gene regulation. *Mol. Phylogenet. Evol.* **5:** 13-17.

Britten, R. 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205:** 177-182.

Britten, R. 2004. Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc. Natl. Acad. Sci.* **101:** 16825-16830.

Brosius, J. 1999a. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107:** 209-238.

Brosius, J. 1999b. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238:** 115-134.

Burset, M., Seledtsov. I.A., and Solovyev, V.V. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28:** 364-4375.

Camblong, J., Iglesias, N., Fickentscher, C., Dieppois, G., and Stutz, F. 2007. Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in *S. cerevisiae. Cell* **131:** 706-17.

Cantrell, M.A., Filanoski, B.J., Ingermann, A.R., Olsson, K., DiLuglio, N., Lister, A., and Wichman, H.A. 2001. An ancient retrovirus-like element contains hot spots for SINE insertions. *Genetics* **158:** 769-777.

Castillo-Davis, C.I., and Hartl, D.L. 2003. GeneMerge – post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **7:** 891-892

Chao, H., Spicer, A.P. 2005. Natural antisense mRNAs to hyaluronan synthase 2 inhibit hyaluronan biosynthesis and cell proliferation. *J. Biol. Chem.* **280:** 27513-27522.

Chen, J., Sun, M., Kent, W., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R., and Rowley, J. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* **32:** 4812-4820.

Chen, J., Sun, M., Hurst, L., Carmichael, G., and Rowley, J. 2005. Genome-wide analysis of coordinate expression and evolution of human *cis*-encoded sense-antisense transcripts. *Trends Genet.* **21:** 326-329.

Cordaux, R., Udit, S., Batzer, M., and Feschotte, C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci.* **103:** 8101-8106.

Crampton, N., Bonass, W., Kirkham, J., Rivetti, C., and Thomson, N. 2006. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Res.* **34:** 5416-5425.

Curwen, V., Eyras, E., Andrews, T.D., Clarke, L., Mongin, E., Searle, S.M.J., and Clamp, M. 2004. The Ensembl automatic gene annotation system. *Genome Res.* **14:** 942-950.

Dai, L., Wang, X., Yao, X., Lu, Y., Ping, J., and He, J. 2007. Enhanced therapeutic effects of combined chemotherapeutic drugs and midkine antisense oligonucleotides for hepatocellular carcinoma. *World J. Gastroenterology* **13:** 1989-1994.

DeBarry, J., Ganki, E., McCarthy, E., and McDonald, J. 2006. The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. *Mol. Biol. Evol.* **23:** 479-481.

DeCerbo, J., and Carmichael, G. 2005. Retention and repression: fates of hyperedited RNAs in the nucleus. *Curr. Opin. Cell Biol.* **17:** 302-308.

Donner, H., Tönjes, R.R., Van der Auwera, B., Siegmund, T., Braun, J., Weets, I., Herwig, J., Kurth, R., Usadel, K.H., and Badenhoop, K. 1999. The presence or absence of a retroviral long terminal repeat influences the genetic risk for type 1 diabetes conferred by human leukocyte antigen DQ haplotypes. *J. Clin. Endocrinol. Metab.* **84:** 1404-1408.

Edgar, R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** doi:10.1186/1471-2105-5-113.

Faghihi, M., and Wahlestedt, C. 2006. RNA interference is not involved in natural antisense mediated regulation of gene expression in mammals. *Genome Biol.* **7:** doi: 10.1186/gb-2006-7-5-r38.

Frischmeyer, P.A., and Dietz, H.C. 1999. Nonsense-mediated mRNA decay in health in disease. *Hum. Mol. Genet.* **8:** 1893-1900.

Fu, L., Ma, W., and Benchimol, S. 1999. A translation repressor element resides in the 3' untranslated region of human p53 mRNA. *Oncogene* **18:** 6419-6424.

Galante, P., Vidal, D., de Souza, J., Camargo, A., and de Souza, S. 2007. Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol.* **8:** doi:10.1186/gb-2007-8-3-r40.

Gasior, S., Wakeman, T., Xu, B., and Deininger, P. 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **357:** 1383-1393.

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. *Nature Gene.* **25:** 25-29.

Gotea, V., and W. Makałowski, W. 2006. Do transposable elements really contribute to proteomes? *Trends Gene.* **22:** 260-267.

Hanahan, D., and Weinberg, R.A. 2000. The hallmarks of cancer. *Cell* **100:** 57-70.

Hartner, J., Schmittwolf, C., Kispert, A., Muller, M., Higuchi, M., and Seeburg, P. 2004. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J. Biol. Chem.* **279:** 4894-4902.

Häsler, J., and Strub, K. 2006. *Alu* elements as regulators of gene expression. *Nucleic Acids Res.* **34:** 5491-5497.

Hedges, D., and Deininger, P. 2007. Inviting instability: transposable elements, double-stranded breaks, and the maintenance of genome integrity. *Mutation Res.* **616:** 46-59.

Hedges, D.J., Callinan, P.A., Cordaux, R., Xing, J., Barnes, E., and Batzer, M.A. 2004. Differential *Alu* mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14:** 1068-1075.

Henz, S.R., Cumbie, J.S., Kasschau, K.D., Lohmann, J.U., Carrington, J.C., Weigel, D., and Schmid, M. 2007. Distinct expression patterns of natural antisense transcripts in *Arabidopsis. Plant Physiol.* **144:** 1247-1255.

Higuchi, M., Maas, S., Single, F., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P. 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406:** 78-81.

Hillman, R.T., Green, R.E., and Brenner, S.E. 2004. An unappreciated role for RNA surveillance. *Genome Biol.* **5:** doi:10.1186/gb-2004-5-2-r8.

Hu, J., Lutz, C., Wilusz, J., and Tian, B. 2005. Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. *RNA* **11:** 001-009.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P., Pagni, M., and Sigrist, C. 2006. The PROSITE database. *Nucleic Acids Res.* **34:** D227-D230.

IHGSC, International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Jackson, D., Pombo, A., and Iborra, F. 2000. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *The FASEB Journal* **14:** 242-254.

Jen, C., Michalopoulos, I., Westhead, D., and Meyer, P. 2005. Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol.* **6:** doi:10.1186/gb-2005-6-6-r51.

Jordan, I., Rogozin, I., Glazko, G., and Koonin, E. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19:** 68-72.

Jukes, T.H., and Cantor, C.R. 1969. Evolution of protein molecules. Pp. 21-123 in H. N. Munro, ed. *Mammalian protein metabolism*. Academic Press, New York.

Jurka, J. Krnjajic, M., Kapitonov, V.V., Stenger, J.E., and Kokhanyy, O. 2002. Active Alu elements are passed primarily through paternal germlines. *Theoretical Pop. Biol.* **61:** 519-530.

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

Kazazian, H. 2004. Mobile elements: Drivers of genome evolution. *Science* **303:** 1626-1632.

Kim, D., Kim, T., Walsh, T., Kobayashi, Y., Matise, T., Buyske, S., and Gabriel, A. 2004. Widespread RNA editing of embedded *Alu* elements in the human transcriptome. *Genome Res.* **14:** 1719-1725.

Kimelman, D., and Kirschner, M.W. 1989. An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in *Xenopus* oocytes. *Cell* **59:** 687-696.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111-120.

Kiyosawa, H., Yamanaka, I., Osato, N., and Kondo, S. RIKEN GER Group, GSL Members, Hayashizaki, Y., 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13:** 1324-1334.

Kiyosawa, H., Mise, N., Iwase, S., Hayashizaki, Y., and Abe, K. 2005. Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res.* **15:** 463-474.

Knee, R., and Murphy, P. 1997. Regulation of gene expression by natural antisense RNA transcripts. *Neurochem. Int.* **31:** 379-392.

Korber B. 2000. HIV Signature and sequence variation analysis. *Computational analysis of HIV molecular sequences*, Pp. 55-72. in A. G. Rodrigo and G. H. Learn, eds. Kluwer Academic Publishers, Dordrecht, Netherlands.

Kramer, C., Loros, J., Dunlap, J., and Crosthwaite, K. 2003. Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature* **421:** 948-952.

Kreigs, L., Churakov, G., Jurka, J., Brosius, L., and Schmitz, J. 2007. Evolutionary history of 7SL RNA-derived SINEs in supraprimates. *Trends Genet.* **23:** 158-161.

Kumar, M., and Carmichael, G. 1998. Antisense RNA: Function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. and Mol. Biol. Rev.* **62:** 1415-1434.

Lehner, B., Williams, G., Campbell, D., and Sanderson, D. 2002. Antisense transcripts in the human genome. *Trends Genet.* **18:** 63-64.

Leon, A. 2004. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *J. Clin. Psychiatry* **64:** 1511-1514.

Lev-Maor, G., Sorek, R., Levanon, E.Y., Paz, N., Eisenberg, E., and Ast, G. 2007. RNA-editing-mediated exon evolution. *Genome Biol.* **8:** doi:10.1186/gb-2007-8-2-r29.

Liang, H. Zhou, W. and Landweber, L.F. 2006. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. *Nucleic Acids Res.* **34:** W381-W384; doi:10.1093/nar/gk1272.

Li, Y., and Su, B. 2006. No accelerated evolution of 3' UTR region in human for brain-expressed genes. *Gene* **383:** 38-42.

Li, Y., Qin, L., Guo, Z., Liu, L., Xu, H., Hao, P., Su, J., Shi, Y., He, W., and Li, Y. 2006. *In silico* discovery of human natural antisense transcripts. *BMC Bioinformatics* **7:** 10.1186/1471-2105-7-18.

Lipman, D. 1997. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* **25:** 3580-3583.

Loning, F., Ma, W., and Benchimol, S. 1999. A translation repressor element resides in the 3' untranslated region of human p53 mRNA. *Oncogene* **18:** 6419-6424.

Lorenc, A., and Makalowski, W. 2003. Transposable elements and vertebrate diversity. *Genetica* **118:** 183-191.

Maglott, D., Ostell, J., Pruitt, K., and Tatusova, T. 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **35:** D26-D31.

Makalowski, W., Mitchell, G., and Labuda, D. 1994. Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* **10:** 188-193.

Makalowski, W., Zhang, J., and Boguski, M.S. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6:** 846-857.

Maquat, L.E., 2005. Cell science at a glance. *J. Cell Sci.* **118:** 1773-1776.

Medghalchi, S.M., Frischmeyer, P.A., Mendell, J.T., Kelly, A.G., Lawler, A.M., and Dietz, H.C. 2001. *Rent1*, a *trans*-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum. Mol. Genet.* **10:** 99-105.

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

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. and Evol.* **3:** 418-426.

Nekrutenko, A., and Li, W. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* **17:** 619-621.

Nikaido, M., Rooney, A.P., and Okada, N. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci.* **18:** 10261-10266.

Nishikura, K., Yoo, C., Kim, U., Murray, J.M., Estes, P.A., Cash, F.E., and Liebhaber, S.A. 1991. Substrate specificity of the dsRNA unwinding/modifying activity. *EMBO J.* **10:** 3523-3532.

Ohhata, T., Hoki, Y., Sasaki, H., and Sado, T. 2008. Crucial role of antisense transcription across the Xist promoter in Tsix-mediated Xist chromatin modification. *Development* **135:** 227-235.

Osato, N., Suzuki, Y., Ikeo, K., and Gojobori, T. 2007. Transcriptional interference in *cis* natural antisense transcripts of human and mice. *Genetics* **176:** 1299-1306.

Pearson, W.R., and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444-2448.

Peters, N.T., Rohrbach, J.A., Zalewski, B.A., Byrkett, C.M., and Vaughn, J.C. 2003. RNA editing and regulation of *Drosophila 4f-rnp* expression by *sas-10* antisense readthrough mRNA transcripts. *RNA* **9:** 698-710.

Piriyapongsa, J., Polavarapu, N., Borodovsky, M., and McDonald, J. 2007. Exonization of the LTR transposable elements in the human genome. *BMC Bioinformatics* **8:** doi:10.1186/1471-2164-8-291.

Prescott, E., and Proudfoot, N. 2002. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci.* **99:** 8796-8801.

Pruitt, K., Tatusova, T., and Maglott, D. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33:** D501-D504; doi:10.1093/nar/gki025.

Reuter, S.M., Dawson, T.R., and Emerson, R.B. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399:** 75-80.

RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564-1556.

Rouzic, A., Boutin, T., and Capy, P. 2007. Long-term evolution of transposable elements *Proc. Natl. Acad. Sci.* **104:** 19375-19380.

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A., and Ast, G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals *Alu*'s unique role in shaping the human transcriptome. *Genome Biol.* **8:** doi:10.1186/gb-2007-8-6-r127.

Sen, L., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P., Dyer, M., Cordaux, R., Liang, P., and Batzer, M. 2006. Human genomic deletions mediated by recombination between *Alu* elements. *Am. J. Hum. Genet.* **79:** 41-53.

Shabalina, S. A., Ogurtsov, A.Y., Lipman, D.J., and Kondrashov, A.S. 2003. Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3' UTRs. *Nucleic Acids Res*. **31:** 5433-5439.

Shendure, J., and Church, G. 2002. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.* **3:** 0044.1-0044.14.

Simons, R.W. 1988. Naturally occurring antisense RNA control – a brief review. *Gene* **72:** 35-44.

Sleutels, F., Zwart, R., and Barlow, D. 2002. The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810-813.

Smit, A. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743-748.

Smit. A.F., Tóth, G., Riggs, A.D., and Jurka, J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive elements. *J. Mol. Biol.* **246:** 401-417.

Smith, C., Watson, C., Ray, J., Bird, C., Morris, P., Schuch, W., and Grierson, D. 1988. Antisense RNA inhibition of polygalactouronase gene expression in transgenic tomatoes. *Nature* **334:** 724-726.

Sorek, R., Ast, G., and Graur, G. 2002. *Alu*-containing exons are alternatively spliced. *Genome Res.* **12:** 1060-1067.

Standart, N., Dale, M., Stewart, E., and Hunt, T. 1990. Maternal mRNA from clam oocytes can be specifically unmasked *in vitro* by antisense RNA complementary to the 3'-untranslated region. *Genes & Dev.* **4:** 2157-2168.

Steigele, S., and Neiselt, K. 2005. Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes. *Nucleic Acids Res.* **33:** 5034-5044.

Tang, W., Kinken, K., and Newton, R.J. 2005. Inducible antisense-mediated post-transcriptional gene silencing in transgenic pine cells using green fluorescence protein as a visual marker. *Plant and Cell Physiology* **46:** 1255-1263.

Tatusova, T.A., and Madden, T.L. 1999. Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. *FEMS* **174:** 247-250.

Terryn, N., and Rouze, P. 2000. The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci.* **5:** 394-396.

Thornburg, B., Gotea, V., and Makalowski, W. 2006. Transposable elements as a significant source of transcription regulating signals. *Gene* **365:** 104-110.

Tufarelli, C., Stanley, J.A., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., and Higgs D.R. 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.* **34:** 157-165.

van de Lagemaat, L., Landry, J., Mager, D., and Medstrand, P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19:** 530-536.

Vanhee-Brossollet, C., and Vaquero, C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211:** doi:10.1016/S0378-1119(98)00093-6 .

Vardy, L., and Orr-Weaver, T.L. 2007. Regulating translation of maternal messages: multiple repression mechanisms. *Trends Biol.* **17:** 547-554.

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R., and Makalowska, I. 2004. Mammalian Overlapping Genes: The Comparative Perspective. *Genome Res.* **14:** 280-286.

Wacheck, V., and Zangemeister-Wittke, U. 2006. Antisense molecules for targeted cancer therapy. *Oncology Hematology* **59:** 65-73.

Wagner, E.G., and Simons, R.W. 1994. Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* **48:** 713-742.

Wallace, M.R., Anderson, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W., and Collins, F.S. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353:** 864-866.

Wang, H., Chua, N., and Wang, X. 2006. Prediction of *trans*-antisense transcripts in *Arabidopsis thaliana*. *Genome Biol.* **7:** doi:10.1186/gb-2006-7-10-r92.

Wang, X., Gaasterland, T., and Chua, N. 2005. Genome-wide prediction and identification of *cis*-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biol.* **6:** doi:10.1186/gb-2005-6-4-r30.

Wheeler, D., Church, C., Federhen, S., Lash, A., Madden, T., Pontius, J., Schuler, G., Schrimi, L., Sequeira, E., Tatusova, T., and Wagner, L. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31:** 28-33.

Woolf, T.M., Chase, J.M., and Stinchcomb, D.T. 1995. Towards the therapeutic editing of mutated RNA sequences. *Proc. Natl. Acad. Sci.* **92:** 8298-8302.

Wu, M., Li, L., and Sun, Z. 2007. Transposable element fragments in protein-coding regions and their contribution to human functional proteins. *Gene* **401:** 165-171.

Xuezhong, C., Hagedorn, C.H., and Cullen B.R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10:** 1957-1966.

Yelin, R., Dahary, D., Sorek, R., Levanon, E., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K., and Rotman, G. 2003. Widespread occurrence of antisense transcription in the human genome. *Nature* **21:** 379-386.

Yin, Y., Zhao, Y., Wang, J., Liu, C., Chen, S., Chen, R., and Zhao, H. 2007. antiCODE: a natural sense-antisense transcripts database. *BMC Bioinformatics* **8:** doi:10.1186/1471-2105-8-319.

Zhang, Y., Liu, S., Liu, Q., and Wei, L. 2006. Genome-wide *in silico* identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. *Nucleic Acids Res.* **12:** 3465-3475.

APPENDIX A

AVERAGE PERCENT COVERAGE OF HUMAN SABRS BY INCLUDING ALL ALTERNATIVE TRANSCRIPTS TRANSPOSABLE ELEMENTS (TEs)

(≥80% SABR COMPLEMENTATION)

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
ALU (S)	273	97	84	369,230	1	369,230
CAM1_GG (L)	388	9	5	3	1	2
CHARLIE1 (D)	384	19	3	1	0	1
CHARLIE1A (D)	368	14	4	2	0	2
CHARLIE1B (D)	353	15	10	1	0	1
CHARLIE26a (D)	372	14	16	2	0	2
CHARLIE3 (D)	140	100	5	2	0	2
CHARLIE4 (D)	424	18	4	1	0	1
CHARLIE5 (D)	364	11	1	2	0	2
CHARLIE8A (D)	371	9	10	2	0	2
CHESHIRE_A (D)	228	15	12	2	0	2
CHESHIRE_B (D)	374	20	22	1	0	1
ERV11_MD_I (E)	1,246	10	1	1	1	0
ERV33_MD_I (E)	580	16	1	1	1	0
ERV36_MD_LTR(E)	1,003	10	15	1	0	1
GGERVK1 (E)	569	15	1	1	1	0
GGLTR8A_LTR (E)	949	7	7	1	1	0
GOLEM_B (D)	336	8	2	2	0	2
HARLEQUIN (E)	592	9	1	1	0	1
HERV49I (E)	358	13	1	2	0	2
HERV70_I (E)	203	92	2	61	1	61
HERVIP10FH (E)	592	29	3	1	0	1
HERVK14CI (E)	381	18	1	1	0	1
HERVL (E)	656	100	12	1	0	1
L1 (L)	141	90	2	10	0	10
L1-2a_MD (L)	401	11	2	1	1	0
L1A_OC (L)	1,136	7	1	1	1	0
L1HS (L)	269	98	29	12	0	12
L1M2_5 (L)	323	9	1	1	0	1
L1M3C_5 (L)	3,561	2	4	1	1	0
L1M4B (L)	355	11	1	1	0	1
L1MA10 (L)	374	17	7	3	0	3
L1MA1 (L)	344	12	4	6	0	6
L1MA2 (L)	343	13	4	1	0	1
L1MA3 (L)	350	9	3	1	0	1
L1MA4A (L)	370	15	5	2	0	2
L1MA5A (L)	339	10	3	4	0	4
L1MA7 (L)	354	12	4	1	0	1
L1MB1 (L)	346	10	4	5	0	5
L1MB2 (L)	297	17	6	4	0	4
L1MB3 (L)	341	15	6	3	0	3
L1MB5 (L)	356	10	4	2	0	2
L1MB8 (L)	344	10	4	3	0	3
L1MC4 (L)	356	11	1	5	0	5
L1MD1 (L)	351	10	4	2	0	2
L1MD2 (L)	354	14	4	1	0	1
L1MDA_5 (L)	339	12	1	15	0	15
L1MDB_5 (L)	366	13	3	1	0	1
L1ME4A (L)	382	16	7	2	0	2
L1ME5 (L)	386	13	9	4	0	4
L1MEC_5 (L)	398	15	2	1	0	1
L1ME_ORF2 (L)	360	13	1	19	0	19
L1PA10 (L)	228	99	25	11	0	11
L1PA11 (L)	104	100	11	1	0	1
L1PA13 (L)	307	30	9	11	0	11
L1PA15 (L)	409	99	44	3	0	3
L1PA2 (L)	297	93	31	12	0	12

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
L1PA3 (L)	297	94	30	10	0	10
L1PA4 (L)	237	92	24	20	0	20
L1PA5 (L)	162	93	16	5	0	5
L1PA7 (L)	329	96	35	17	0	17
L1PA7_5 (L)	788	100	46	1	0	1
L1PA8 (L)	345	9	3	3	0	3
L1PB1 (L)	159	85	15	1	0	1
L1PB2 (L)	348	9	3	1	0	1
L1PB4 (L)	330	12	5	9	0	9
L1P_MA2 (L)	389	15	1	3	0	2
L1PREC1 (L)	288	99	4	4	0	4
L1PREC2 (L)	272	29	1	8	0	8
L2A (L)	531	27	2	5	2	3
L2B_ME (L)	400	14	16	1	0	1
LTR12 (E)	207	99	31	1	0	1
LTR13 (E)	334	100	33	1	0	1
LTR14B (E)	390	100	66	2	0	2
LTR14C (E)	370	9	6	2	0	2
LTR17 (E)	326	100	42	1	0	1
LTR19B (E)	583	99	100	1	0	1
LTR2B (E)	326	14	9	35	0	35
LTR3 (E)	125	98	28	1	0	1
LTR35_MD (E)	440	20	11	1	1	0
LTR38C (E)	242	19	7	2	0	2
LTR5 (E)	281	100	29	1	0	1
LTR66 (E)	367	11	7	6	0	6
LTR7B (E)	271	47	17	10	0	10
LTRIS_Mm_LTR (E)	106	31	7	1	1	0
MADE1 (D)	340	10	42	2	0	2
MER11B (D)	1,090	99	100	1	0	1
MER11D (D)	632	100	73	1	0	1
MER1A (D)	459	62	53	2	0	2
MER1B (D)	186	96	53	1	0	1
MER2 (D)	355	20	18	29	0	29
MER2B (D)	357	12	12	21	0	21
MER30 (D)	327	66	61	3	1	2
MER31I (D)	330	13	1	2	0	2
MER3 (D)	357	10	16	7	0	7
MER34B (D)	1,545	23	66	1	1	0
MER34B_I (D)	390	19	1	2	0	2
MER41A (D)	362	15	10	1	0	1
MER44B (D)	373	14	10	2	0	2
MER44D (D)	129	100	19	1	0	1
MER45R (D)	162	31	3	1	1	0
MER51I (D)	348	13	1	6	0	6
MER5A1 (D)	356	2	62	1	1	0
MER5B (D)	367	12	24	2	0	2
MER63B (D)	395	22	20	1	0	1
MER65I (D)	360	14	1	2	0	2
MER66C (D)	344	8	5	4	0	4
MER77 (D)	363	10	6	1	0	1
MER85 (D)	284	76	91	9	1	8
MER9 (D)	498	95	92	2	0	2
MER96B (D)	324	11	9	3	0	3
MIR3 (S)	366	17	32	2	0	2
MIR (S)	756	21	30	10	3	7
MIRb (S)	346	12	16	11	1	10
MLT1A1 (E)	351	11	9	5	0	5
MLT1A (E)	333	14	12	1	0	1
MLT1B (E)	369	15	14	3	0	3
MLT1C1 (E)	383	14	12	1	0	1
MLT1C (E)	348	13	10	1	0	1
MLT1E (E)	210	30	10	1	0	1
MLT1F (E)	358	18	11	6	0	6
MLT1H1 (E)	935	4	8	1	0	1
MLT1H (E)	348	13	8	1	0	1
MLT1K (E)	352	14	8	1	0	1
MLT1L (E)	354	10	6	1	0	1
MLT2B2 (E)	318	11	7	2	0	2
MSTA1 (E)	161	25	8	1	0	1

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
MSTA (E)	263	48	21	7	0	7
MSTB1 (E)	182	96	41	1	0	1
MSTC (E)	380	14	12	2	0	2
PABL_AI (E)	279	14	1	3	0	3
PB1 (S)	336	10	25	1	0	1
RLTR11A2_LTR (E)	164	47	14	1	1	0
RLTR24_MM (E)	234	16	6	1	0	1
RNLTR17_I (E)	238	71	2	1	1	0
SVA2 (S)	236	92	99	2	0	2
SVA (S)	223	76	8	118	0	118
THE1A (E)	338	76	54	9	0	9
THE1B (E)	183	100	51	3	0	3
THE1BR (E)	453	43	16	6	0	6
THE1C (E)	342	12	11	7	0	7
THE1D (E)	293	23	13	5	0	5
TIGGER7 (D)	340	14	2	10	0	10
TOTALS	**273**	**97**	**84**	**369,960**	**25**	**369,935**

SINE (S)
LINE (L)
DNA transposon (D)
Endogenous retrovirus (E)

APPENDIX B

AVERAGE PERCENT COVERAGE OF MOUSE SABRS INCLUDING ALL ALTERNATIVE TRANSCRIPTS BY TRANSPOSABLE ELEMENTS (TEs) (≥80% SABR COMPLEMENTATION)

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
B1 (S)	137	95	92	179	1	178
B1F (S)	181	84	97	201	3	198
B1_Mm (S)	158	88	93	4,822	0	4,822
B1_Mur1 (S)	166	94	94	44	1	43
B1_Mur2 (S)	154	93	92	96	1	95
B1_Mur3 (S)	151	91	91	187	0	187
B1_Mur4 (S)	148	94	92	711	1	710
B1_Mus1 (S)	160	89	94	11,806	0	11,806
B1_Mus2 (S)	164	88	96	8,029	1	8,028
B1_Rn (S)	154	90	92	2,602	1	2,601
B2 (S)	187	94	84	173	0	173
B2_Mm1a (S)	190	94	91	744	0	744
B2_Mm1t (S)	189	94	91	411	1	410
B2_Mm2 (S)	178	95	85	857	0	857
B2_Rat1 (S)	180	95	90	287	0	287
B2_Rat2 (S)	182	94	90	1,599	0	1,599
B2_Rat3 (S)	171	95	86	233	0	233
B2_Rat4 (S)	179	95	89	701	1	700
B2_Rn2 (S)	128	94	64	123	0	123
B2_Rn (S)	221	17	22	1	0	1
B3 (S)	306	75	55	6	2	4
B3A (S)	717	35	86	2	2	0
BC1_Ma (S)	2,304	2	29	1	1	0
BGLII_A_LTR (E)	186	21	8	1	0	1
CR1-G (L)	2,170	2	1	1	1	0
ERV24_MD_I (E)	222	23	1	1	1	0
ERV46_MD_I (E)	1,401	3	1	1	1	0
ERVB1_3-I_RN (E)	246	20	1	1	0	1
ERVB2_1-I_MM (E)	1,028	13	1	2	1	1
ERVB4_1-LTR_MM (E)	424	99	80	1	0	1
ERVB5_1-I_MM (E)	183	17	1	1	0	1
GGERVK10 (E)	1,150	3	1	1	1	0
Harbinger-2_XT (D)	249	14	1	1	1	0
HERVR (E)	323	37	2	1	0	1
HITCHCOCK2_LTR(E)	383	17	10	1	1	0
IAPEZI (E)	136	100	2	1	0	1
IAPLTR1a_MM (E)	141	88	35	1	0	1
IAPLTR1_Mm_LTR(E)	284	96	72	5	0	5
IAPLTR2b_LTR (E)	284	84	72	2	0	2
ID_Rn1 (S)	938	6	55	1	1	0
L1MC4 (L)	270	21	2	1	1	0
L1_MM (L)	246	96	4	163	0	163
L1PA14_5 (L)	546	12	3	3	0	3
L1R2_RN (L)	104	91	13	1	0	1
L2-2_ME (L)	195	36	3	1	1	0
L2A (L)	105	80	3	1	1	0
LTR12E (E)	590	7	4	1	1	0
LTR16A1 (E)	1,234	12	34	1	0	1
LTR6_MD (E)	1,150	5	7	1	1	0
LTRIS3 (E)	144	99	27	2	0	2
LX (L)	198	94	18	39	0	39
LX3 (L)	173	98	17	1	0	1
MEN (S)	187	85	54	798	1	798
MER106B (D)	273	44	12	1	1	0
MER20 (D)	1,864	5	51	1	1	0
MER21C (E)	687	8	6	1	1	0
MER52D (E)	829	7	3	1	1	0

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
MERVL_LTR (D)	144	99	29	7	0	7
MLT2F (E)	1,485	5	10	1	1	0
MT2B (E)	202	58	18	6	0	6
MT2B2_LTR (E)	130	97	24	3	0	3
MT2C (E)	127	92	31	19	0	19
MTA (E)	335	98	82	29	0	29
MTAI (E)	106	100	10	3	0	3
MTA_Mm_LTR (E)	225	98	55	51	0	51
MTB_Mm_LTR (E)	170	86	35	7	0	7
MTB_Rn_LTR (E)	240	98	58	45	0	45
MTC (E)	185	100	44	1	0	1
MTEb_LTR (E)	346	25	24	1	1	0
ORR1A0 (E)	122	100	35	5	0	5
ORR1A2_LTR (E)	183	99	55	18	0	18
ORR1A3_LTR (E)	140	99	42	2	0	2
ORR1A4_LTR (E)	144	95	41	8	0	8
ORR1AI (E)	189	93	9	1	0	1
ORR1A_Rn_LTR (E)	222	98	66	17	0	17
ORR1B1 (E)	128	100	33	1	0	1
ORR1B2 (E)	88	98	23	1	0	1
ORR1BI (E)	186	76	10	218	0	218
ORR1D1_LTR (E)	976	35	94	1	1	0
ORR1D (E)	201	20	10	1	0	1
PB1D10 (S)	521	17	44	6	1	5
PB1D9 (S)	599	49	88	4	2	2
RLTR10C (E)	114	100	26	1	0	1
RLTR10E (E)	101	95	30	1	0	1
RLTR13D (E)	114	96	11	1	0	1
RLTR13D3 (E)	119	100	11	1	0	1
RLTR16 (E)	2,304	4	18	1	1	0
RLTR17_MM (E)	218	17	4	5	0	5
RLTR25_MM (E)	180	22	5	4	0	4
RLTR35_MM (E)	196	30	12	6	0	6
RLTR44C_LTR (E)	523	98	95	1	0	1
RLTR6I_MM (E)	232	21	1	1	0	1
RLTR9E (E)	304	100	73	1	0	1
RLTRETN_MM (E)	221	99	63	1	0	1
RMER17D2_MM (E)	196	21	5	1	0	1
RMER19 (E)	226	19	6	2	0	2
RMER19A_LTR (E)	119	66	10	1	0	1
RNLTR7_I (E)	198	21	1	8	0	8
RSINE1 (S)	919	14	85	1	1	0
RSINE2 (S)	181	90	45	569	0	569
RSINE2A (S)	187	88	49	208	1	208
SINEB1_MU (S)	160	81	95	8	0	8
Tigger2f (D)	323	14	1	1	0	1
URR1 (D)	144	96	57	36	0	36
URR1B (D)	126	98	55	5	0	5
TOTALS	**166**	**89**	**90**	**36,177**	**43**	**36,134**

SINE (S)
LINE (L)
DNA transposon (D)
Endogenous retrovirus (E)

APPENDIX C

AVERAGE PERCENT COVERAGE OF COW SABRS BY INCLUDING ALL ALTERNATIVE TRANSCRIPTS TRANSPOSABLE ELEMENTS (TEs) (≥80% SABR COMPLEMENTATION)

TE	Average SABR length	Average % SABR coverage by TE	Average % TE found in SABR	Number of SABRs	*cis*	*trans*
ARMER1 (Repeat)	140	94	42	13	0	13
ART2A (S)	268	98	48	56	0	56
ART2B_BT (S)	99	100	20	1	0	1
BCS (S)	118	92	36	26	0	26
BDDF2 (L)	620	33	5	37	0	37
BOV2 (S)	306	95	50	460	0	460
BOVA2 (S)	161	96	56	1,777	1	1,777
BOVB (S)	370	89	9	41	0	41
Bov-tA1 (S)	146	98	62	270	0	270
BOVTA (S)	146	97	67	1,313	0	1,313
Bov-tA2 (S)	137	97	62	589	0	589
Bov-tA3 (S)	151	96	68	427	0	427
BTALUL1 (S)	179	97	54	77	0	77
BTCS (S)	304	92	93	80	0	80
L1_BT (L)	225	90	17	23	0	23
L1_Felid (L)	312	9	3	1	0	1
L1MB7 (L)	344	15	6	2	0	2
L1MC4 (L)	358	16	2	1	0	1
L1MC5 (L)	359	18	3	1	0	1
L1P_MA2 (L)	124	99	2	1	0	1
LINE_CH (L)	172	64	6	2	0	2
MAR1b_MD (S)	157	92	53	2	0	2
MER2 (D)	165	19	9	1	0	1
RNLTR17_I (E)	242	33	1	1	0	1
RNLTR21_I (E)	445	13	1	1	1	0
RTE1_LA (L)	326	99	10	1	0	1
TOTALS	**175**	**96**	**59**	**5,205**	**2**	**5,203**

SINE (S)
LINE (L)
DNA transposon (D)
Endogenous retrovirus (E)

APPENDIX D

EXCERPTS FROM THE GO SLIM MAP FILE FROM THE EBI WEBSITE

! Mapping of GO terms to GO Slim Terms 24-JUL-2007
! Slim Term Source: GOA Database @ EBI
! Version: $
! Date: $Date: $
! Format: GO_id <tab> GO_slim_id
GO:0000049 GO:0003676
GO:0000182 GO:0003676
GO:0000217 GO:0003676
GO:0000339 GO:0003676
GO:0000340 GO:0003676
GO:0000341 GO:0003676
GO:0000342 GO:0003676
GO:0000400 GO:0003676
GO:0000401 GO:0003676
GO:0000402 GO:0003676
GO:0000403 GO:0003676
GO:0000404 GO:0003676
GO:0000405 GO:0003676
GO:0000406 GO:0003676
GO:0000739 GO:0003676
GO:0000900 GO:0003676
GO:0003676 GO:0003676
GO:0003677 GO:0003676
GO:0003680 GO:0003676
GO:0003681 GO:0003676
GO:0003684 GO:0003676
GO:0003688 GO:0003676
GO:0000146 GO:0003774
GO:0003774 GO:0003774
GO:0003777 GO:0003774
GO:0008569 GO:0003774
GO:0008574 GO:0003774
GO:0060001 GO:0003774
GO:0060002 GO:0003774
GO:0000009 GO:0003824
GO:0000010 GO:0003824
GO:0000014 GO:0003824
GO:0000016 GO:0003824
GO:0000026 GO:0003824
GO:0000030 GO:0003824

APPENDIX E

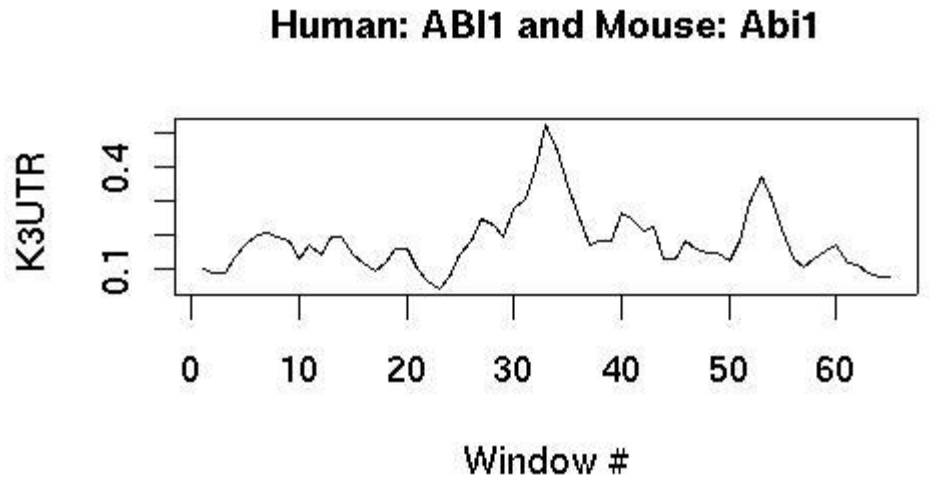
EXCERPT FROM THE GO.terms_and_ids FILE

PROVIDED BY THE GENE ONTOLOGY

```
!version: $Revision: 1.521 $
!date: $Date: 2007/08/15 10:47:59 $
!
! GO IDs (primary only) and text strings
! GO:0000000 [tab] text string  [tab] F|P|C
! where F = molecular function, P = biological process, C = cellular component
!
GO:0000001      mitochondrion inheritance       P
GO:0000002      mitochondrial genome maintenance        P
GO:0000003      reproduction    P
GO:0000005      ribosomal chaperone activity    F
GO:0000006      high affinity zinc uptake transmembrane transporter activity  F
GO:0000007      low-affinity zinc ion transmembrane transporter activity   F
GO:0000008      thioredoxin     F
GO:0000009      alpha-1,6-mannosyltransferase activity  F
GO:0000010      trans-hexaprenyltranstransferase activity       F
GO:0000011      vacuole inheritance     P
GO:0000012      single strand break repair      P
GO:0000014      single-stranded DNA specific endodeoxyribonuclease activity     F
GO:0000015      phosphopyruvate hydratase complex       C
GO:0000016      lactase activity        F
GO:0000017      alpha-glucoside transport       P
GO:0000018      regulation of DNA recombination P
GO:0000019      regulation of mitotic recombination     P
GO:0000020      negative regulation of recombination within rDNA repeats   P
GO:0000022      mitotic spindle elongation      P
GO:0000023      maltose metabolic process       P
GO:0000024      maltose biosynthetic process    P
GO:0000025      maltose catabolic process       P
GO:0000026      alpha-1,2-mannosyltransferase activity  F
GO:0000027      ribosomal large subunit assembly and maintenance      P
GO:0000028      ribosomal small subunit assembly and maintenance      P
GO:0000030      mannosyltransferase activity    F
GO:0000031      mannosylphosphate transferase activity  F
GO:0000032      cell wall mannoprotein biosynthetic process     P
GO:0000033      alpha-1,3-mannosyltransferase activity  F
GO:0000034      adenine deaminase activity      F
GO:0000035      acyl binding    F
GO:0000036      acyl carrier activity   F
GO:0000038      very-long-chain fatty acid metabolic process    P
```

APPENDIX F

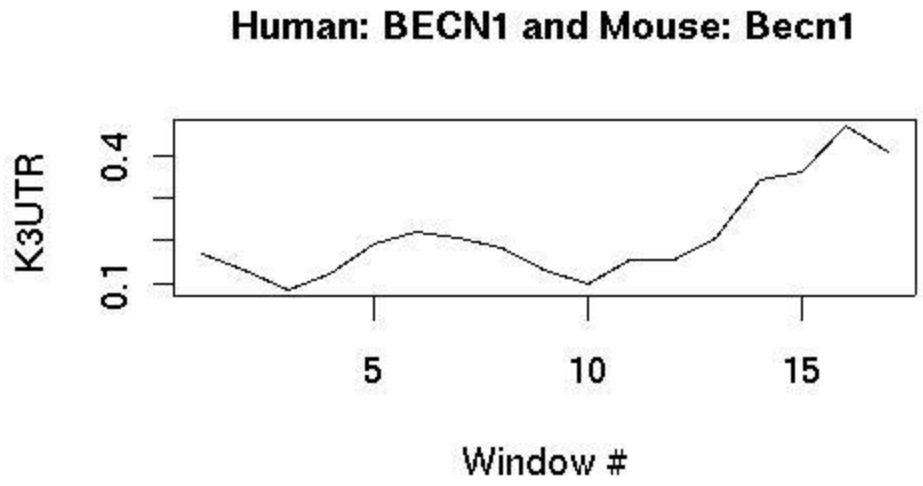
SEQUENCE ALIGNMENTS OF SELECTED HUMAN AND MOUSE HOMOLOGOUS GENES USED IN THE SLIDING WINDOW ANALYSIS




```
MUSCLE (3.6) multiple sequence alignment

ABI1      -TTTTTTTTTTTCTTTTGAAGTAGATTCTTATTACTCAGTCATACTGTGGGACTATTATG
Abi1      TTTTTCTTTTTTCTTTTCATGTAGG---TTATTACTCCGTCATACTGTGGGATTA-TATG
           **** *********** * ****    ********* ************** ** ****

ABI1      GTTAACAGAACTGTCTTAATATGTTTTAAAATGTGCCCATATTTTC-AGAACATGCTGTT
Abi1      GTTAACAGAATTGTTTTAATG---TTTAAAATGTGCCCATATTTTCAAGGACA---TGTT
          ********** *** *****    ********************** ** ***   ****

ABI1      TTATTGGTA-AATTGAATGTCTACCTGTAAGCATAAATCTTTGAGGCAGTTTATGTATTG
Abi1      TTATTGGTATATTTGGATGTCTACCTGTAAGCATAAATTTTGGAGGCAGTTCAAACATTG
          ********* * *** ********************** ** ********* *   ****

ABI1      CTGAATAGCAATTTATACAAGAAGCTGTCCATAACTGATTATGCTTATGTACTTACTTAC
Abi1      CTGAGCAGCAGTTTATA--------TGCCTATAATTGATTATGCATATG----TACTCAC
          ****  **** ******        ** * **** ********* ****    **** **

ABI1      ACATTTTTAACTTTATGACCAGCCTAAATATTCTGGGGGAAGTGGGGTATAATATTTAAC
Abi1      ACCTTGCTAAGTTTATGACCAGCCTAAAACTTCTGGGG---ATTGGGTATTATGTTTAAC
          ** **  *** *****************  ********    * ****** ** ******

ABI1      GAATCATGATTCAGATTGTACCATTACATGTTTCAGTGCAGCATGGTTACTAACGCTATG
Abi1      AAATCATGGTTCAGAATGCACCATTACATGTTTCAGTGCAGCATGGTCACTAACATTGTG
           ******* ****** ** ***************************** ******  * **

ABI1      TCAGACTAATATT-AAAATCAGAAAATTTAAATGCTGGTGCTGGTCAGACTTTTTTTGTT
Abi1      TCAGACTAATAGGAAAAAACAGAAAACGTCAATGCTGGTGCTGGTCATACTTTTGGT-TT
          ***********   **** *******  * ***************** ******  * **
```

```
ABI1      AGATTCTCTCATTTAAAAAAAATACTGTTTGTTTAAAGCATGCATAAAAATTTATGTATT
Abi1      CAATTCTCAT-TTTTAAAAAAATACTG--TGTTTAAAGCATGCATAAATTTTTATGTATT
            ******   *** ************  ******************  **********
ABI1      GAAATATACTTAAAAATTCAAGATGCTTCCCATTTGTGTAATATTTACCTGGAGGACTCG
Abi1      GAAATATACTTAACAATTCAAGATGCTTCCAATTTGTGTAACGATTATCTGGAGTACTCA
          ************ *************** **********   *** ****** ****

ABI1      TACTTAGGTGTCTTAACGTGAATTGAGTCTCC-AAGGTCTCCATGTGAAACAAGCAAAAA
Abi1      TAC--------------------TTGAGTCTCCTAAGCTCTCCATGTGAAATGA-------
          ***                    ********** *** *************  *

ABI1      GAGAATTATCTGTAATGTTGTAATTTGTACCTAAGTTTTTTAATGAGTGAAATTTGCATT
Abi1      -AAGACTATCTGTAATGTTGTAATTTGTATCTAAGTTTTTTAATGAGTGAAATTTGCATT
           *  * ********************** ******************************

ABI1      ATAAACTTTTTCCATTCATAAATACATAAGTGAACCAAAGGTTTTTGTCCTTTCCTTCAC
Abi1      ATAAA-TTTTTCCATTCATAAATACATAAGTGAACCAAAGGATTTTGGCCTCTCCTTTAC
          ***** ************************************ ***** *** ***** **

ABI1      TGATTTGCTTTAAAAA-------------AAATAAAAG--------------------AT
Abi1      TGGTTTGCTTTAATTATGTATGCTAGTGCATATGCATGCACACCCTCCCACCCCTTTAAG
          ** **********  *             * **  * *                     *

ABI1      AATGATTTATTGCAGAATTATGATTCTATTTTCTCAATATGTTAACTTGGAAAAAAATTT
Abi1      AAAGATTTATTGCAGAATTATGGTTCTATTTTCTCAATGCATCAGTTTTGAAAATATTTT
          ** ******************* ***************   * *  ** ***** * ***

ABI1      TAGCCTTATCTTAATCTGTCCCAACAG-CAATGTGACGGATTTTTGCAGATTCAAAATCT
Abi1      TAGCTTAATCTTAATGTGTCCAAACAGTCAATGTGACAGAATTTTGCAGA----------
          **** * ******** ***** ***** ********* ** *********

ABI1      GCAATGGTTATTTACAAGTCAATCTACTGAATTCCTTTTTTAAATAATCTTTTGAAACTA
Abi1      ------------------------------------------------TTTTGGAGCT-
                                                          ***** * **

ABI1      AGAAAATGTGTCAAATTGTGTGCATT-CATTCTGTAGGTAAAATTCTTA--AGGATTGCC
Abi1      --AGACTGTGTCAGAGTAAGCACATGATATTCTGGGGGGCAACATCTGAAGGAGGAGTGCC
            * * ******* * *  *  ***   ******  ** ** **    *  **** ****

ABI1      ATGTCAGTCCTTCAGTTGTACAGTGAACTG-TTGTACACTGGTTCAATGAAAGGAAGTTA
Abi1      ACACCGGTCCTTTGGATTTTAATGAAACTGAATTTACATTGACTCAGTGAAAGGAAA---
          *   * ******  * * *  *   *****  * **** **  *** *********

ABI1      AAAGTACCTTTTACATATTGTAAAAGTGGATACAGTTGATTTGTGAGTAGGCACTCTTTA
Abi1      -----ATCCTTTACATATTGTAAAGATGGATGCAGTTCATTTGTGGATAGGCA-TATTTA
               * * ***************  ***** ***** *******  ****** * ****

ABI1      ATCCATTACCTGGCACTAGCAACATTAGAATTTTAAAATAAAATAATTGGGAAAGAAGGT
Abi1      ATCCATTCCCTGGCACTAGAAACA-TAAAATTT-----TACAATAATTTTGAAAGAAACT
          ******* *********** **** ** *****     ** *******  *******  *

ABI1      GGGTCATGTATTAATCAGTGAACAGAGATTTACCTAACCAACAGACTTGGATTGTCTTTT
Abi1      AGGCCATGTATTAATCTGTAAACAGG--------TAATAAGCAGGTTTGTATTG------
           ** ************ ** *****         ***  * ***  *** ****

ABI1      GACATAATCAAAATGCAACACATGCACTTTGTGTGTCTCCTCTTAATTGAAGGGAGGGC-
Abi1      ----TAACTTTGATGCAACACATGCACTTTGTGTGTCTC--CTTGATTCAAGGGAGGGTG
              ***     ***************************  *** *** *********

ABI1      TGAGGGATGTTTTCTCTTCTTGTCTTGTGTATAATTCTCTATTGCTTAGGATATTAAAGT
Abi1      TGAGGGATA---------CTTCTCTTGTATATAAGACTGTACTGCTTAGGATATTAAAGT
          ********          *** ****** *****  ** ** ******************
```

```
ABI1      AGAGCACTCAAGTGTGGGTTTCTGTGTTATTGAGGATTTGTTTGGAATTCAAATTACAGT
Abi1      AGATCACACAAGTGTGGGTTTC---------------------------------------
          *** *** **************

ABI1      TATGTACTGGATGCTACAGACTTATAACAGCATAGTGAATGGTAAGACTAGTGCAAAACA
Abi1      TATGTACTGAGGGTTGCAGATTTATGACAGCATAGTGA----TGAGAGCAGTGCAAAA--
          *********   * * **** **** ************    * ***  *********

ABI1      GTTATTTCTGAAAATTAAAGACCATTATTGCTACCAAATCAATGTGACTATTTCATATGC
Abi1      --TATTTCTGAAAATTAAAGACCATTGTTCCTACCAAATCAATG-----ATTCCATATAT
            ************************ ** **************     *** *****

ABI1      ATTTT-GCCTTTGTTAATTTTAAACAAACAAAGTATCATTAGTGATCAGCTAGCTACC--
Abi1      ATTTTAACCTTTGTTATTTTTGAAGTACTA---TGTCATTAGCAATCAGATAGCTTTTTA
          *****  ********* **** **  *  *   * *******  ***** *****

ABI1      TTCTACTTTCCATTTTTCAAGTGGATTGTTCTCTTAATTTGTATATAACCTGTTGTCTAA
Abi1      TTATGCATTTCCTTTTTGAAGCAAAATGTTTTCTTCATTTGTAAATAACCTGTTGTCTAA
          ** * * ** * ***** ***   * **** **** ******* ****************

ABI1      ATTTTATGTACAGTCTTTTATAATAAACCATTCTCCTATATG-AAATTTTGGATACTGTT
Abi1      ATTTTATCT--AGTCTTTTATAATAAACTATGCTCCTATATGAAAAATTTGGATATTGTT
          ******* *  ***************** ** ********** *** ******** ****

ABI1      AAAATATAACTCATGTGGACTTCAA---TGCAGACTTCACATTTTTCTGACCTTTACTCC
Abi1      AAAGTATAACTAATGTGGATTTAAATAGTACAGACTTCAAAATTCTCTGACCTTTAATCC
          *** ******* ******* ** **   * ********* * ** *********** ***

ABI1      TACGGT-AAATCAAGACAAACCATTTGAATTTTAACATTGGTTATTTATTTGTGTAATGG
Abi1      TATGGTGAAACAAAGACAAACCACTTGAATTTTAACGTTGATTATTTATTTGTGTAATGG
          ** *** ***  *********** ************ *** ******************

ABI1      CCTTACATGTGGATATTAACAGTGTAAACAATACCACATATAACCATCAAGATACCTTGA
Abi1      CTTTACATGTGGGCATTAACAGTGTAAACAATACCACATATAACCATCAGGATACCTCGA
          * **********  ********************************** ******* **

ABI1      TTATAATTTATAGGACTAGAATGACAGATTCAGGTGGTTTATTAATTACATTTAAAAAGT
Abi1      TTAGAATTTATAGGACTATAATGACA----CTGGT------TTAATTACATTTAAAAAGT
          *** ************** *******    * ***      ******************

ABI1      TTTCTGAAAAAACTTTTGCAATAAAAA-------------------------------AA
Abi1      TTTCCG-AAAAACTTTGCCAATAAAAATTGCCCATCAATAAGGTTGAAATTATTTCTTAA
          **** * *********  *********                                **

ABI1      AATAAAT----------TGCCCATCAA---------------------------------
Abi1      AATGATTTTTGGTTTTCTGCACATGAAGCAAATAGTGTTTCACAAAATCATTAGGCACAG
          *** * *          *** *** **

ABI1      -------------
Abi1      TCTGGTAAAATAG
```

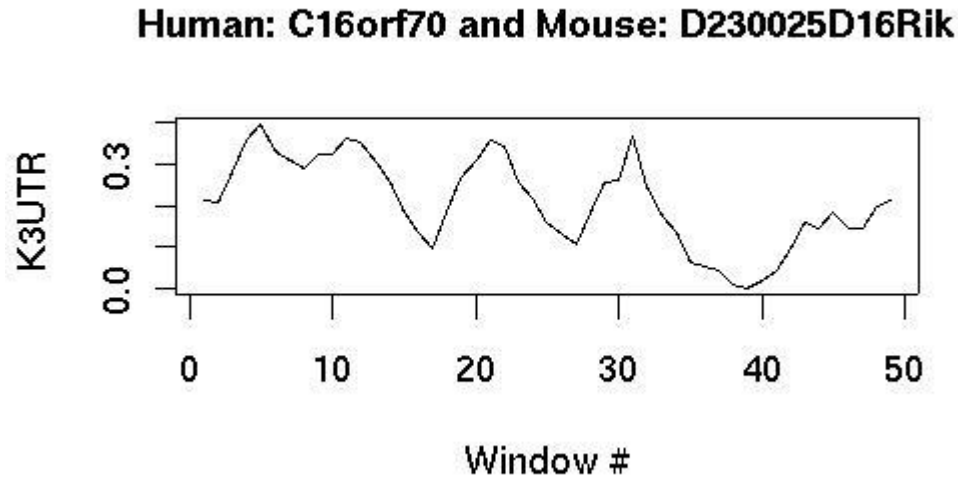


MUSCLE

(3.6) multiple sequence alignment

```
BECN1       CTTTTTTCCTTAGGGGGAGGTTTGCCTTAAAGGCTTTTAATTTTGTTTTGTTTGCAAACA
Becn1       -CTTGCTCCTTA-GGGGATGTTTGCCTTTAAGGTTTTATACTTTGTTTGGTTTGGAAAGA
              **  ****** ***** ********* **** ***  * ******* ***** *** *

BECN1       TGTTTTAAATTAAATTCGGGTAATATTAAACAGTACATGTTTACAATACCAAAAAAGAAA
Becn1       TGCTTTAAATTAAATTTGGGTAATATTAAAC--CACATGTTTACAATACC---------A
            ** ************* **************   ****************         *

BECN1       AAATCCACAAAAGCCACTTTATTTTAAAATATCATGTGACAGATACTTTCCAGAGCTACA
Becn1       AAATCCACAAAAGCTACTTTATTTTCAAATA-----TGACAGATAGTTTCCAGAG-----
            ************** ********** *****     ********* *********

BECN1       ACATGCCATCTATAGTTGCCAGCCCTGGT-CAGTTTTGATTCTTAACCCCATGGACTCCT
Becn1       -TACGCCATGTATAGCAAAGAACCCTGCCATAGTTTTGA--CTCAGCCCCATGCA-TCCT
              * ***** *****     * *****    ********  ** * ******* * ****

BECN1       TTCCCTTTCTTCTCTGAAAAAAACTAATTTAAATTTGCTTTTCTTTTTTTTAACTGAGTT
Becn1       TTCCC--TCTTTCCTGAAAACAACTAATTTAAATTTGCTTTGTTTTCTTTT--TTAAGTT
            *****  ****  ******* ********************  *** ****   * ****

BECN1       GAATTGAGATTGATGTGTTTTCACTGGATTTTTATCTCTCTCAACTTCCTGCACTTAACA
Becn1       GAATTGACGTTAATGTGTTTTCACTGGA-TTTTATCTCTCTCAACTTCCTGCACTTAA-A
            *******  ** ***************** ***************************** *

BECN1       ATATGAAATAGAAACTTTTGTCTTTACTGAGATGAGGATATGTTTGAGATGCACAGTTGG
Becn1       ATTTGAAACAGCAAAGGTT--------TGAGATGAG---ATGCTTGTGGCACACAGTTGG
            ** ***** ** **   **        *********   *** *** *   **********

BECN1       ATAATGTGGGAAAATGACATCTAAGCTTTACCTGGTCACCATGTGATGTGATCAGATGCT
Becn1       GTGATGTGGGGAAAGGACACC---------------------------GGGTCAGGAGTT
             * ******* *** **** *                             * ****  * *

BECN1       TGAAATTTAACAC--TTTTCACTTGGTTCTTATACTGAATGCCGACTCTGCTCTGTGTTA
Becn1       GCAAGTTTAACTCCGTCCTCACTTG----TAGCATTGAATGCCTCCTGTGCTGTCTAGTG
```

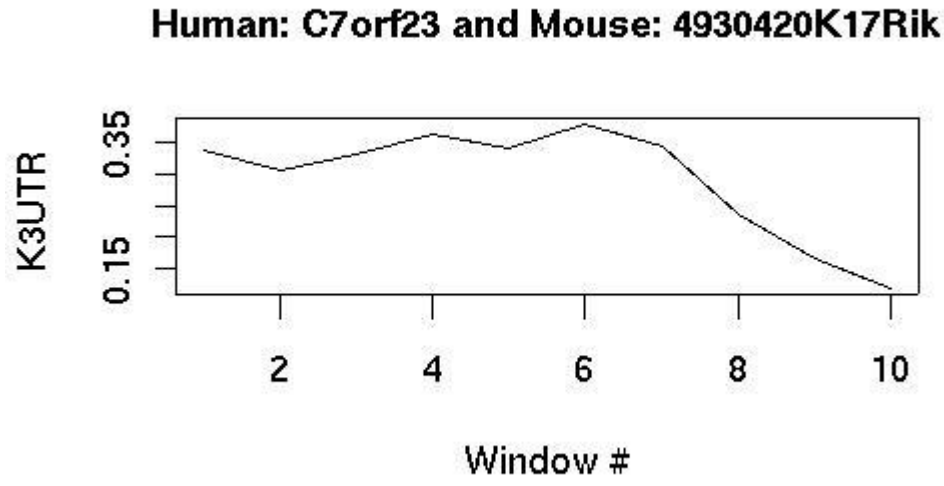
```
             ** ****** *  *  *******    *   * ********  ** **** * *  *

BECN1        GAGATATGAAATGGTGTTTGATACTGTTTGAGACATTATGGAGAGATTTAATTATTTGTA
Becn1        GGACTACAGAATGCTGTTTGATACTGTGTGCGAC---GTGGAGAGATTTAATTATTTGTA
             *   **   **** ************* ** ***    **********************

BECN1        ATAAAAGATTTGCTGCAGTCTGAAAACTGAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Becn1        ATAAAGGATTTGCTATGGTCT----ATT--------------------------------
             ***** ********   ****    * *

BECN1        A
Becn1        -
```




```
MUSCLE (3.6) multiple sequence alignment

C16orf70           GGACACCACCACCCATGCCCCTCTGTCCCGTGGAACTGTGCATCACATCCTGCTCAGTGG
D230025D16Rik      GGACATCACTGT-----TTTCCCTGT-CCATGGAACTGTGCGTCACATCCAGCTCAGTGG
                   ***** ***           * **** ** *********** ******** *********

C16orf70           GCCTCTGTACCACCCTGTGGGTTTTCTTGGACACCTGGCCAGTGCTGAAGGGCTGTTGTG
D230025D16Rik      GTCTCTGTGCCACCCTGTGTGTTTGCTTGGACACCTTGGCAGTGTTGAAGGGCCCTGGGA
                   * ****** ********** **** ********** * ***** ********  * *

C16orf70           AT-------GTTCTGAGGTTGGGCTCAGGCTGGGTGCTCTGCCATGGGCTGAGTGGCCCA
D230025D16Rik      CTCAGAAGGGTTTGGAAGTAGGATTTGGGCTGAGTCCTCTGCCGTGGGACGGAGGACCCA
                    *       ***  ** ** **  *  ***** ** ******* ****  *   * ****

C16orf70           GATATTCTTCTGTCCATCTTTGGCCTGCTGGTGCCAGCAGGGGACACAGACTGCAAAGAG
D230025D16Rik      GATACTCT-------GTCCTAAGCCTGCTGTTGCCAGGAGAGAACACAGTGTGCAAAAAG
                   **** ***        ** *  ******** ****** ** * ******  ****** **

C16orf70           AAGCACAAAGTTTGAGCCTTGATTCCTGGACCCAGGAGCTCTC--AACTAACAAG-GAGG
D230025D16Rik      AAGCACAAACTTAGAGCCATGAT----------GAGGGCTCTCATAACTAACATGTAATA
                   ********* ** ***** ****           * ******  ******** *  *
```

```
C16orf70          CAGGAGAGGTCAACCCTTGTCCCATGCACATTGGGAAGACTTGGGGCTCTTTCTGTGACT
D230025D16Rik     AAGGACACGTCAGCCCCTGCCCTGCTCACATAGGGAAGAC--AGGACTCTTTCTCTAGCT
                   **** * **** *** ** **    ***** ********   ** ******** *  **

C16orf70          GAGGACACAGGCACCCAGGATAAGGACAAGGTCCTGCCTTTGGCTCCAC----ATTGCCA
D230025D16Rik     AAGAAAATA--TTCCCAGGATGAGAACAGGCTCCTGCTTTTGGCCGGACATTGATTGCCA
                   ** * * *    ******** ** *** * ****** ******   **    *******

C16orf70          CATGACCCTTAAGGC-----AAGCAGGTAGCGTGTCCATAGTACTTGGTTGCGACATTTG
D230025D16Rik     CGTGGCCCCTAAGGTGAGCAGAGCAGGTAGCGTGTCGGTAGTACTTGGTTGCAATATTTG
                  * ** *** *****       ***************  ************** * *****

C16orf70          CACTACATCCACTTTAGTTACTTGATGAGCTGGCTGTGGCCAGGGTGGCCCACCTGCCCT
D230025D16Rik     CACTCCATCCACTTTAGTGACTTGATGAGCTGGCTGTGGCTGGTGTGGCCCA-CTGCCCT
                  **** ************** **********************  * ******** *******

C16orf70          TCCCTGTTCCTTTCTTGCACGCGCTCCCTGCCT------GGGCACACCAAACTGGTTGAA
D230025D16Rik     TCCCTG--CTTCTCCTGTGAGTGCTCCTTGCCGTGCCCAGGTGGCACCTAAGTGGATGAA
                  ******  * * ** **   * ***** ****       **   **** ** *** ****

C16orf70          ACACAGCTTTCTACCATCCAGGTACATGCCCAGGCCTGGTTCTCACCCCTGTT---GGAG
D230025D16Rik     --ATCGTTTCCTACCAGGTATGTA-----CCAGGCCTGGTTCCCACTGCTGCTCTGGGAG
                    *  * ** ******   * ***     ************ ***  *** *   ****

C16orf70          TCAGCTTTCAAGATTGCCTGTGCCCCTGCTCTGCCCCATCCTGGCTGGCAGGGCTGCATG
D230025D16Rik     CCAGCCTTTAAGGTCACCTGTGTCTCT----TGGTCCAACCTGGCTGACAGGGCTGCATG
                   **** ** *** *  ****** * **    **  *** ******** ************

C16orf70          TTTCCATCCTGTTTGCCTCTTTTATTTAATGCCAAAGTTTTAGCCAAAGACATCTTCCTA
D230025D16Rik     TCTCCATCTTGTTTGCCTCTTTTATTTAATGCCAAAGTGTTAGCCAAAGACACCTTCCTC
                  * ****** ***************************** ************* ******

C16orf70          CTTTTGTTGTGTTTCAGCATTCTGTTCTGCACTGTGGGCTGGCTCTCTGCCCCAACCCTG
D230025D16Rik     CTTTTGTTACATTACAGCATTCTGTTATGCACTGTGGGCCAGCTCCTTGCCCTAGTTCTG
                  ********   ** ************ ************  ****  ***** *   ***

C16orf70          GGCACTGGCC---CCTGGCTGGGCC--ATTTCATGGCTCAAAGCCTCTGGGGGCTCAAGG
D230025D16Rik     GGCACTGGCCTGGCCTGGCCTGGCCTGGCTACTTGGTTTGAGGC--CTAGAGGCTCAAGG
                  **********   ******  ****    * * *** *  * **  ** * *********

C16orf70          AAGGCTGGGCTGCTCAGTCCCTT-CATGGGTCTTGCTAATGGAAAGTAGCATATATGTGC
D230025D16Rik     AACACTAGGCTGCTTGGTTTCTTACATGGGTCTTGCTAATGGGAAGTAGCATATATGTGC
                  **  ** *******  **  *** ****************** *****************

C16orf70          TTTAAAAATATTAATCCTTTTGAAAAGAACTGAGAAGAAAAATGTATAATTTTATCCCAT
D230025D16Rik     TTT-AAAATAGTAATCCTTTTGAAAAGAATTGAGAAGAGAAACATATAATTTTATCCCAT
                  *** ****** ****************** ******** ***  ****************

C16orf70          TTTTAATATTTTGGTCTAGCAACTTGTGATACATAGATGACAATTTTGTGAGTTTTTCAA
D230025D16Rik     TTTTAATATTTTGGTCTAGCAACTTGTGATACATAGATGACAATTTTGTGAGTTTTTCAA
                  ************************************************************

C16orf70          ATGTGTGTACAGATTTTTGTAAAT-ATGACTCTTTTGTAATTAACTCATGTACAGCCTCA
D230025D16Rik     ATGTGTGTACAGATTTTTGTAAATAATGA--CTTTTGTAATTAACTCATGTACAGCCTCA
                  ************************ ****  ****************************

C16orf70          TCCTGTATAGTTTAATGATGAATGTGCAGGGGACCTGTCTCAGGCTCCTATATGGTTCCT
D230025D16Rik     TCTTGTATAGTTC-ATGATGAATGTGCAGGGGACCTGCCCCAGCTGTCTGTATGGTTCCT
                  ** *********  ********************** * ***    ** **********
```

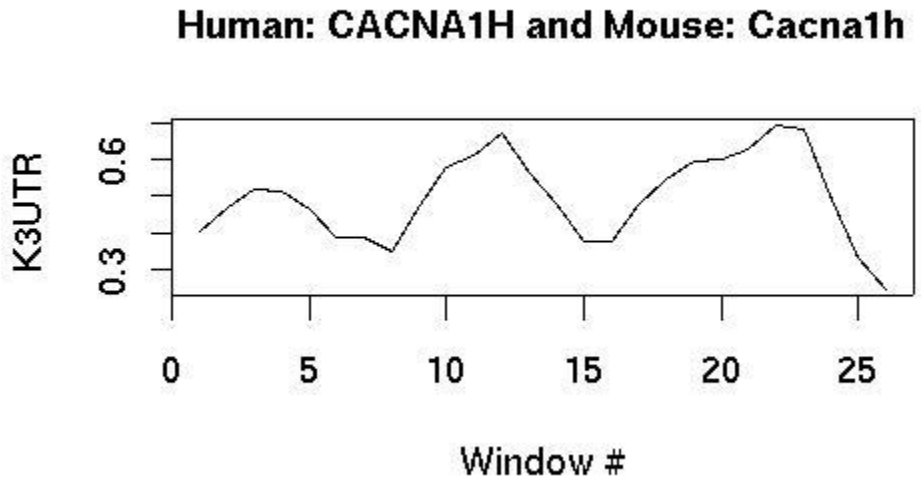
```
C16orf70          GGGCCTTATAGCCAGGTTTGTGTGGCGCTCCCGACTTTTGTGACTGACT-----GGTGTC
D230025D16Rik     GGGCC-TACAGCCAGGCCTGTGTGGCACTCCC-------ATGACTGATTTGACAGGTGTC
                  ***** ** *******  ******** *****        ******* *     ******

C16orf70          TTCCCATTTGGACTGTGGCCTGGCCAGAGCCCCTTGCATATCCCCACTGTCAGGGGCAGC
D230025D16Rik     TTCCCATTTGGACCTTCGTCTGGCCAGAGACCC-TGCACAT-CCCACTGTCA-GGGTAGC
                  *************  * * ********** *** **** ** ********** *** ***

C16orf70          CAGGACTGTTCCCATCCTCCTGGAATGGGGGAACCTTCCTTATCCCCTGCCCACATCCCC
D230025D16Rik     CAGGACTCTTCCCATCCTCGT-GAACAGAGGAGACTTCCT-------------CGTCCCC
                  ******* *********** * ***  * ***  ******             * *****

C16orf70          TCTCCAATAAAGCACCTGTGCCCTCAGCAATGGCCTGCCATGTGCTGTTGCTGGGATGTT
D230025D16Rik     CTTCCAATAAAGCACCTATGCCCACAGTGACAGCGTGCCATGTGCTGCCACTGGGCTGTT
                  *************** ***** ***  *  ** ************   ***** ****

C16orf70          TGTATTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
D230025D16Rik     TGTACTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
                  **** **************************************************

C16orf70          AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
D230025D16Rik     AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA------------------
                  *********************************
```



MUSCLE (3.6) multiple sequence alignment

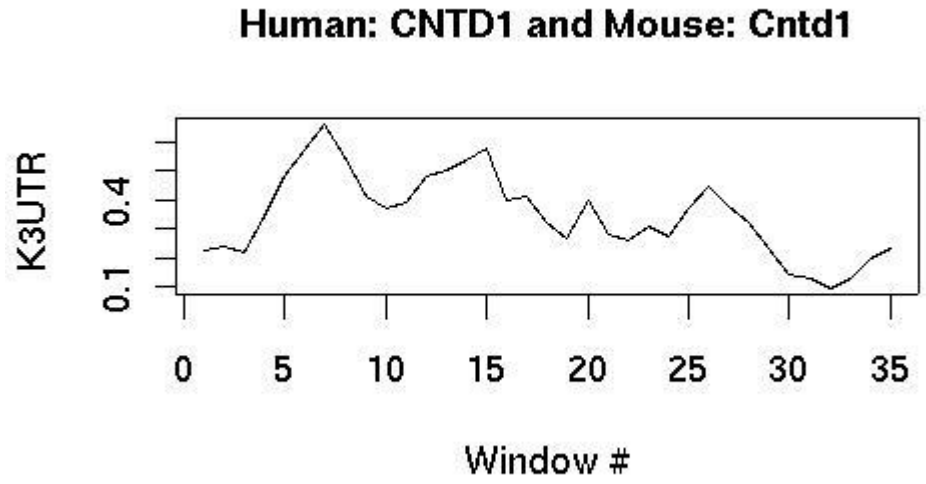
```
C7orf23           GGCTGCCAAGGAGAAGTACTTACCAGGACTCTTCAAAATGATACATTAGGACAGTGAGTA
4930420K17Rik     GGCTGCAGAGACGAAATACTTACCAGGACTCTTGGAAAT------CTATGTGAAT--TTA
                  ******  **  *** ****************  ****       ** *  * *   **

C7orf23           ATTTTTG-GATAAGGTATGCTGAAGAATCTCCTGCAGAAGTCTGATACATGA--TTTTCA
4930420K17Rik     ACTGCTGCAAAAAAGTATGCTGAAGACTCATCGACAGACGTGTGGTACATGATGTTTCCA
                  * *  **  * ** ************ **  *  **** ** ** *******  *** **
```

```
C7orf23         TGTTAATTGTAAATGTTAATTCCCTCTTGCAAGGGAGACATATCCTAGATCACTTTGCTT
4930420K17Rik   TGTTAACTGTAAATCTCAA-TCCCCCCTGCGAGGGGGCTGCAGCATAGAT--CTCTGCTC
                ****** ******* * ** **** * *** **** *    * * *****  ** ****

C7orf23         TTTCTTTAAGGAGCTGATGTTGCACCTAAACATTCCAACCCTTAAAGCTAAAACAGCACA
4930420K17Rik   TCCTTTAAA------------ACGCCCACACATTCCAGCCCTC-------AAATAG--CA
                *   ** **             * ** * ******** ****        *** **  **

C7orf23         AAAAAATTTCACTTTTGAAATGAAATTTT--------------------------------
4930420K17Rik   AGGAATTCTCACTTTTGAAAGAATTTTTTTTTTTTTGTTTTTGTTTTTTTTTTTTTTAACA
                *  ** * ************  *  ****

C7orf23         -TATAATTGTATGGCAAAAGGCTATGTAAAAACAAATCTTGCATCTTAAGACAAATATTC
4930420K17Rik   ATTTAATTGCATGGCTCA---------------AAATCTTGCATTTTAAGACAAATACTC
                 * ****** *****  *               *********** ************ **

C7orf23         TTTTATTTCTGTTAAACTGAATATACAATTGTT----CCCTAGGCAACCAACTTTTGCTT
4930420K17Rik   TTTTATTTCTGTTAAACTGAATATACAATTATTGTTCCCCTAGGCAACCAACTTTTGCTT
                ****************************** **    **********************

C7orf23         ATAACTACAATTTAATTTCACGTTGACAAAACACAGTGAAAAGACAACTTTGTGAAGATC
4930420K17Rik   ATTACTACAGTT-------------------------------------------------
                ** ****** **

C7orf23         TAATTACAATAATAAATAAAATAATTTATACAAAAAAAAAAAAAAAAAAAAAAAAAAAA
4930420K17Rik   ---TTACATTAATAAACAGA-----TGGTG--------------------------------
                   ***** ******* * *     *  *

C7orf23         AAAAAAAAAAAAA
4930420K17Rik   -------------
```




```
MUSCLE (3.6) multiple sequence alignment

CACNA1H       CTCGGGGCTTGGTGCCGCCCACGGCTTTGGCCCTGGGGTC-TGGGGGCCCCGCTGGGGTG
Cacna1h       -ACTGGGCTGTGTGTC-CACAGGGCTTTGGCATTGAGGTTGTTGGCTCCCTGCAGGGTGG
                * *****  *** * * ** *********  ** ***  * **  *** ** ***  *
```

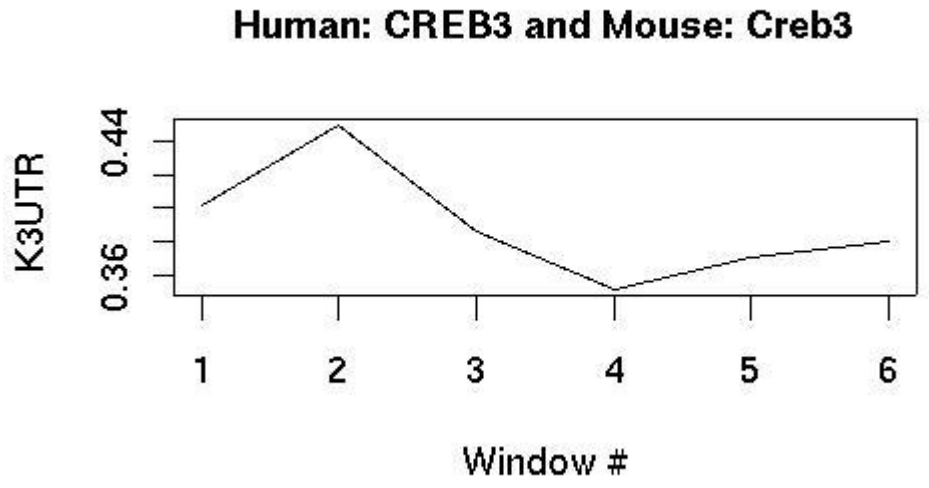
```
CACNA1H      GAGGCCCAGGCAGAACCCTGCATGGACCCTGACTTGGGTCCCGTCGTGAGCAGAAAGGCC
Cacna1h      TAGCCCCAGTCAGGGCTATGAGTGGACCCTGGCTTAGGCCCCACCAAG-GCAGAGGGACC
              ** ***** ***  *  **  ********* *** ** ***  *  * *****  * **

CACNA1H      CGGGGA---------GGATGACGGCCCAGGCCCTGGTTCTCTGCCCAGCGAAGCAGGAGT
Cacna1h      GGGAAATACTATCCCGGGAGAGGCAGCAGACGTTCCGTCTCTGCACCATGACACAGGAGC
              **  *         **  ** *   *** *  *   ******* *   **  ******

CACNA1H      AGCTGCCGGGCCCCACGAGCCTCCGTC-----CGTTCTGGTTCGGGTTTCTCCGAGTTTT
Cacna1h      AGCC-TCGGGCCCCACGAGCCTCCCTCGTGGTGATTCAGGTTTGGGTTTTCCTGAGTTTT
             ***   ****************** **       *** **** ******  * *******

CACNA1H      -GCTACCAGCCGAGGCTGTGCGGGCAACTGGGTCAGCCTCCCGTCAGGA-GAGAAGCCGC
Cacna1h      AACCACCACCAGAAGTTGTACCAG-GACCAGGTCATCAGTC--TCAGGAGGAGATACTGT
               * **** * ** * *** *  *  **  ***** *   *  ****** ****  * *

CACNA1H      GTCTGTGGGACGAAGACCGGGCACCCGCCAGAGAGGGGAAGGTACCA-GGTTGCGTCCTT
Cacna1h      GTCC-TGAGA--AGGACCAGAAATTCCTCATGGGCAGGAGGGCACCACAGCTCCATCCAT
             ***  ** **  * **** *  *  *  **  *   *** ** ****  * * * *** *

CACNA1H      TCAGGCCCCGCGTTGT---TACAGGACACTCGCTGGGGGCCC-TGTGCCCTTGCCGGCGG
Cacna1h      GTGACACACAGGTTGCCGATAGGGAGTACACGCTTGAGCCCCTTGTGCCCTGGTGGGCAG
                   * *  ****    **  *   ** **** * * *** ******** *  *** *

CACNA1H      ----CAGGT--TGCAGCCACCGCGGCCCAATGTCACCTTCACTCACAGTCTGAGTTCTTG
Cacna1h      ACATCGGGTCCTGTAGCCGCCACAACCC-ATATCACCTTCGTTCACGGTCT-CGTTCCTG
                 * ***  ** **** ** *  *** ** ********  **** ****  **** **

CACNA1H      TCCGCCTGTCACGCCCTCACCACCCTCCCCTTCCA---GCCACCACCCTTTCCGTTCCG-
Cacna1h      TCCATC-ATCA----CTCGGCATCCTTCCCTCTCACAGTACCCCACCCCCATCATTCCAT
             ***  *  ***    ***  ** *** ****  **     * ******    * ****

CACNA1H      ---CTCGGGCCTTCCCAGAAGCGTCCTGTGACTCTGGGAGAGGTGACACCTCACTAAGGG
Cacna1h      CCTCTTAGGTGGTTTAGAAACTTTGCAGTGACCCTGGGAAGGGC--CACATCACCAAGGA
                **  **   *     **   * * ***** ******  **   *** **** ****

CACNA1H      GC-CGACCCCATGGAGTAACGCGCC-CGGCCCCGATGCGAATCAGGCCTCCCCTACATCT
Cacna1h      GTACTGGCCCATGCAATAAGACGTCACAGTCCCAACAGAGGCGGGGCTTCCCCTCCATCC
             *  *   ****** * ***  ** * * * *** *         *** ****** ****

CACNA1H      GGGGGCGTTGGCCGCGAGATTCCCATTGACACC--TTTGTTTCGTGTGCT------TTTA
Cacna1h      CAGGCTGTCAGGCTCAGCGTTCAT-TTGACATCCATTTGCTTTATGTCATCCGTTCTACA
               **  **  * * *    ***   ****** *  **** **  ***  *      *  *

CACNA1H      AATTCAGGTTAAATGTTGCAATAATCTGATGCAGAAGACTCAGCTTCTCAAG--------
Cacna1h      AATTCAGGTTAAATGTTGCAATAATCTGATGCAGAAAACTTGGCTTCCTAAGTCAAAGCT
             ************************************ ***  *****  ***

CACNA1H      --GGAGAGGGAGGGGGCGGAGCG-----GAATAAATAGTAACTTATTTAAGAAATGCAAA
Cacna1h      AAGGGGAGGGGAGGGGCAAAGCAAAGCTGAATAAACACTAACTTATTTAAG---------
               ** *****  *****  ***      ******* * *************

CACNA1H      AAAAAAA
Cacna1h      --AAAGA
               *** *
```



MUSCLE (3.6) multiple sequence alignment

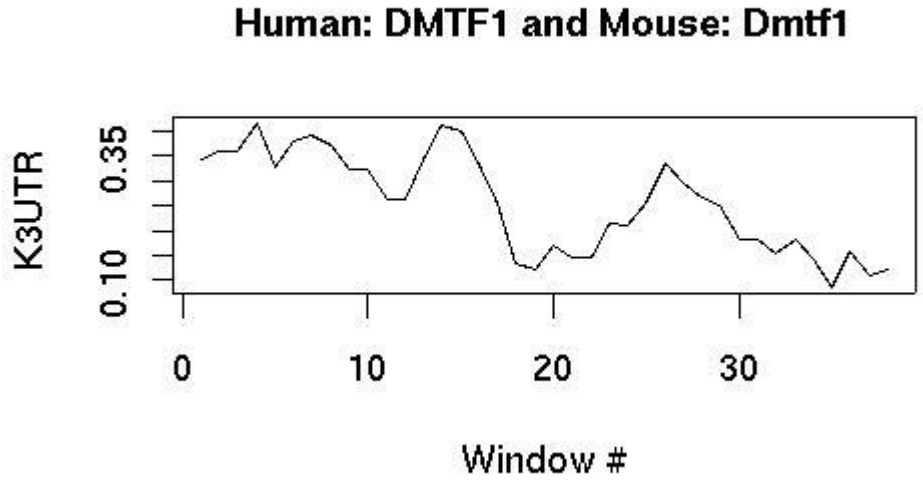
```
CNTD1       GGGAGGCTGAATCCACCAAATATAAACAGCCATCCGTCACTGCACTCATGCCTCCCTCTG
Cntd1       GGGTGGCTG-ACCCGC---------ACTGCCTT--GTCACTGGACC----TCTTTCCTTG
            *** ***** * ** *         ** *** *  ******* **      **  *  **

CNTD1       TTTACTTTCATACTAAGGGTAC-AAAAATTCCAAGTCTCTTTTGAACTGTATTTTGTATG
Cntd1       TTTACTTTTATACTCAGGGTACAAAAAATTCCAAATTTCCTTTGAACAGTGTTTTGTATT
            ******** ***** ******* ********** * ** ******* ** ********

CNTD1       CCAATTTCATGCTTA----TTTTTCCTTTATCAGAGAGAGTTAAGGTGGACGAGCATGCC
Cntd1       CCAGTTTCAGGCATGTTATTTTTTTCTACATCAGAGATAGTTATGGGAGACAAACATATC
            *** ***** ** *     ***** **  ******** ***** **  *** * ***  *

CNTD1       CTTTTTGTCATATCAGCCT---GAAAATGTTAAAAAGCTAGGTGGA---GACAGATTAGT
Cntd1       TGTTGTGTCATGGTAGGCTAAAAAAAACATCAAAGAATTGGACTGAGTTGATGATGAAAC
              ** ******   ** **    ****  * *** *  * *   **   **      *

CNTD1       TGTTTCATTTTTGTTTAACAAGGTATTTATACTTT--TAGCTTAATTTCATTAAGAGG-A
Cntd1       TAGTTCATTTTTGTTTGATAATACTTTTAATTTTTCCTAAGTTAATCTAACTAAGAGGTA
            *  ************* * **    ****   ***  **  ***** * * ******* *

CNTD1       ACATCAGGCATTGCAATCAGTATTAATCAGGGGCTCAAATACAGACTATCTG-GGTGACC
Cntd1       ATATCAG--ATAGAGATAAGTATT--TTGGGGG----ACTTTAGAGCATTTTCGATGACC
            * *****  ** *  ** ******  *  ****    * *  ***  ** *  * *****

CNTD1       TTGACTAAGCATCAAGGAGGTAGCCTTTATT-------TCCCCTTAAAATTAGTTTAACA
Cntd1       TTGACTGAATCTCTTTGAAGTAACCTTTCTTTCCTTGCTCCCCCTACAGCTGGTTGTTTT
            ****** *   **   ** *** ***** **       ***** ** *  * ***

CNTD1       TCTCTGTT-----------------------------------------CCATTATTCAG
Cntd1       GTTCTGTTTTGTTTTGTTTTGTTTTGTTTTATTTTGTTAAGACAGGGTCTCAATACTCAG
              ******                                        ** ** ****

CNTD1       AT-------------------CTACACAAACAAGGCT------------------TCCTC
Cntd1       CCCTGGCTGCCTTGCAACTCACTATACACACCAGGCTGGGCTGAAATTCAGAGGGTCACT
                               *** *** ** *****                  **
```

```
CNTD1       AACAGCTATCTATTTT----TACTAGAG------------------TCTTTTTTTAAAAC
Cntd1       TGCCTCTCTCTGCCTCTCAGTGCTGGGGCTAAATACCATCATTTGTTTTTCTCTTTAAAC
              *  ** ***   *     * ** * *                 * ** * ** ****

CNTD1       TAAAACTAACTCTAAAGAAGTTTCAACAGAATTTCCACATACCTGCATTCATTAGAACTT
Cntd1       TAAGAC-----------ATGTTTCAACAGAATTTCTA-----CTGAACTGACTACAACTC
            *** **           * ***************** *     *** * * * ** ****

CNTD1       GATTCTCCCAGAATACAAAGTACTCTATTTTAAAGAAAAACCCAACAGTGCACCCCTGGG
Cntd1       AATTCTTTCAGAATACAAAATACTCAATTTTAAGGAAAAACATACAGGCGTACCACTTGG
             *****  ************ ***** ******* *******  *   * * *** ** **

CNTD1       CAGTTTTCAGACTGCAGCAAATCTTTTATTACAAATAATTAAATCTCTCCATAATGTCTC
Cntd1       TAATTAATAGACCATAGCAAATCCTTTATTACAAATAATTAAATCTCTCC---ACGTCGC
             * **   ****   ******** *************************   * *** *

CNTD1       AAACAGTATCAAACACCATTTCATATCTCTAACACAGAGCAGAGTCGGCATTCAGTATAA
Cntd1       ACACAGTATCAAACAGCATTCTGTAGTCCCACTAGACAGCACAGGAGGCATTCAATGCTA
            * ************** ****   **   * *  * * **** **  ******** *   *

CNTD1       GAACCAAGTGAAAA--GTGTTAAATTTCAAGCATCTGATCACATCACATGGTGACCAGGT
Cntd1       ----CAAGTGAGGACGGAGTTAAACTTGCAACTCCTGACC--------------------
                ******* *  * ****** **  * *  **** *

CNTD1       AAAGCTTAGATGTCATTTTCCCACATTATCCAACTGTGCATCTCAAACATATCCTCATCT
Cntd1       -------CGGTGTCCTTTCCCCACATCACCCAACTGTGTGCCACAAGCAT---CTCATCT
                  * **** *** ******* * *********   * *** ***   *******

CNTD1       CAGTAAAGACAAAAGTTTCTATTTCATATTGTTAAGTGCAGGAAGTTGAGAGAGATAAAA
Cntd1       CA--------AACCTTTGCTGTTTCAAATT-TTAAGTGCAGGAAGTTGAGAGAGAT-AAA
            **        **   ** ** ***** *** ************************ ***

CNTD1       ATCCAGTGAAAACACATCAATCTCAATTCAACTCAGTTAAAAAAAAGAAAAGCAAATTTA
Cntd1       ATCCAGTGAAAACACATTAACGTCAATTCAACTTAA--AAAAGAAAACAAAGCAAATTTA
            ***************** **  *********** *   **** ***  ************

CNTD1       AATTAGTTTTTTTCAGAGAAGAAAGGGAAAGGAGTCCATGGGGTTAAGAATCAAAACTG-
Cntd1       AATTAGTTGTTTTCAGGAAAG--AGGGAAAGGA-TGCATGGGGCT--GAGTCAAAACTAT
            ******** *******  ***  ********** * ******* *  ** ********

CNTD1       ACCAGGGCTGGCAACTATAGATGGCATGTTGTAGCTCTGGAAAGTATCTGTCACATGATA
Cntd1       GGCAGGGTTCTTTGCTATACATGGCGT------ACTCTGGAAACTATCTGTC-----ATA
              ***** *     ***** ***** *       ********* ********     ***

CNTD1       TTTTAAAATAAAGTGGCTTTTGTGG
Cntd1       TTTGAAAATAAAGTAGCTTTTGTGG
            *** ********** **********
```




```
MUSCLE (3.6) multiple sequence alignment

CREB3       ATATGAGGATATGTGGGGGGTCTCAGCAGGAGCCTGGGGGGCTCCCCATCTGTGTCCAAA
Creb3       GTGTGAGGATGTG-GGGTGTCCTCAGCCTGA--------GGACTCCTGTCTGTATTCAAA
             * ******* ** *** *  ******  **        **   **  ***** * ****

CREB3       TAAAAAGCGGTGGGCAAGGGCTGGCCGCAGCTCCTGTGCCCTGTCAGGACGACTGAGGGC
Creb3       TAAAAAGGAGCAGGGGAAAGCTGGCCTTTGCTCACGTGCTCTGTTAGGATGCCCAGGAAC
            *******  *  **  *  *******   ****  **** **** **** * *   *  *

CREB3       TCAAACACACCACACTTAATGGCTTTCTGGGTCTTTTATTTGTACCC--ATGTGTCTGTC
Creb3       TCAGACACACCACACTAACTAGTTTTCTGAGTATTTTATTTATACCCATATGTATCTGTT
            *** ************ * * * ****** ** ******** *****  **** *****

CREB3       ACACCATGAATGTACCTGGGGAAATCAACTGACCTCCCTGAACATTTCACGCAGTCAGGG
Creb3       GCCCCGTGAATAGATTTGGAGACATTTACTGGCCT---TGAATACTTCATCCATGGAAGG
             * ** *****  *  *** ** **  **** ***   **** * ****  **   * **

CREB3       AACAGGTGAGGAAAGAAATAAATAAGTGATTCTAATGCTGCCTAAAAAAAAAAAAAAAAA
Creb3       ACTGGGTAGGGTAAGAAATAAATAAGTGGTTCCTC-------------------------
            *   ***  ** ***************** ***

CREB3       A
Creb3       -


<Human: DHPS, Mouse: Dhps; no graph, UTR too short for sliding window>


MUSCLE (3.6) multiple sequence alignment

DHPS        GCGGCTGCGGTCCCAG--GAAGGTCTTACCCCCTCTTCTATTTATTAATTTGCAGACCCA
Dhps        GAAGAT-TGGTAAAGACTGAGGCTTTTTGCCACACCTTTATTTATTACTTTACATG-CCA
            *  * *  ***       ** * * **  ** * * * ********* *** **   ***
```

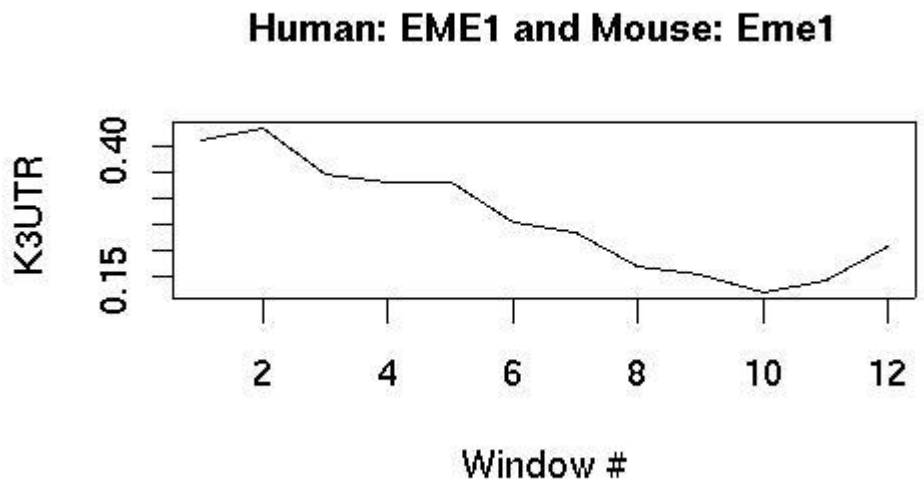
```
DHPS        GCCCCTCC-----CCTACTTTTTGGTCAGCTACGTCTCTAGAATAAGATGG-----TATC
Dhps        GCCCCTCCCTAGGCCCACTCCTTGGTCAGCAGCATCTCTAGAATAA-ATGGCCCTTTGGT
            ********     ** ***  *********  * ************ ****     *

DHPS        TGAAGTCCTTAAAAAAAAAAAAAAAAAAAAAAAAA
Dhps        TGGGTTTTCTGAGTC----------------------
            **   *   * *
```



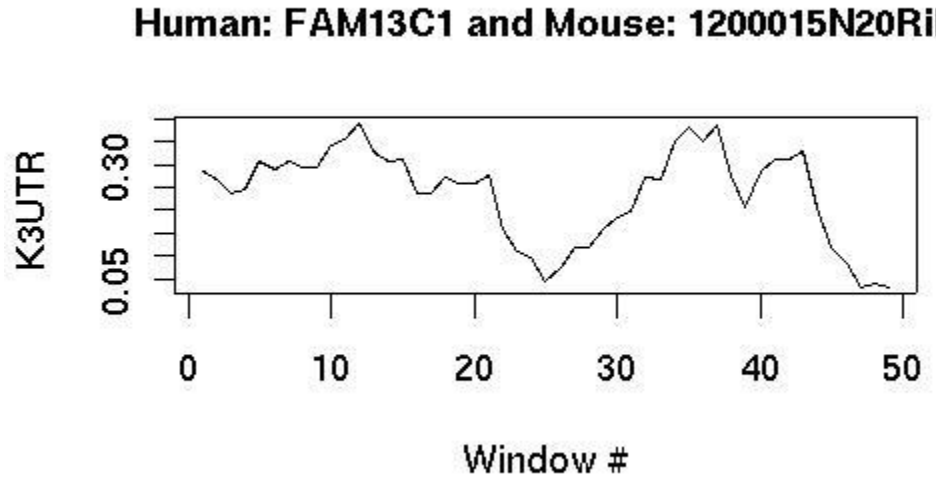
MUSCLE (3.6) multiple sequence alignment

```
DMTF1       AATAATTCTTAGAAATAGGCAGTTCAAGCAAAGAAGGCACACTGTTAATTACAACCTCTT
Dmtf1       ----ATTATTAGAAACAGGTACTTC-----AAGAAGCCACATTG-TGACTACATTGTCTC
                *** ******* *** * ***     ****** **** ** * * ****   ***

DMTF1       CAAAGAAATAGGAGCAACCCCCAAGAGGCTTAATTTACCAATT----------TAAATAG
Dmtf1       CAAAGAA--AGGAGCCA-TCCCAGGAGTTGTGGTTTGCCATTCCTCTGGCTTGTACTTAG
            *******  ****** *  **** ***   *  *** *** *           **  ***

DMTF1       CCACAGTCCTTAAGCCACACACATTGTTGCTGCTATGACTTTTTACCTCCTTTAAACACA
Dmtf1       CTGCCATGCTTAAGCCATGCACATTGTTGCTGCTGTTACTTTTACCTCCTTCTCAGTAGA
            *  *  * *********  *************** * ******  *  * * * *  * *

DMTF1       TCATCTGAGGTTGAGTTTTATGACAGTATGTAGTTGAGTGGAGGCTGGGAGTTTTAAGCA
Dmtf1       TCATCTGAGGTCCAATTTTATAACAGT-TGTTAT--GATGGAGGATAGG------AAGTG
            ***********  * ****** ***** ***  *    ****** * **      ***

DMTF1       TAAAT--CCCTGTTTAGT---GTTACATGGGAATAAGGAATTTCATTCACTTCAGCCACT
Dmtf1       TGAATTGCCCAGACTTGTTAGGTTTTATGTCAAGAGGGAGTTGCAGTCACTGCAGCTACT
            * ***  *** *  * **   ***  ***  ** * *** ** ** ***** **** ***

DMTF1       AAGAAAAGTTTAGAATCACGAAAGCTTAACTGCTGTGGTTTAAAGTACAGTT--------
Dmtf1       TAT-----------ATCACCAGAGCTTAACTACTCTGGTTTAAA-TATAAGTAGTAATAC
             *            ***** * ********* ** ********* ** *  *

DMTF1       ---TCTCTAAAGATCAGACATGGCACTGTCTCCTCTCAAGC----CTGGTTGTAGTTCAG
Dmtf1       TCATCTCTGCAG-TTAGACACAGCTCTGTCCAGACTCAAGCTGGGCTGGTTGTAATTCAG
               *****  ** * *****  ** *****    *******    ********* *****
```

```
DMTF1       ATGAGTCTTTTCAACATGGTCTTCAACATGGTCTAGAGCTTACCAGTGATCTTCTGATCT
Dmtf1       CT-------------------------TATG---------TTAGCAGAGACTTTCAGA---
             *                          ***         *** *** **  *** **

DMTF1       TCAAGAAGACTAAGTTTGAGACT-------------TGACCAGCATTCAAGTATAGAGAC
Dmtf1       -----AAGATTAACTTAGGAGTTGATGACATGACCATGACCAGCATCCAGATACCACATT
                 **** *** ** *    *             ********** **  **

DMTF1       CTAGGAGGTGGTCTTGTGGTGGTACATTTGGTTAACCCATTGCTGGCAGTGGGAGCTGAT
Dmtf1       ATGGGGGGTGGTTGTG-----------CTGGTTAACCCATTGCTGGCAGTGGGAGCTGAC
             * ** ******  **            ******************************

DMTF1       TTAGGCAGGGTAAACAGG-AAAGCATTA-AAAGTT---AAAATTCACTACAGGTTTTTTG
Dmtf1       TGAGGCAGGGTAAAGAAGCAAAGCATTACAGAGTTAAGAAAATTCACTACAGGTCTTTGA
            * ************ * * ********* * ****   ***************** ***

DMTF1       TT----ACTTTTAAAGGGAATATGGATAAGCATAGTAACAAAACCCACCAGAATCTAAGC
Dmtf1       TTTTTAACTTTTAAAGGGAACGTGGTTAAA-GTGGTAACAAAA----------------C
            **    **************  *** ***   * *********                *

DMTF1       AGTTTTCA--CCCCCTCAGAAACCACTGTCATTAGTTTACAAAGTTAGCACTTTGAAGTA
Dmtf1       AGTTATCAACTCTCATCAGAAACCA--GTCATTAGTTTAC--AGCTAGCACTTTGAAGTG
            **** ***   * * **********  *************  ** ***************

DMTF1       AAACTAAATGAGGAAGGAAGTAATGTTACCTATCCTTGAT-ACCATGAC--CATTTATTA
Dmtf1       AAATTAAATGAGGGGAGAGGTCATATTGCCTGTGTCTGATAACCATGACTTTTTTTATGA
            *** *********   ** ** ** ** *** *   **** ********    ***** *

DMTF1       GATGTTTTGCTATATAAATTACCGAGAGAATAGTT-------------------TGTCAT
Dmtf1       CATGTTTTAC-ATATAAATTGCCCAGAAGATAGTGAGGGGTATTAATTAGAGTATGCTGT
             ******* * ********* ** ***  *****                    **   *

DMTF1       CCACTTAGTGTGTTAGCTGGT-GGGGTACAATATAACCTCTCATCTCAGGCTATTTTAAA
Dmtf1       TCACTTCGTGCATTACCTGCTGGGGGTACACTATAGCTTCTC-CCTCAGGC-ACTTTTAA
             ***** ***  *** *** * ******** **** * ****  ******* * *** **

DMTF1       AAAACAATATTTGCTTCTATAACAAAAGGAAACAAATCTAAGAATCATTCCTGTACTACA
Dmtf1       AAAACAAAATTTGCTTCTGTATCAAAAGAAAACAAATC-AAGAGTCATTCCT--------
            ******* ********** ** ****** ********* **** ********

DMTF1       GAAGGGTTAAGGCAAAGGTAGCCTTTTGGGCTTTTTAATGAATATGACCCCTATAGAAAA
Dmtf1       --AAGGTAAAGCCAAAACTAGCCTGTAGGGC--TTTAATG-ATATGATCCTTATAGAAAA
              * *** *** ****  ****** * ****  ******* ****** ** **********

DMTF1       GTC--------------AAGAAAAAAAAACCCTTGTATAAATTATTTTATTTATTATTGTA
Dmtf1       GTCCAAAAAAAGGAGGAAAAAAAAAGGAAACTTGTATAAATTATTTTATTTATTATTGTA
            ***              ** ******  *  ******************************

DMTF1       ATTAGATCTTCA----CAAAGTTGTCTTTTCACTGTGT-TTTGTCAACGTGAAATTAAAT
Dmtf1       ATTAGATCTTCACAATCAAAGCTGTCTTTTCACCATCTGTTTATTAATGTA-----AAAC
            ************    ***** ***********  * * *** * ** **      ***

DMTF1       TGTAGTTATAAGCAAAAGTTGGTTGCCTAGGG----AACAATTGTATATTCAGTTTAACA
Dmtf1       TGTAGTAATAAGCAAAAGTTGGTTGCCTAGGGGAACAATAATTGTATATTCAGTTTAACA
            ****** ************************    ** **********************

DMTF1       GAAATAAAAGAATATTTGTCTTAAGATACAA---------------------------AA
Dmtf1       GAAATAAAAGAGTATTTGTCTTAAAATGCAAGATTTTGAGCCATGCAATTAAATTGTTAA
            *********** ************ ** ***                           **

DMTF1       AAAAAAAAAAAA-------------------------
Dmtf1       AAAAAAAAAAAACAAAAACAAAAAAAAAAAATTCTTTC
            ************
```




```
MUSCLE (3.6) multiple sequence alignment

EME1       TTCTAGCCCTCAGGGATGAGGATGAAAAGCTGGAAACTTCCACTTCCCCAACCTCAGAGC
Eme1       -----------------------------------------ATATCCC--ACCTTGGCTA
                                                     *  ****  ****  *

EME1       CTGACTGTAATGAAGAGACTGGCAGCACCTCCTGGAACACAAGCCTAGGTGAGGCCCAGT
Eme1       TTGACAGTA---GACAGGCCAGAAGTACCTCCTTG-GTACAG--TTAGG---GAGCAAGT
            **** ***    * ** *  * ** ******* *   ***    ****   *  * ***

EME1       CTTTCTTGGGTCTTATTATTTGTGAAGGTCTCTCTGCCTGTCGGCTGGGGCAGAGA---C
Eme1       CTCTC-------------------------ATCTCTG----------GGAACAGAGAAGGC
           ** **                         ******          **  ******   *

EME1       TGAAATACTGCCACCTACCTTTGGCAT----TTAATGTTCCTCTCCTGGCAA---AAATT
Eme1       TGAGTCACTTGAATCTACCTCCACCATAGCACTCATGTTCCTCTCCCGGCAAGCGAGA--
           ***   ***   * ******    ***     * ************ *****   * *

EME1       CACTGCCACAGACAAACCACCCCCACTCCTACCCAGCCAGCCCTCAAAACACAAAGGAAC
Eme1       ---TGCCATGGAC-AGCCA-TCCCACCCCATCCCAGTTGACCCTCACAGCACAAAGGAG-
              *****  *** * ***  ***** **  *****    ****** * *********

EME1       AAAGACAGTCCACTCAGACACTTATTTAATAACTGTAGAAATCCAAAAGAATTAGCATCA
Eme1       AAAGGCAGACCACTCAGACATGGATTTAATAATTGTAGAAATCCAAGA-AATAAGCATCA
           **** *** ***********   ********* ************* * *** *******

EME1       AATCTTGAAGTCGTGAGTGAA------GCTGCGGGTTGGCTTGACTGGGCTCAGCCACTG
Eme1       AATCTCGAAGTCA-GAGTGAACTCTTGCCTGCGGGTTGGCTTGACTACGCCCAGCCACTG
           ***** ******  *******       ******************  ** *********

EME1       AGCTGCCTCAACCGGCCAAGGAACGGGATTATGATGACTATGCGGACTTCTATATTGTCT
Eme1       AGCTGCCTCAACCAGCTAGGGAGC-----TATGATG---AGGCTGACTCC--TGTTTTCA
           ************ ** * *** *     *******   * ** **** *  * ** **

EME1       TCATCTCAT----TGTGTGTATTATGTATTTAGTTTCAATAAAGCATTTGTACCAATG
Eme1       TGATGTCACCATATGTATGTAGTATGTATTTTGTCTCAATAAAGCCTTTGTACCT---
           * ** ***     *** **** ********* ** ********** ********
```

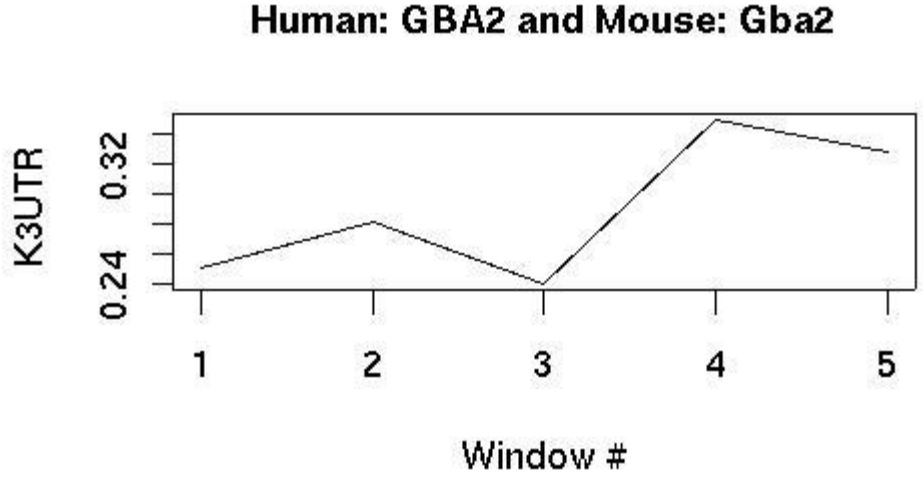


MUSCLE (3.6) multiple sequence alignment

```
FAM13C1          GGTTCAGGAAATGTTATGATCACTTTCACCCATGATATAAAGTAAAGTTTATTTTCCTCT
1200015N20Rik    GGTTCA------GCTCTGCTCACTTCC-CTCTCAAGATGAAGTCAAGTTCATTTTGCTCT
                 ******      * * ** ****** * * *   * ** **** ***** ***** ****

FAM13C1          GCCATCCTTGCTAAGTAGTTTTGACACAATGAAAATGGAAGCACTTTAGTGGTAGTATTA
1200015N20Rik    ACTGTTCTTGCTACGTAGTTTTGATGTACTGAAAACGGAAGCACTTTAGAGATAGTATTA
                  *  * ******* **********   * ****** ************ * ********

FAM13C1          GCTGTTTTTAAGAAGGAATAGCAAGTTTAATTATATACAAGGAGAAGGGATTTAAACGGG
1200015N20Rik    GCTGTTTTGACAAAGGAAAAGCTAGTATAATTATACCCAA--CCAAGGCATTTAAATAGG
                 ******** *  ****** *** *** ********  ***    **** *******  **

FAM13C1          GGGAAGAATACAACAGGTAGCCATATAATTGGGAAAAA—ATTCAGTGTCCTCCATGCCA
1200015N20Rik    G------AAATAACAGGTAATGTTGGGATTGGCAAAAAGCATTTAG---CCTCCATGCCC
                 *      * * ********    *   ***** *****  *** **   **********

FAM13C1          AGCAGAAAACTCATAGTC-AATACAAGTATTTTTAAAAATGTCTAATATTTTATCAAATC
1200015N20Rik    AACTGAAATC-CAGAGTCAAATGTAAGTGTCTCAAAAACTGTCTAATATCTTGTCAAACC
                 * * **** * ** **** ***  **** * *  **** ********** ** ***** *

FAM13C1          TAAA---TAACATAGCTAGGACACTTGTTAGGGAAAGTTTATTTAGTATCCAAAGACTGT
1200015N20Rik    TAAAATCACACGTAGGCAGGCTACCTGCTAGCGAAGATTTA-GTAGTGGCCCCAAATTGA
                 ****     ** ***  ***  ** ** *** ***  ****  ****  **  * * **

FAM13C1          TTATGTTGATGTATGGAAAAGAGCATGATTTTAAAAAATCAATCATAGGAGGAAAAGAAA
1200015N20Rik    TTTTATTGATGTATGG-AAAGAGCATGGTTTAAATACATTGGT-----GAGACAAAGAAA
                 ** * *********** ********** *** ** * **   *     ***  *******

FAM13C1          ---TTCGCTTTTCAAGTAGGAAGGAATACAGCTAGCAAGAAAGCAATTTATTTGAAACTT
1200015N20Rik    ATTTTTGCTCTTCACCCAGGGAGGAAAACAGCTAGAAGGAAGGTAATTTA----------
                    ** *** ****   *** ***** ******** * *** * ******
```

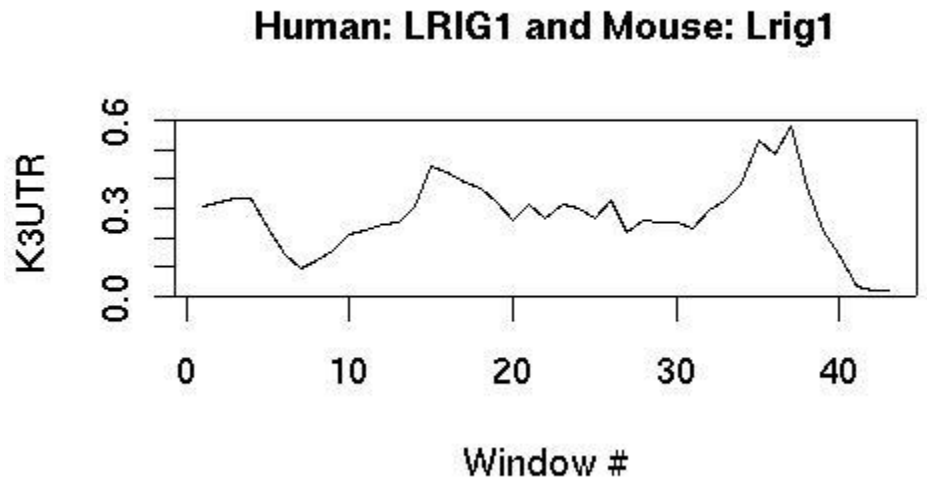
```
FAM13C1          CTAATGGATTTTTGAGTGATAAAACATTTACTACCTTGTCCTTTAAGTCTGCTAGGCTCT
1200015N20Rik    --------------AGCGAGACAACATCCACTACCTTGTCCTGTAAGCCTG-TAGGCGTT
                             ** ** * *****  ************* **** *** *****  *

FAM13C1          CAGTACCCTAAAATAAACTAGATTGTGTTGCTATT--TTTTTTCTTTCTCTATAAAAATA
1200015N20Rik    CAGCACTCTAAAAT-AACTCGACTGTGTCACTGCTAATCTCTTATTTCTATATGAAAGTA
                 *** ** ******* **** ** *****  **  *  * * ** ***** *** *** **

FAM13C1          ACACATTATTTTATCCGTTATTTGAAATTTTACATTTCTGGTTACCAAAGTTCATTCTGA
1200015N20Rik    GTGCATTATTTTATCTGTTATTTGAAAATTTACATTTCCAGTTACCAAAGTTCACTTGGA
                    ************ *********** **********  ************** *  **

FAM13C1          TAGCATGTACTTTGTGAATTATTATCTTTGTCTATAACTGACAGATGTTTATATTAAAAT
1200015N20Rik    TGGCATGTACTTTGTGA---ATTATCTTTGTCTATGACTGACAAATGTTTATATTAAAGT
                 * ***************   *************** ******* ************** *

FAM13C1          AAAATATTGTATTAAAAATTTAAAATAGGTATTTTGGATAGATATGTGTCTGTAGTATAT
1200015N20Rik    AAAATA-------------TTAAAATAGGTATTTTGGATAG--ATGTGTCTGCAGTA-AT
                 ******             **********************  ********* **** **

FAM13C1          AATCTAATGTGTCCATAGTATTATTGCTAATCTTTTGGTTTACTA---TAAGATGATATA
1200015N20Rik    AACCTAATGTGACCATAGTACTATTGCTGATTTTCTTGTTTCCTATACTAAGATTATATA
                 ** ******** ******** ******* ** ** * **** ***   ****** *****

FAM13C1          ACTATTTTTTCATTGGGAATATACATTTTTCTTAATGTTCCAACATCTATACTTTGTAAA
1200015N20Rik    ACTATTTTTCCATTGGGAATATGCATCTTTCTTAATGTTCCAAGGTCTGTACTTTATAAA
                 ********* ************ *** ***************  *** ****** ****

FAM13C1          GTCAAA-ACATTTCCCATGAGCTGTAGTTATTCATCCTTCTGTACAAAATGAAAAGTTTG
1200015N20Rik    GTGAAATAACTTTCTTGTGAGCTGCATTCA-------TTCTGCACAGGGTGAAAAGCTTG
                 ** *** *  ****   ******* * * *       ***** ***   ******* ***

FAM13C1          GAAATTGTTTGCCCTGATACCTTGAAAAAGAAGCCAGAATATTTATTTGCTTCATCAACT
1200015N20Rik    GAAA----TCACCAAGAGACCT-------------------TTTATTTGCTTTA-----T
                 ****    *  **  ** ****                   *********** *     *

FAM13C1          TCAGTGTATATCATTTTGTGTTATTTTATACGAAAACATGTTTATTATTTTCATTTTTGT
1200015N20Rik    TTAGTGTATCTTGCT--------TTTTCTCCACTCACACACTTACTGGTTTGGTTTTTGT
                 * ******* *   *        **** * *    ***   *** *  ***  *******

FAM13C1          AAAAGGAAGTAAAAGGTCAACATTTTCTCTCATGTACCAACCTTGTTTGTATTTCTATTT
1200015N20Rik    AAAATCAAGTAACAGGTAGACATTTTCTCAT---TATCAACATTGTTCCTATTTCTGTTT
                 ****  ****** ****  **********     ** **** *****  ******* ***

FAM13C1          TCTGTAATGTTTAAGTATGATGTTGAAGAAATTCACATTCTCTTATAGTTTGGATGGGAA
1200015N20Rik    TCTGTCATGTTTAAGCATGGTGTTGAAGAAACTCAGATTTTCCTATTGCTTGGGCGGAAA
                 ***** ********* *** *********** *** *** ** *** * ****  ** **

FAM13C1          GA-----CTATTGACTATTTCAGAAACAGACTTATTTCAGAGGCTTATTGTTTTCTCTGT
1200015N20Rik    AAAAAACCCAACCCAGATTCTAGACACAGAC-TATGTCAGAGGCGTGTTGTTCTCTCTGT
                  *     * *      ***  *** ****** *** ******** * ***** *******

FAM13C1          ATTTACCTAATATTTTATAACTTTTATGAATCAGAATAATGTCCTTCATAAATTTGTTTA
1200015N20Rik    ATTCACCTGCTG----------TTTATGAATCAGAGTAATGGCCTTCATAAATTTGTTTA
                 *** ****  *           ************* ***** ******************

FAM13C1          ATTGAAGTCATCTACTTCTAACAGGACAGATACACAACTATTTGAGGTTTACAAATTACA
1200015N20Rik    ATCGAAGTCATCTACTTGTAACAGGACAGATACACAACTATTTGAGGTTTACAAATTACA
                 ** ************** ******************************************

FAM13C1          TCTTTGATAAGGGAAATGGTTTCGTGACATGTACACAGTTGCTATTAAAATGTAACTCTA
1200015N20Rik    TCTTTGATAAGGGAAATGGTTTCGTGAGATGTATACAATTGCTATTAAAATGTAAC----
                 *************************** ***** *** ******************
```

```
FAM13C1          TATATTCTATATGATTGTAAATATTTTATACAACAATACAAATAAAATATTTTTCTATTA
1200015N20Rik    TATATTCTATATGATTGTAAATATTTTATACAAC-GTACAAATAAAATATTTTTCTATTA
                 ********************************  ************************

FAM13C1          TAAAAAA
1200015N20Rik    T------

                 *
```



MUSCLE (3.6) multiple sequence alignment

```
GBA2       -GCCGTCTGAACTGTGGGAGGGAAGTGCTAACAGCCCAGCCTCCAGCCTGGCCTTTCCTC
Gba2       ATCCCTCTGAACTGTGGA--------------GGCCCAGCCTCCACACTGGAC-CTCCTC
             ** ************               ************  **** *  *****

GBA2       CTTCCCCTCTGAACCTCCTGCAACCCTGAGCCATCAGGACAATCATACCCCTTCCCTTCT
Gba2       CTTCCTCCCACAA-GTCCTGCAGCCCTGAGCCAATAGGACAATCGCGCCCC------TCT
           ***** * *  **  ******* **********  *********   ****      ***

GBA2       CTCCACCCAATTGTGCCAGTAAATGGGGGTTGAGGGTGACCTAGGCAGCATTAGAATCAC
Gba2       TCCTTCCCAGCTGTGCCAGTAAAAAGGGGCTGAG-------------------GAACCAC
             *  ****  ************  **** ****                   *** ***

GBA2       TTATTTATTTCTTTCCTCACCTGTTCCCTGACTGCGTGAAATGTTCAGGGAGGTCAGTTG
Gba2       TTATTTATTTCTTACCCTACCCAGTCCTTCCATGGATGAAGTATTCA---AGGCCAGTAA
           ************ **  ***   *** *   **  **** * ****   *** ****

GBA2       ATTTCCCCAGGTACATTCATGGTGTGACAG--ACACATGGGTACAAATAAAAGACCCAGA
Gba2       ATGTCTCCAAATCTATTCACGGGGCAACAGATACATATGGGTATAAATAAAATACTC---
           ** ** ***  *  ***** ** *  ****  *** ******* ******** ** *

GBA2       AAGCCAAAAAAAAAAAAAAAAAAAAAAAAAAA
Gba2       --------------------------------
```

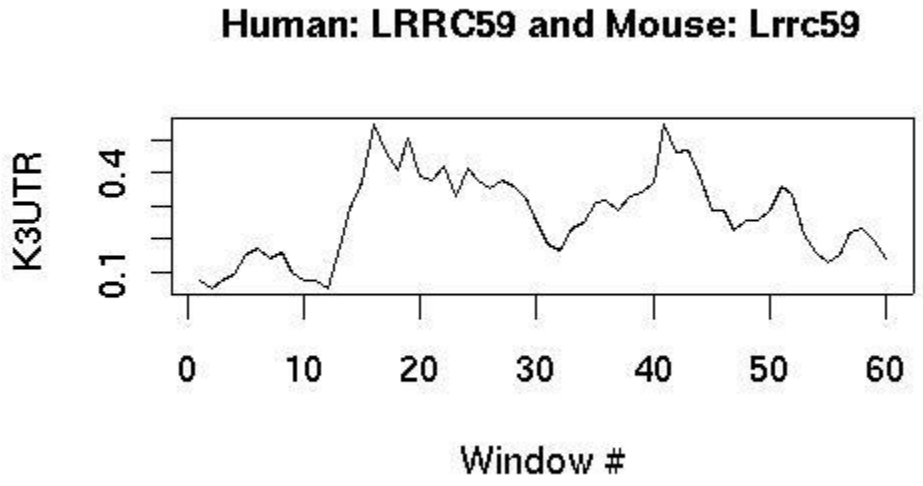


MUSCLE (3.6) multiple sequence alignment

```
LRIG1       GTTTTGTCTACCTCAGTTCTTGTCATACCAATCTCTACGGGAAAGAGAGGTAGGAGAGGC
Lrig1       GCTTCATCTACCTCAGCTCTT-TTAAAGCAGTCGCTACAGGAA--AGAGGTAGGAGAGGC
            * **  ********** **** * * * ** ** **** ****  ***************

LRIG1       TGCGAGGAAGCTTGGGTTCAAGCGTCACTCATCTGTACATAGTTGTAACTCCCATGTGGA
Lrig1       CGTG-GAAAGCTTGTGTCCAGACGTCACTC-----CGGACAGTC-CGACTTCCGTGTGGA
             * * * ******* ** **  ********        * ***    *** ** ******

LRIG1       GTATCAGTCGCTC----ACAGGACTTGGATCTGAAGCACAGTAAACGCAAGAGGGGATTT
Lrig1       ATGTCAGTCAGTCGGCTAGAGGAGTGGCATCTGGAGCTCAG-AAACGTGAGAGACTATTT
             * ******  **    * **** * * ***** *** *** *****  ****   ****

LRIG1       GTGTACAAAAGGCA-AAAAAAGTATTTGATATCATTGTACATAAGAGTTTTCAGAGATTT
Lrig1       GTGTACAAAAGGCAGGAAAAAGTATTTGATACCACTGTACATAAGAGTTTTCAGAGATTT
            **************  *************** ** ************************

LRIG1       CA-------TATATATCTTTTACAGAGGCTATTTTAATCTTTAGTGCATGGTTAACAGAA
Lrig1       CATATATATTATATATCTTTTACAGAGACTATTTTAATCCTTAGCGCATGGGT-------
            **       ****************** *********** **** ****** *

LRIG1       AAAAATTATACAATTTTGACAATATTATTTTTCGTATCAGGTTGCTGTTTAATTTTGGAG
Lrig1       --------CCCGCTTTTGGTGATA----TTTTCACATTAGGTTGCTGTCTAATTTTGGAG
                     *  *****   ***    *****  ** ********** ***********

LRIG1       GGGGTGGGGAAATAGTTCTGGTGCCTTAACGCATGGCTGGAATTTATAGAGGCTACAACC
Lrig1       AGGGTCAGG-AATCGTTCTGGTGCCTTAATGCACAGCTGGAATTCA-------GGTAAAC
             ****  ** *** *************** ***  ******** *          ** *

LRIG1       ACATTTGTTCACAGGAGTTTTTGGTGCGGGGTGGGAAGGATGGAAGGCCTTGGATTTATA
Lrig1       ACATC----------AGCTGTTGCTGGCGGTTGGAAAGGAGGGCA--TCTTGG--CTGTG
            ****           ** * *** **  ** *** ***** ** *   *****   * *
```

```
LRIG1      TTGCACTTCATAGACCCCTAGGCTGCTGTGCGGTGGGACTCCACATGCGCCGGAAGGAGC
Lrig1      TGGCACTTA--------------------GTGGTCAGATTCGA-GTGCACTGGCCAGACC
           * ******                     * ***  ** ** *  *** * **   ** *

LRIG1      --TTCAGGTGAGCACTGCTCATGTGTGGATGCCCCTGCAACAGGCTTCCCTGTCTGTAGA
Lrig1      TATGCAGGCGTG----------GTGTTG-----------GCAG-----CCTCACTGTAGA
             * **** * *          **** *            ***     ***  *******

LRIG1      GCCAGGGGTGCAAGTGCCATCCACACTTGCAGTGAATGGCTTTTCCTTTTAGGTTTAAGT
Lrig1      GC--AGGGTGCACGGGTC---------TGTAGT----------------------------
           **   ******* * * *         ** ***

LRIG1      CCTGTCTGTCTGTAAGGCGTAGAATCTGTCCGTCTGTAAGGCGTAGAATGAGGGTTGTTA
Lrig1      -----------------------------------TGAGGACGAGATTGAGAGCTGTTA
                                               * ***   *** **** * *****

LRIG1      ATCCATCACAAGCAAAAGGTCAGAACAGTTAAACACTGCCTTTCCTCCTCCTCT-TATTT
Lrig1      ATCCACCAAATGCAATGGCTCAGAACAATTAAGCACTGCCTTTCGTCTTTCTTTGTAACT
           ***** ** * ****  * ******** **** ********** ** * ** * **  *

LRIG1      TA--TGATAAAAGCAAATGTGGCCTTCTCAGTATCATTCGATTGCTATTTGAGACTTTTA
Lrig1      CACTTGGTAAAAGCAAATGT---CTT----GTCTCCTTCAGTGGTTCTTTGAAGCTT---
            *  ** *************   ***    ** ** ***  * * * *****  ***

LRIG1      AATTAAGGTAAAGGCTGCTGGTGTTGGTACCTGTGGATTTTTCTATACTGATGTTTTCGT
Lrig1      -AGTGAGGCGGAGGCTGCCAGTGTTGGTACCTGTGGATTTTCCAATAGTGAGGGTT----
            * * ***   *******  ********************* * *** *** * **

LRIG1      TTTGCCAATATAATGAGTATTACATTGGCCTTGGGGGACAGAAAGGAGGAAGTTCTGACT
Lrig1      ----------------------------------------------AGCTCTGCC-
                                                          ** **** *

LRIG1      TTTCAGGGCTACCTTATTTCTACTAAGGACCCAGAGCAGGCCTGTCCATGCCATTCCTTC
Lrig1      -TCCAGGGCCCGCTAAGTTCTGCT---------GAGCAGGCCTGTCCACGCTGTTCCTT-
            * ******   ** * **** **         *************** **  ******

LRIG1      GCACAGATGAAACTGAGCTGGGACTGGAAAGGAC---AGCCCTTGACCTGGGTTCTGGGT
Lrig1      --ACCGGTGAGACTGGGCTTGGACTGTACAAGACTTGAGTCCTAGACCTGGGTGT-----
             ** * *** **** *** ****** * * ***   ** *** *********

LRIG1      ATAATTTGCACTTTTGAGACTGGTAGCTAACCATCTTATGAGTGCCAATGTGTCATTTAG
Lrig1      ---ATTTGCACTTTGGGGCCCTGTTGCTAGCCATCTTGTGAGTGCCAATGTATAATGCAG
              *********** * * *  ** **** ******* ************* * **  **

LRIG1      TAAAACTTAAATAGAAACA------AGGTCCTTCAAATGTTCCTTTGGCCAAAAGCTGAA
Lrig1      TAAAGACTACATGGAAACACAAACTCAGTCCTTAAAAC---------CCCAGGAGTTGAG
           ****   ** ** ******        ****** ***           ***  ** ***

LRIG1      GGGAGTTACTGAGAAAATA---------GTTAACAATTACTGTCAGGTGTCATCACTGTT
Lrig1      GGGAAGTGTTGGCACAGTGATGGCCTGTGTCACCGTGTAGTGTCATAGGCCATCACTGTG
           ****  *  **  * * *          ** * *   ** *****   * *********

LRIG1      CAAAAGGTAAGCACATTTAGAATTTTGTTCTTGACAGTTAACTGACTAATCTTACTTCCA
Lrig1      CCAGA-GCAAGCACGCTGAGAAGGCTAGTCTCCATAGTTGA-TGGTTAATGTTACTTCCA
           * * * * ******  * ****   *  ***  * **** * **  **** **********

LRIG1      CAAAATATGTGAATTTGCTGCTTCTGAGAGGCAATGTGAAAGAGGGAGTATTAC-TTTTA
Lrig1      CAAAATATGTGAATTTGCTGCTTCTGAGAGGCAATGTGAAAGAGGAAGTATTACTTTTTA
           ********************************************** ******** *****

LRIG1      TGTACAAAGTTATTTATTTATAGAAATTTTGGTACAGTGTACATTGAAAACCATGTAAAA
Lrig1      TGTACAGAGTTATTTATTTATAGAAATTTTGGTACAGTGTACATTGAAAA-CATGTAAAA
           ****** ****************************************** *********
```

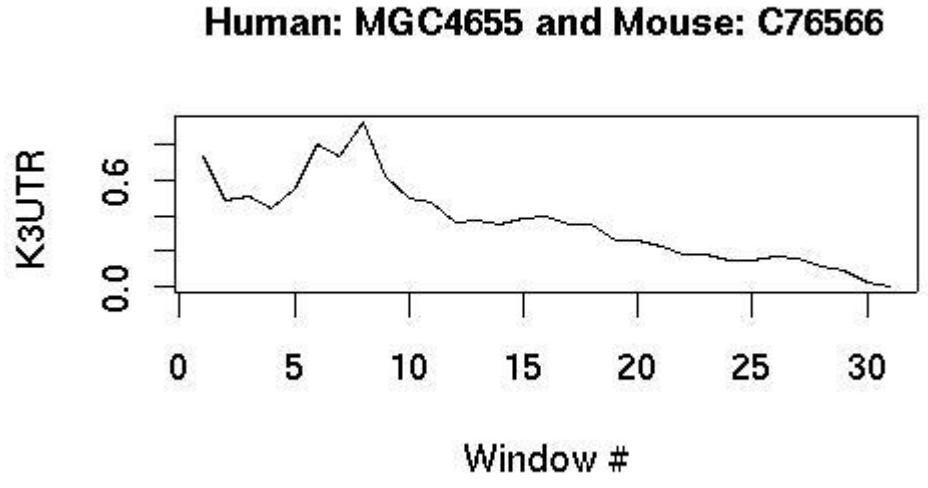
```
LRIG1       TATTGAAGTGTCTAACAAATGGCATTGAAGTGTCTTTAATAAAGGTTCATTTATAAATGT
Lrig1       TATTGAAGTGTCTAACAAATGGCA-TAAAGTGTCTTTAATAAAGGTTCATTTATAAATAC
            ********************** * ******************************

LRIG1       CAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Lrig1       TAA-------------------------------------------------
             **
```




```
MUSCLE (3.6) multiple sequence alignment

LRRC59      GCTTGTCCCCAGCACCTGCTGCCTCCCAGCCTTGGAGTTTGGATTCCTATGGAATTGGGT
Lrrc59      GCTCACCCTCAGCACCCGCTGCCTCCCAGCCTTGGAGCTTGGATTCCTATGGAATTGGGT
            ***   ** ******* ******************** **********************

LRRC59      TCTGCTGGACACAACCTCTTTTTAGCATCAGACCTACCTGCCATCATCAAATGGCTGCAG
Lrrc59      TCTGCTGGACACAACCTCTTTTTAGCGTCAGACCTACCTGCCATCATCACATGGCTGCTG
            ************************** ********************** ******** *

LRRC59      ATTGGTACATGAGACCTCCTCTTTGTAGGACTTCTTCATTCCTTAGTCAGGGTTCCCTGA
Lrrc59      G-TGGTACTTGAGATCTCCCCTTTGTAGGACTTCT--GTTCCATAGTCAGGGTTCCCTGG
              ****** ***** **** ***************   **** ****************

LRRC59      AGGAATGAGGAGAAATGGGAGGTGGCGGGGGGGCGTGGGGGGCAGTTACCTGCATGCCTA
Lrrc59      TGGAATGAGGAGAAATGGAGAGGGGGGAGGAAG----------AGTTACCTGCATGCCTA
             *****************   * ** * **  *          *****************

LRRC59      AAGGAGTAGGCTTGGGGGTGGGGAGAGAGAAAACATAGCCTTTTCTAGTTGTTATATAAA
Lrrc59      AAGGAGTAGGCTTAGGGGTGGGGAGAGAGAAGGCATAGGCTTTTCTA---GTTATACAAA
            ************* *****************  ***** ********   ****** ***

LRRC59      GCTGTGTAAAGGCAAGGCTCGTTTCTACTAAATGGTCAGCTGTCACTACATTTATACTTT
Lrrc59      GCTGTGT-AAGGCAAGGTTCCTTTCTACTAAATGGTCAGCTGTCACTACATTTATACTTT
            ******* ********* ** ************************************

LRRC59      TGTATGCCACAAACC----CTTTCATTCCTCCCTGGGAATCAGGGTAGATCAGGAGGAAC
Lrrc59      TGTATGTCATAAACCCTTTCTTTCATTCCTCCCTGGGTAACCAGGACAATCGGAGGGCAG
            ****** ** *****    ****************** * *  **   *** *  ** *
```

```
LRRC59      TGGG----GGGGACTAGAACACC----ACGCTCAGTAAATCCAGTCTAAACTGGGAGGTA
Lrrc59      TGTGTTACTGGGATTAGAGGACTAGCAATACTGGGT-AACCCAGCCTAAGCTGGGAAGGT
            ** *     **** ****  **     *  **  ** ** **** **** ****** *

LRRC59      GGGGTATTCCTGTTTTCTTTAGGACCTCAGAGATGTAAGCATTTTAGCAGCCACACAAAA
Lrrc59      GACGTAATAC--GTTTCTTTAAAGATTCAGTCAGTCAAGCAGTTTAGCAA-TATCCAAAA
            *  *** * *   ********     ****  *   ***** *******   *  *****

LRRC59      TCTCTGGCTATGAAAGGGACTTCATGACCATCCAGTCCAATATAACACTTGCAGACAGAG
Lrrc59      TGTCTGGCT--------------------GTTTGGTCCAGTGTCACATGTTGAGGC--AG
            * *******                     *   ***** * * ***  *  ** *  **

LRRC59      AAACTGAGGTCT--TCCATGACTTGCC--TAGTCTCCCAGCTAGTTTGAGGCAAAACTGG
Lrrc59      GAGGTGCCGTCTCATCCATGACTTGCCAGCAGTGTTCAGGCTAGTCTGAGGCACAACCAG
             *  **  ****  *************   *** * *  ****** ******* ***  *

LRRC59      ATTCCCACTCTGGTATTCTTTCTTCCCTTTACATCATTTTCCCTCCTTTATAATGTCCTG
Lrrc59      GCTCTCATTCCAGTTTTCCTCCTTCCCTTTATGTCATTCGACCTCCTATATAATACCCAA
              ** ** **  ** *** * **********  *****   ****** ******  **

LRRC59      AGAGACCAGAACTCACACCAGAATCGATTATTCCTCAGGTGAAGCATAGACTCTTTCATG
Lrrc59      GAGGAT--GAACTCACACCAGAGTTG------TCTCAGCTAAAGC---GAATCTTTCATA
               **   ************** * *       ***** * ****   ** ********

LRRC59      GTAGACAGATTTCACGACTCAGAGATAGAAATCTCTTGCTATCATCAGGTCACGGG----
Lrrc59      ACAG-----TCTTACCACCCACAAATAG--ATCTCATATCATCA--GGGTCCTGGGAAAC
              **     * * ** ** ** * ****  ***** *   ****   ****  ***

LRRC59      -CAGCTCCTGTGGAGTCCTGCC-CAAC--TTATGTGGCTTCCATAAAATGGCAACAGTCC
Lrrc59      TTAACTCCTGTGGAATTTTGCCTCAACTTTTAAATGGCTTCCACAAAATGGCAGCAGGCC
              * ********** *  **** ****  ***  ********* ********* *** **

LRRC59      AGGCTCCTTGCCT-AATTTTAGAGCATTAACTCCCTAATTGCCAGTAAGCAAGGAGGTGG
Lrrc59      GGGTTCCTTGCCTCAGTTTTAGAGCATTAACTCCCGAATGGCCTGGAAGCAGAG----GG
             ** ********* * ******************* *** *** * *****  *    **

LRRC59      ATCTCTGCAAACCTACACTGTCTATGACAGCTCTAGTTGTACTTGGTGTGACTAAATACC
Lrrc59      AACTCTGCAAAGCCGTACTG----TGGCTGCTCCACTAGCACTCGGTGTGATGAAATACC
            * ********* *   ****    ** * **** * * * *** *******  *******

LRRC59      TCAAAGGCAACCTGCTTCTGCAGGTTTTGAAGTGTCAGCTTCATAAGACACTGAGGTTTA
Lrrc59      TCAAAGGCAACCTA--------------GAAAC-TTGACTTCATGGAA---GCAGGTTTA
            *************               ***   *   ******   *     *******

LRRC59      GAATTGTTTGATTCTAGACC-ATAACTGAAGGGCATAAATGGAAACA------GGATATG
Lrrc59      GAG----TTGATTCTGGACCTGTAACAGAAGGGAAGGAAGGAAGAGAACAGTGGGAGTCA
            **     ******** ****  **** ****** *  ** * * * *      ***

LRRC59      AAGGGAAAC-AAGTAGCATCATGGAGCTGAAAAGTGGTGCATCACCCAATGGCTAGCACA
Lrrc59      AAGGGAAACAAAGTCACATGGTGGCGCTGCAGAGTGATACATAACTAAAC-ATTAGCACA
            ********* ****  ***  *** **** * **** * *** **  **    *******

LRRC59      GACAAGGATCACACTGTCCATTCTCTTGTCTGCTA-AATTAAGCATTTTCTTGCCTCCTT
Lrrc59      -ACCAGGGCAGAGCCGCCCACCTCCCTG-CTGCTACAAGAAAGCATTCTC------CCCT
             ** ***      * * ***    * ** ****** **  ******* **      ** *

LRRC59      TGCTTCATCTTTTCACAACAGCTGGATAGAGGGATCAGAAATGACTGTGTCATGGTGCTC
Lrrc59      TTCCCAGTCTCTTCACAACAGCTGCATGAAGGGATCAGAAACAACTGTGTCTCGGTGCTT
            * *    *** ************* **  ************  ********  ******

LRRC59      ATTCACTGCAAACTCCCAGTTGCAAGCTCCTTGGCTCCCCCGGAGGGAGCAAGAATC--T
Lrrc59      ACTTGCTAAAAACTCCCATTTGCAAGCTCCCTAG----------AGGAGCAAGGACCTGT
            * *  **  ********* ********** * *           ******** * *  *
```

```
LRRC59      CATAGTTCAGAGACACAGAGGGCCTTTTAGCCCTAATGACCT-TTTGGATGGGACTGCAA
Lrrc59      CATAGTTCAG-----CAG-------TGTAGCCCTAGTGGCCCATCTGGATAGGTCTGCCA
            **********     ***       * ******** ** **  * ***** ** **** *

LRRC59      CTCATGACTATCCTGATATTAGAAGAAAGGACTTTGTTAATCTTCTCCCCC--ATAGCTC
Lrrc59      TTCAAAACTGTCCTGACACTGGAAGAAAGGCCTTGG--ATTTTTCCTCCCCAGACAGTTC
             ***  *** ****** * * ********* *** *  * * ***  ****  * ** **

LRRC59      TGCTGCGTAGGTCTACATCT-------------TACTCAGAATCACTACACATTCCTTTA
Lrrc59      TGCCATGTAGGCCCACACCTCCTCAGCATCACCTCCTCAGCATCACTACACGCTCCTTTA
            ***   ***** * *** **             * ***** **********  *******

LRRC59      GTCTTCCTCCAAGCTCCAGAGCCATTGGTACAAATGCTTTATTGAAACTAAATACATAAT
Lrrc59      GTCTTCCTCCAAGCTCCACAGCCATTGGTACAAAGGCTTTATTGAGACAAAATACATACT
            ****************** ************** ********** ** ********* *

LRRC59      ACACACA----ATGAGATGAAGACAATATAGAAGTCCGCATAGTCATCATAATCCCGTTC
Lrrc59      ACATACATATGGTGACATCATGAAA--ACAGGAGTCAGCC---TCATCATA-----GCTC
            *** ***     *** ** * ** *  * ** **** **    ********     * **

LRRC59      CTTGGCCGGTTGAGGCAGCTCAGTGGCTGAGCCCAGTCAAGCCAACCCGCAG------CT
Lrrc59      CCTAGCTGGTTGAGGCAGCTCAGTGGCTGGGCGTAGTCAAGCCAACCCGCAGGCAAGAGT
            * * ** ********************** **  ******************       *

LRRC59      TCACTCACGACTTCAAGATTTGATGCTAATTCTTTTGGATTTCTACAGTTATTAAATAAG
Lrrc59      TCACTC-TGACTTCGAGATTTGATGCTTATT-TCTTGGATTTCTACAATTATTAAATCCA
            ******  ****** ************ *** * ************* *********

LRRC59      TGTCTGAGTGGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Lrrc59      TGTCTGAGTGGTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA----
            ***********  ******************************
```



MUSCLE (3.6) multiple sequence alignment

```
MGC4655     CTCCCCACTACAGCCCCAAGCTCCTA--ACTCAGACCCAGAATGGAGCC-----------
C76566      TTATCTCTCACGGTATCAAACTTCTCTTACTCAGCTGCAGGACAGAACATGAAAGAACCC
             *  *    ** *   *** ** **   ******   *** *  ** *
```

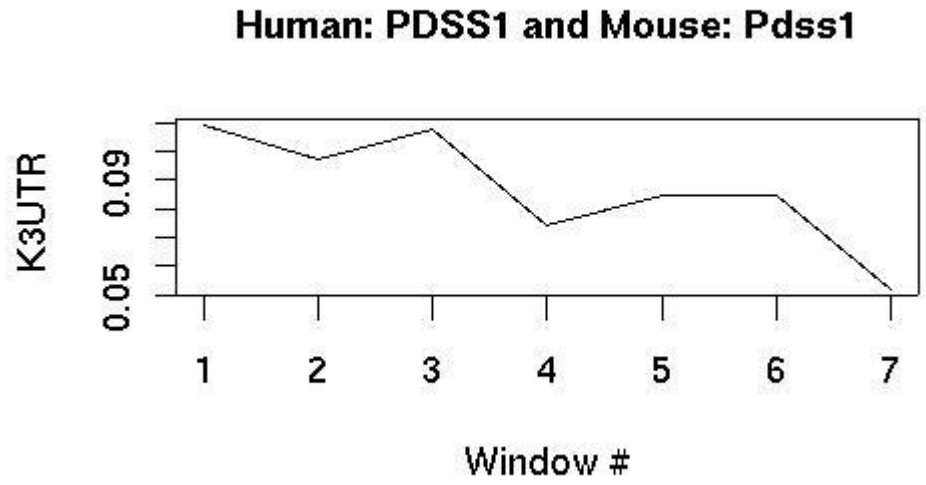
```
MGC4655     -----GGTTTCCCAGATTA--TTGCCGTGTATGTGGTTCTTCCCTGATCACCAGGTGCCT
C76566      AAAGGGGTTTTCCATGTGGGTTTATGGTGTACATGACTTTTCTCCA--------------
                 ***** ***  *    **   *****  **  * *** *

MGC4655     GTCTCCACAGGATCCCAGGGGATGGGGGTTAAGCTTGGCTCCTGGCGGTCCACCCTGCTG
C76566      ------------------------------------------------------------

MGC4655     GAACCAGTTGAAACCCGTGTAATGGTGACCCTTTGAGCGAGCCAAGGCTGGGTGGTAGAT
C76566      ------------------------GATCACTCACTGAGAGAGCTAGGGCTGG------GAT
                                  * * ** *  **** **** * ******      ***

MGC4655     GACCATCTCTTGTCCAACAGGTCCCAGAGCAGTGGATATGTCTGGTCCTCCTAGTAGCAC
C76566      GA-------------AACAA-----ACACAAGTGGATATACCC-----------------
            **             ****      * *  *********  *

MGC4655     AGAGGTGTGTTCTGGTGTGGTGGCAGGGACTTAGGGAATCCTACCACTCTGCTGGATTTG
C76566      -------------------ATAGCCAGGCCTCAG------------CCCTCGTGGACCTG
                              * **  ** ** **            * **  ****  **

MGC4655     GAACCCCCTAGGCTGACGCGGACGTATGCAGAGGCTCTCAAGGCCAGGCCCCACAGGGAG
C76566      G-------------------------------------------CAGGCATTTTCGTATA
            *                                           *****      *

MGC4655     GTGGAGGGGCTCCGGCCGCCACAGCCTGAATTCATGAACCTGGCAGGCACTTTGCCATAG
C76566      GTTACTAGCTTCCTTCCTTCTCTCATTGGATTCCCAG--------------TGGCCATAT
            **     *  ***  **  * *    ** ****                 * ******

MGC4655     CTCATC---------TGAAAACAGATATTATGCTTCCCACAACCTCTCCTGGGCCCAGGT
C76566      CTTATCTGTTCCCTTTGAGGTCTGGTGTTAGAGCTAACTAAGCCCTTC--AGGTCGAGGT
            ** ***         ***   * * * ***    *  *  * **  **   ** * ****

MGC4655     GTGGCTG-AGCACCAGGGATGGAGCCACACATAAGGGACAAATGAGTGCACGGTCCTACC
C76566      TTGCCTGTAGCATGTGGGGCGGGGCCAT------------CATGTGT-CACCGTCTTGCC
             ** *** ****   ***  ** ****              *** ** *** *** * **

MGC4655     TAGTCTTTCCTCACCTCCTGAACTCACACAACAATGCCAGTCTCCCACTGGAGGCTG---
C76566      CA--CTATACACAC-------ACACACAAAACCATGCCAGTGTACTTCTGAAGGGTGACC
             *  ** * * ***       ** **** *** ******** * *  *** *** **

MGC4655     TATCCCCTCAGAGGAGCCAAGGAATGTCTTCCCCTGAGATGCCACCACTATTAATTTCCC
C76566      CATCCCCCTCCCGG----AAGGAGT-------CCTAAGATTCCTCCACTATAAGATTGCT
             ******     **    ***** *       *** **** ** ******* *  ** *

MGC4655     CATATGCTTCAACCACCCCCTTGCTCAAAAAACCAATA-CCCACACTTACCTTAATACAA
C76566      CATATTCTTCAACTCCAGCTTCA---AGGAAATCAACAGCTCACATT----TTAGTACAA
            ***** *******  *  * *     *  *** *** * * **** *    *** *****

MGC4655     ACATCCCAGCAACAGCACATGGCAGGCCATTGCTGAGGGCACAGGTGCTTTATTGGAGAG
C76566      ACAGCCCAGTGGCAGCACATGGCACGCTGTCACTGTGGGCATAGGTGCTTTATTGGAAGG
            *** *****   ************ **  *  *** ***** ***************  *

MGC4655     GGGATGTGGGCAGGGGATAAGGAAGGTTCCCCCATTCCAGGAGGATGGGAACAGTCCTGG
C76566      GGGACG-------------AGGAAGTCTCCTCTGAT-CACGAGGATGGGAAGAGTCCTGG
            **** *             ******  *** *   * ** ********** ********

MGC4655     CTGCCCCTGACAGTGGGGATATGCAAGGGGCTCTGGCCAGGCCACAGTCCAAATGGGAAG
C76566      CTA-CCCTGACAGT-GGGATGTGC-AGGGTCTCTGGCCAGACGAAGGTCCAAATGGGAAG
            **  ********** ***** *** **** ********** * *  ***************

MGC4655     ACACC-----AGTCAGTCACAAAAGTCGGGAGCGCCACACAAACCTGGCTATAAGGCCCA
C76566      ACACCTGTCAAATCAGTCA-------TGGGAGTGCCACACAGGCCTGGCTGT-AGGCCCA
            *****     * *******        ***** ********  ******* * *******
```

```
MGC4655     GGAACCATATAGGAGCCTGAGACAGGTCCCCTGCACATTCATCATTAAACTATACAGGAT
C76566      GGAACCATACAGACAGCTGGGGCAGGTCCCCTGCACATTCATCA-TGAACTATACAAGAT
            ********* **    *** * ********************** * ********* ***

MGC4655     GAGGCTGTACATGAGTTAATTACAAAAGAGTCA-TATTTACAAAAATCTGTACACACATT
C76566      GAGGCTGTACATGAGTTAATTACAAA—AGTCATTATTTACAAAAATCTGTACACACATT
            *************************  ***** *************************

MGC4655     TGAAAAACTCACAAAATTGTCATCTATGTATCACAAGTTGCTAGACCAAAATATTAAAAA
C76566      TGAAAAACTCACAAAATTGTCATCTATGTATCACAAGTTGCTAGACCAAAATATTAAAAA
            ************************************************************

MGC4655     TGGGATAAAATTATAAAAAAAAAAAAAAAAAA
C76566      TGGGATAAAAT--TAAAAAAAAAAAAAAAAA
            ***********  ******************
```

<Human: MORG1, Mouse: 1500041N16Rik; no graph, UTR too short for sliding window>

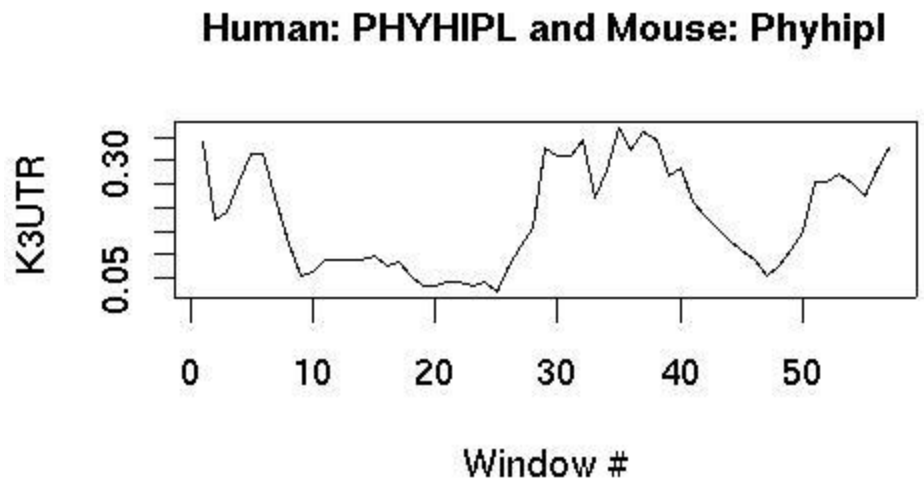
MUSCLE (3.6) multiple sequence alignment

```
MORG1           AGCCAGGGGACCCACCAACAGGACCAAGGACCGAGACACAGACATGGAAGGACT--TCAG
1500041N16Rik   GGCTTACGGACCTGAC------ACCAAGGATGCAG--ACTGACTCAGAAAACCCAACCAA
                 **    *****   *      ********   **  ** ***   ***   *    **

MORG1           ATACCATCTTATTCTAGAGACGTAGCTGACCAAAAAGTAG-----GGGAGGGGCTGGGTC
1500041N16Rik   AGGGCCATTTATTCTAGAGATGCTGCTGACCAAGGAGTGGGCCTAGGGAGGGGCT-GGCA
                *   *   ************ *  *********  *** *     ********** **

MORG1           TGCAAATTAATAAATAGAAGAGGGGGTAAGACCTTCCTGGGACCGCAAAAAAAAAAAAAA
1500041N16Rik   TGTAAAGTAATAAATAAAGGTGTGGCAAAAAGCCTC---AGTCTTTACCAA---------
                ** *** ********* * * * **  ** * * **    * *   *  **

MORG1           AAAAAAAAAA
1500041N16Rik   ----------
```



MUSCLE (3.6) multiple sequence alignment

```
PDSS1       CAACTCTTTCTGTTCTTTCTGGCAGCTATCTTACCAGACTGTGCCTAAAGAATTTTGTGG
Pdss1       CAATTCCTCCTCTTCTTTCTGGCAGCTATTTTACCAGACTGTGCCTAATG-ATTTTGTGA
            *** ** * ** ***************** ****************** * ********

PDSS1       AATACACTTTGTTTGCTTCATGTGCAGATAACCAAAAATCATTTTAAAAGATA---TCAA
Pdss1       AACAC----TATTTGCTTCATGTGCAGAAAACCAAAAATCATTTTAAGAAATAATTTCAA
            ** **    * ***************** ****************** * ***   ****

PDSS1       ACTTATTGATGGGCAATTTATTTTTTTTTATTGCAAAAGTTTTTTCAGAAAACTTTTTAA
Pdss1       CCTTATTGATGGGCAA--------TTTTTATTGGCAAAG-TTTTTCGGAAAACTTTTTAA
             ***************        *********  **** ****** *************

PDSS1       ATGTAATTAATAAACCACCTGAATCTGTCATTCTAGTCCTATAAATTATAATCAAGGTAT
Pdss1       ATGTAATTAA------ACCAG----TGTCATTATAGTCCTATAAATTCTAATCGAGGTAT
            **********      *** *    ******* ************** ***** ******

PDSS1       CTTGATGGTTATATGTGGTATTGTTTACACTGTTAATATCCACATGTAAGGCCATTACAC
Pdss1       CCTGATGGTTATATGTGGTATTGTTTACACTGTTAATGCCCACATGTAAAGCCATTACAC
            * ***********************************  ********** **********

PDSS1       AAATAAATAACCAATGTTAAAATTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Pdss1       AAATAAATAATCAACGTT------------------------------------------
            ********** *** ***

PDSS1       AAAAAAAAAAAAAAAAAAAAAAAAAAA
Pdss1       -----------------------AAAAA
                                  *****
```

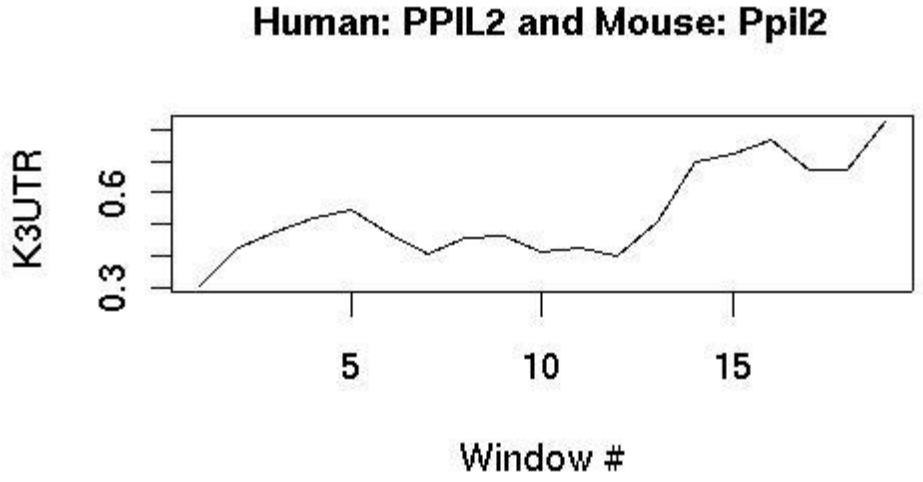


MUSCLE (3.6) multiple sequence alignment

```
PHYHIPL      TGCCCACTTTTCTTATTCTTACTCAGCC-------CCTTTTCCTCCCTTAGGAGCATTGG
Phyhipl      ---ACACCACTTTCGCCCCTTCCCCGCCCCACACACCCTGTCTTCCCATAGGAGCATT-A
                 ***   * *    * * * * ***       ** * ** **** **********

PHYHIPL      TCCTCTGTTGTCCATTTTTATCACCAGATGTTTTCCACTGAAGCATGCACATGCCACTGT
Phyhipl      CCCTCTGTTGTCCA--TCTGCCATCAGATG-TCCCCACTGAAGCATGCATATGCCGCTGT
              *************  * *  ** ****** *  *************** ***** ****

PHYHIPL      CACCAAAA-CAAACAACTACCACTTTCCAAATTTCATTCAGAACCATTTTAGTGTTTTCC
Phyhipl      CACCAAAAGCAAAC-----------TC----TGTCATTCAGAGCC-TTTCCGTGTCTCCT
             ******** *****           **    * ********* ** ***  **** * *

PHYHIPL      TATTCCCTACCCCTCCCACTACTTTCAATGATGAAATACCCTAAGTTAAGTTCTCCTTTT
Phyhipl      CTTGCCCGCCCTTGTCCACCACTTTCAACGATG----TCCCCAGGTTG--------TCCT
               * ***  **    **** ******** ****     *** * ***         *  *

PHYHIPL      GACACTTTATTGCCTAGATGCTGCAATGTTTTTATGTTTCCTTTATGCCAAACATAACAA
Phyhipl      GACACTGT-TTGCCTAGATGCTGCAATGCTTTTATGTTTCCTTTATGCCAAACATAACAA
             ****** * ****************** ******************************

PHYHIPL      AACAGTTATATAGACTGCTTTAGCAAAGTACACAATAATGGCTTATAAATGGTGTTTAAA
Phyhipl      GACAGTTATATAGACCGCTTTAGCAAAGTACAAAATAATGGCTTATCAATGGTGTTTAAA
              ************** *************** ************* *************

PHYHIPL      TATGCATTCATTTTAATCTACTGAACAAATATTGGGATAACTCCAAACCGCATGAAAGGG
Phyhipl      TGTGCATCCATTTTAATCTACTGAATAAATATCGGGATAACTCCAAACCGCATGAGAGAG
             * ***** ***************** ****** ********************** ** *

PHYHIPL      TGTAATTGCAGCATGAGTTAAATCTT-GAAGCCTAGAATTGTCAGCACTTCCAACTTCTT
Phyhipl      AGTAATTGCAGCATGAGTTAAATCCTAAAAGCCTAGACTTGTCAGCACTTCCGCCTTGTT
              ********************** *  ********* **************  *** **

PHYHIPL      GTTTGACAGTGTTATTTATGTTATTTATATTAGCTAACAGGAAACAACTACTGTG-TTTC
Phyhipl      GTTTGACAGTGTTATTTATGTTATTTATATTAGCTAACAGGAAACAGTTACTGTGCTTTC
             ******************************************  ******* ****
```

```
PHYHIPL      AACATAATAAATATAATAGAAAAATATTTTATTTGTATTGTTGTATAAAATATTTACAAT
Phyhipl      AACATAATAAATATAATAGAAAAATATTTTATTTGTA-CGTTGTATAAAATATTTACAAT
             ************************************  ********************

PHYHIPL      CATATAGAATATATAGAGTTACATTTTAATAGCAACTGTGTACATGTCACGAAACCATTT
Phyhipl      CATATAGAATAT----AGTTACATTTTAATAGCAATTGTATACATCTCACGAAACCATTT
             ************    ****************** *** ***** **************

PHYHIPL      CCCTTATCAAAGATGTAATTTGTAAACCTCAAATAGTTGTGTATCTGTCCTGTTAGAAGT
Phyhipl      CCCTTATCAAAGATGTAATTTGTAAACCTCAAATAGTTGTGTATCTGTCCTGTTACAAGT
             ************************************************** ****

PHYHIPL      AGATGACTTCAATTAAACAAATTTATGAAGGACATTATTCTGATTCATAAAAGTTATAAA
Phyhipl      AGATGACTTCGATTAAACAAATTTATGAAGGCCATTACTCTGATTCAT----------AA
             ********** ******************** ***** **********          **

PHYHIPL      ATATTAGGTAAATACAGAGAAAACAATAAGCCTCTGAAATAAGTCTGTTTCTGAAATAGT
Phyhipl      ACAGCAGGTGAATACAGAGAGAACAACACGCCTCTGACAT-AGTCTGTGTCTAGAATCTG
             * *  **** ********** ***** * ******** ** ******* ***  ***

PHYHIPL      CAATAG-----TCTTCCCATCCAAACTATAAGAGAATGTGAATTTCTTCAACATCATACT
Phyhipl      GGTTGGGTTTTTTTTTCCGCCCAAGCAATAGGAAAATCTGAGTTTCTTCAACACCATGCT
                * *     * ** **  **** * *** ** *** *** *********** *** **

PHYHIPL      TAAACATTACAGAAAATAGAAATACAAACAAGGTTGGTACATGAGAGAAAATGTTGACCT
Phyhipl      TAAACATGACAGAAAACAGAAATAGGAACAATGTTGATA---ATGAGAAAATGTCTACCT
             ******* ******** *******  ***** **** **     **********  ****

PHYHIPL      TTTACTTCCTTTTACAAAAATGAAAATAATAAACATGTTTTCGTATAAAATAACACAAAA
Phyhipl      GTTACTTGATTTTACAAAAACCAAACCAGTAAGTGTGTGAGTGGAGAA--------AAAG
              ******  ***********  ***  * ***   ***    * * **        ***

PHYHIPL      TGATATACACTGAAGTTGATGAAGCAAATAAATATTCTGGCTTCTTTTTCAAGGTATCAG
Phyhipl      CAAGATACACTAA-----ATAAAGCAAATAA-------------------AAGGTCTCTT
               * ******* *     ** **********                   ***** **

PHYHIPL      GGCAAACAATTTCCAAACTTTTCATTTTGTACAGAAGGATGAATAACTACAGCTCATGGG
Phyhipl      GG----TGATTTCCAAGCTTTTCACCCTGTGCA-------GAATGAATGCAGCTCACAAG
             **      ******** *******   *** **       **** * * *******   *

PHYHIPL      AAATGT-TTTGACTTTACAAAGTATAGATGTTGGAACATTAAGAAAAATGTATATTCCCA
Phyhipl      AAAGTTATTTCACTTTATAAAGTACAGACCTTGGAACATTAAGAAAGATGCATATTCCCA
             ***  * *** ****** ****** ***  **************** *** **********

PHYHIPL      ATGAAAAAATAGTTATATCATCT---TATAGTAAACCAAAAGATTAGCAATAATACTATG
Phyhipl      ATGGAAAAATAGTTATATAATCTTAGTATAGGAAACAAGAAAATCAGCAATAGTACTATG
             *** ************** ****   ***** **** * ** ** ******* *******

PHYHIPL      GACACATTAGATTATATACTACAGACACATATCTATCCAAAATACCTATTTTAAATTTTT
Phyhipl      GTCACATTAGGTTAT-TACTGCAGACAC--ATCTATCCAAAATACCTATTTT--------
             * ******** **** **** *******  **********************

PHYHIPL      AATACAATATTTTATTTTAATATAAACATCTGTCAGTTATAGACAAAGATAATAATTCAC
Phyhipl      -----AATATTTTACTTTAATATAAACATTTGTCAGTCATAGACAAAG---ATAATTCAC
                  ********* ************** ******* **********   **********

PHYHIPL      AAAGTACATGCTATCAGAATGAACTTTGGTAACCAGAAATGTAAAATTTCAAATAACGGA
Phyhipl      AAAGTACATGCCATCCAAGTGAACTTTGGTAACTGGAAATGTAAATTTTCAAATAACAGA
             *********** ***  * **************  ********** ********** **

PHYHIPL      TAAAATAATGTGTTATTTTTATAGAGAAAGAAAAAA--AATAGCAACACAATCTAGTTTA
Phyhipl      TAAAATAATGCACTACTTTCATATAGAAATAAGAGATTACCAGTGACACAGTCGAG-TTA
             **********   ** *** *** ***** ** * *  *  **  ***** ** ** ***
```

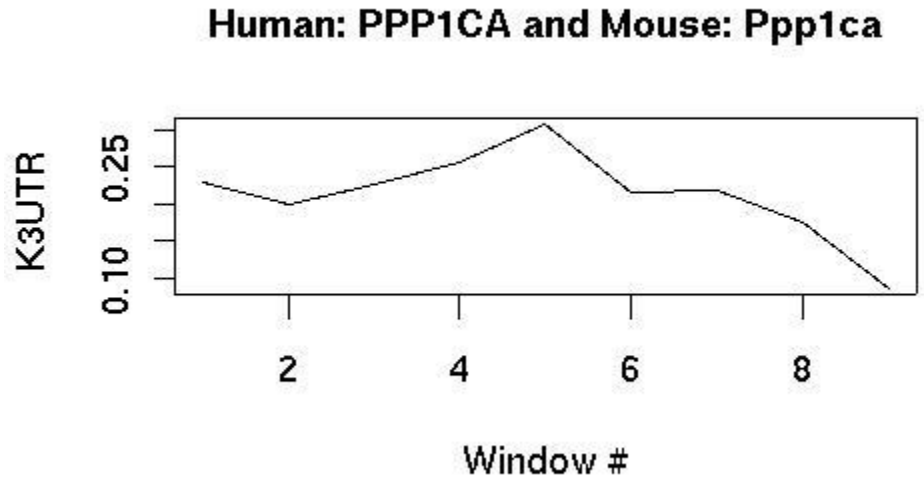
```
PHYHIPL      TTTTAGGGTACTGAGAGCCTAGCAGACTTAAAGGACAAGGTAGTAAATGTTTTATCACTC
Phyhipl      TTTTAGAGTGCTGAACGCCTA-CAGGCTTACAGGACAAGGTAGTGGATGTTGTCTCGCT-
             ****** ** ****  ***** *** **** *************  ***** * ** **

PHYHIPL      AAAAATCCATTAGAAGTTTCAAATAAATTGCTTTCTTGCTAGCTGTATTCCTTCCTACTT
Phyhipl      ----------------------TAAATTACCTTCCTTCTAGCTGTTTTCCTCCCTGGGT
                                  ****** * *** * ******** ***** ***   *

PHYHIPL      GAAAAGCGAA---TTTCTTTTCCTCCTATGATTGATTTTTTAAAATCATGCTCTTTTCCA
Phyhipl      GAAGAGCAGAAATTTTCTTTGTCTC-----ACCAATGTATTTAAACCATGCTC-TTTCCA
             *** ***  *   *******  ***     *   ** * ** *** ******* ******

PHYHIPL      TACATCAACATAAACAGTCTTTGGATACTAAATAAACTTTCCCTAACAAGAAAAAAAAA
Phyhipl      TACATCAATAAAATCAATTTGGGGCCACTA-----------CC----------------
             ******** * ** ** * *  **  ****           **
```



MUSCLE (3.6) multiple sequence alignment

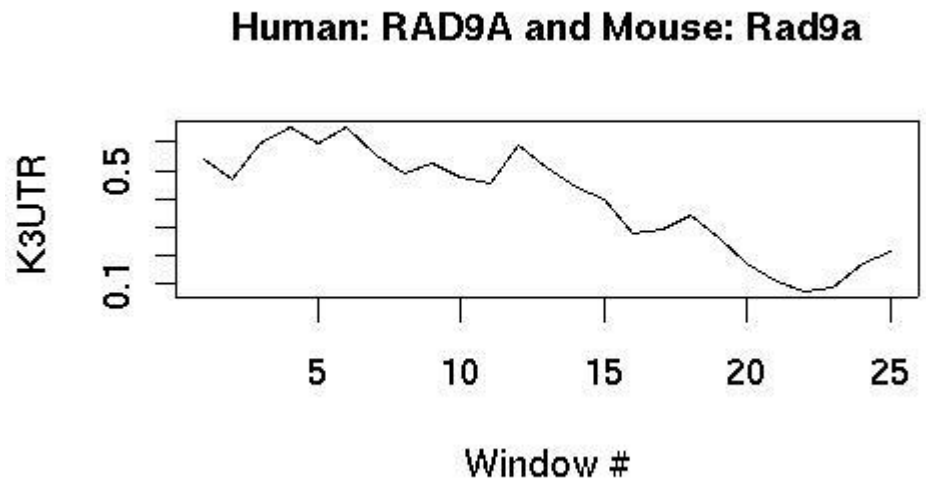
```
PPIL2        CAGCAGCAACTTGCAGGCTGTACCTCTGCCCTTCCTTTTCTCATCAATCACTGATGCTGA
Ppil2        CAGCAG------GCTGGC----------------------CTCAGGGATC----CTGTGGA
             ******      ** ***                      ****   ***     **  **

PPIL2        AGCTGCAGGCCTGAGCCCTTTGTCTCCCTGGATGCTGGGTGGCGCCTCATCTGCATCTCT
Ppil2        TGCTGCAGGAACGAAC----------------------GTGGC------------TCGCC
              ********   ** *                      *****            ** *

PPIL2        GCCTCACCC-CATCCACTGCCACAGGCTGCCTGATGACCACTAGAGGTATGTCTGCCCCT
Ppil2        GCCGCCCCCACCTCCACCCCCA-----------------------------------CCCT
             *** * *** * *****  ***                                     ****

PPIL2        CGTCACCCTGCTGCACACCAATCTGTGGCCCTTCATCATGCTAAGAACAAGAACTGCGCC
Ppil2        GGGAACCTTGCTG----------TGTGAC----------------GACAGGAGCT-----
              *  *** *****          **** *                *** ** **

PPIL2        ATGGCTGGCTCCTTCTCTTCTCCAGCCCATCCCTCTGCAGCCTGTCATCCCTGTCTGTGA
Ppil2        -----TTATTTTATACCTCCCACAG-ACACGCCACAGAGGGCTGACAGCTTTGGCTAGGA
                  *   *   *  ** *  ***  **  ** * *  * *** ** *  ** **  **
```

```
PPIL2      CCATTGGTCGGGCCCCTGGGCTCTAGAGTGACTTTTGACGCCCTCCATCCCTCCCGCCAG
Ppil2      -----GACAGGG-------------------------------------------------AG
                *   ***                                                  **

PPIL2      GCACTGTCCTCCGCAAGGCCTGGTGCAGCCCTGGCAGTAACTGGCTTGTAAGAGGCTCAG
Ppil2      GC------------CAAGGC--------GCTCT-----------GTTTGCTCGAGCTT---
           **           ******        ** **           * ***   ***  *

PPIL2      ACACCAAGCTGGGCCTGCAGAGGAGGGGCACAGTAGGACACAGTGACTGCCCAGGTGTCC
Ppil2      -------GCTGGTCCTGCTG----------------------------TTCTTGGTGAAC
                *****  ***** *                                *  ****  *

PPIL2      ACACACCTGTAGGCCTCTGAGCCAGCGTCCAGGGTACAGGTGCGGGTGGTGGGGATGAAG
Ppil2      TC---CCTGTGGGT-------------TCCACAGCAC---TGCAGTC-----------AG
            *   ***** **              ****  * **   *** *             **

PPIL2      GCCTGACCAGGGAGGGAGAAGCAGGTTTGGAGAGGACCCTGTGCCCACCCTGACAGACAC
Ppil2      CCTTGCTCAGCAGCCGACA------------------CCTGTCCCCA---GGGCAGGCA-
            * **  ***     ** *                  ***** ****    * *** **

PPIL2      CCTGGCTGGCCCTGACTGACTGTATTCTCTGGCCACATTCAAGTCCCCCATTGGTGGGGG
Ppil2      ------------------GCTGTGTCCTC---------TCCTGTCCTCCA---------G
                         **** * ***         **  **** ***         *

PPIL2      CAGAGAAGTAGGACCAGGCCATCCTTGGCTCCAGAGCTCGAAGACCCCAAGACAGCCCTC
Ppil2      CACAGAGGACAGACCA--------------CCAGGGCTTGCA------------------
           ** *** *   *****              **** *** * *

PPIL2      TGCTCTCAGCGGCGCCACAGAGAGCCTGGGCTCAGCCTTCTGCATC----AGGACATGGC
Ppil2      ---TTTC-----------------CATGGG-TCGGTTGTCTATAGCTTAAATGGTAGTGT
              * **                 * **** ** *   ***  * *    * *  *  *

PPIL2      CTCGTCCACTGAGGGCACGATTTAAACATTTGACATCAGAAGCTTTATTTGTAAACCTCA
Ppil2      TTTAGACAGTAAAGACAC----------TTGGATTTTGAAAAC--TGCCTGATGAGCTCA
            *    ** * * * ***          ** **  *   ** *  *   **   * ****

PPIL2      CACAGATAAGGACCAAGGGCTGGCGGTGTGGCCAGAGGACAGGGGAAGCTGAAGGCCCCG
Ppil2      T-CAGAAA---ATAAAGGACTTTTG---------------AGGGGAAA--AAAAATCTCA
             **** *   *  **** **   *               *******    **   * *

PPIL2      TGCTTGAGCTCGGCAGTCCTGCTCCTTGCAGTGAAGCCACCATGGGTGACCGTCCAGCCT
Ppil2      TTTCTGATCTCTTCAA----ATACCTGACA--GATGCCATCAAGAATAAGCATT------
           *   *** ***  **        ***  **  ** **** ** *  * * * *

PPIL2      CACCCGGTGGCCTGCACAGTGAGGGAAGGGCTTCAGGGCCATCTGCTCCCAGGGCAGGGG
Ppil2      --------------------------AAGGTATAAAAATACTATGTATACA------GTA
                                   * **  * *      * **    **      *

PPIL2      ACAGGCCACCAAGGACCTTTGGCAAATGAAGGTTTACATTTCTGTAGTTTGTTTGTTTTA
Ppil2      ACAAGACAACAAGG---------AAATAAA---CCACATTTCT-----------------
           *** * ** *****         **** **     ********

PPIL2      GAGCTTAATTTGTAGTTTTTTAGCTATTAAAACCATTTGAATTTTTAACGACCTGAAAAA
Ppil2      ------------------------------------------------------------

PPIL2      AAAAAAAAAAAAA
Ppil2      -----AAAGCAAA
                ***  ***
```



MUSCLE (3.6) multiple sequence alignment

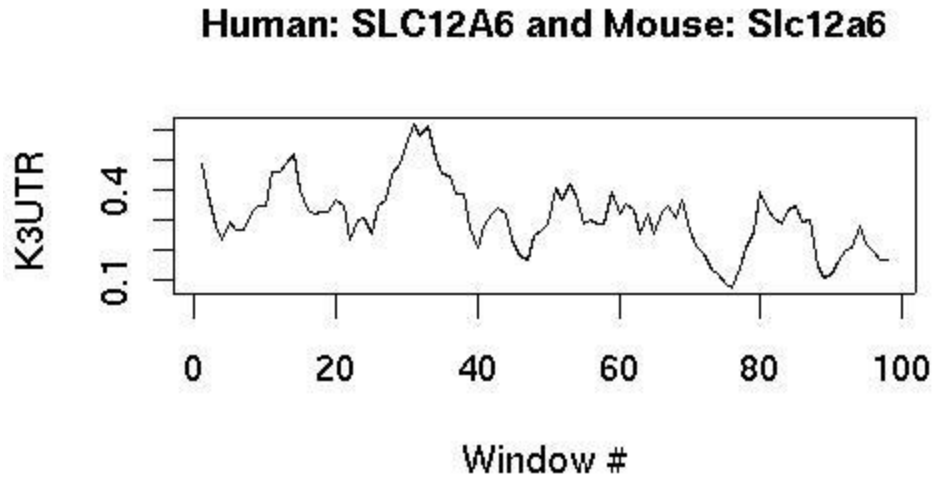
```
PPP1CA      CCCCCGCACACCACCCTG-TGCCCCAGATGATGGATTGATTGTACAGAAATCATGCTGCC
Ppp1ca      CCTCCATGTGCTGCCCTTCTGCCCCA-------GATCGTTTGTACAGAAATCATGCTGCC
            ** **     *  ****  *******       *** * **********************

PPP1CA      ATGCTGGGGGGGGGTCACCCCGACCCCTCAGGCCCACCTGTCACGGGGAACATGGAGCCT
Ppp1ca      AT---------GGGTCACACTGGCCTCTCAGGCCCACCCGTCACGGGGAACACACAGCGT
            **         ******* * * ** ************ *************   *** *

PPP1CA      TGGTGTATTTTTCTTTTCTTTTTTTAATGAATCAATAGCAGCGTCCAGTCCCCCAGGGCT
Ppp1ca      TAA-GTGTCTTTCCTTTA-TTTTTTAAAGAATCAATAGCAGCATCTAATCTCCCAGGGCT
            *   ** * **** ***  ******** ************** ** * ** *********

PPP1CA      GCTTCCTGCCTGCACCTGCGGTGACTGTGAGCAGGATCCTGGGGCCGAGGCTGCAGCTCA
Ppp1ca      CCCTCCCACCAGCACCTGTGGTGGCTGCAAGTGGAATCCTGGGGCCAAGGCTGCAGCTCA
             * ***  ** ******* **** ***  **  * ********** *************

PPP1CA      GGGCAACGGCAGGCCAGGTCGTGGGTCTCCAGCCGTGCTTGGCCTCAGGGCTGGCAGCCG
Ppp1ca      GGGCAATGGCAGACCAGATTGTGGGTCTCCAGCCTTG--------CATGGCTGGCAGCCA
            ****** ***** **** * ************** **        ** ***********

PPP1CA      GATCCTGGGGCAACCCATCTGGTCTCTTGAATAAAGGTCAAAGCTGGATTCTCGCAAAAA
Ppp1ca      GATCCTGGGGCAACCCATCTGGTCTCTTGAATAAAGGTCAAAGCTGGATTCTC-------
            **************************************************

PPP1CA      AAAAAAAAAAAAA
Ppp1ca      -------------
```



MUSCLE (3.6) multiple sequence alignment

```
RAD9A       A-CCAAGAACCTGAAGCCTGTACCCAGAGGC-CTTGGACTAGACGAAGCCCCAGCCAGTG
Rad9a       ATTTGAGAAATTAAAGTCTGTGTTCAAAGGCTCACGGACCAGAAGCAGGCCCGCCCA---
            *    ****  * *** ****   ** **** *  **** *** * ** ***  ***

RAD9A       GCAGAACTGGGTCT-----CTCAGCCCTGGGGATCAGAAAGGTGGGCTTGCTGGAGCTGA
Rad9a       -CTGAGTTGGGCGTCAGTCCTTAGACATGGACATGGAAATGGAAA-------------AA
             * **  ****  *     ** ** * ***  **   ** **                *

RAD9A       GCTGTTTCACTGCCTCTCGCAGGCCCCAGCTGGCTGTC-ACTGTAAAGCTGTCCCACAGC
Rad9a       G-----------------GCATACCCCACTTGGCTGTCTGCTGCAGAGCTGCC---TAGG
            *                 ***  *****  ********  *** * ***** *    **

RAD9A       GGTCGGGCCTGGGCCGT----TATCTCCCCACAACCCCCAGCCAATCAGGACTTTCCAGA
Rad9a       AACCATGTTTGGTCTGTGTTCTGTATTTATTCAGCCTATA-CCAATCATGTCT-------
               *  *  *** * **    * * *     ** **   * ******* * **

RAD9A       CTTGGCCCTGAACTA----CTGACGTTCCTACCTCTTATTTCTCATTGAGCCTCAGGCTA
Rad9a       -CTGGTGCCCACCTGTCACCTGACAGATC--CCTCATCCTTTACTTTGAGGCTCTAGA-A
              ***  *  * **     *****    *  **** *  **  * ***** ***  *  *

RAD9A       TACTCCAGCTGGCCAAGGCTGGAAACCTGTCTCCCTCAGGCTCACCTTCCTA----AGGA
Rad9a       TACAAATTCTGGCCAAAGCCAGAAACTTAGCTTCCTCAGTCTCCTGCCTCTAAAACAGGA
            ***     ******** **  ***** *  ** ****** ***      ***    ****

RAD9A       AAATGTCATAGTAGGTGCTGCTGGCCCCTGGTGATCCAGCTTCTCTGCCAATCATGACCT
Rad9a       AGAT------------------------CAGTATTGAACACCCTCTGCCAATCATGAAC-
            * **                        **  *  *    *************** *

RAD9A       GTTCCTTCCTGAAGTCCTGGGCATGCATCTGGGACCCCCGTG-GAGCTGACAAGTTTTCC
Rad9a       --------------TCCCCGCCA---------GACCCTCCTGTGGGCCAAAGGAACTGTT
                        ***  * **         ***** * ** * **  *      *
```

```
RAD9A       TTGCTTTCCTGATACTCTTTGGCGCTGACTTGGAATTCTAAGAGCCT-TGGACCCGAGTG
Rad9a       TCTCTTTCCTAACA--CTTGGGCCTTGGCCTAGAATCCTAATAGCCTCTGGACCTGATTG
            *  ******* * *  *** ***  ** * * **** **** ***** ****** ** **

RAD9A       TGTG--GCTAGGGTTGCCCTGGCTGGGGCCCGGTGCCGAGA-----------CTCCCAAG
Rad9a       TGTGTCACTGGTGTTGCCCTGGCTGGGCCCATGTTCAGCCACTCAGTCTGCCCTCCCAGG
            ****   ** * *************** **  ** * *  *           ****** *

RAD9A       CGGCTCTGTGCAGAAGAGCTGCCAG-GCAGTGTCTTAGATGT-------GAGACGGAGGC
Rad9a       TGGCTTTGTACCAAAGAACTGCTGGCCCACTGTCTGGGACGGGACAAGAGAGACAGAGGC
             **** *** *  **** ****  *  ** *****  ** *        ***** *****

RAD9A       CATGGCGAGAATCCAGCTTTGACCTTTATTCAAGAGACCAGATGGGTTGCCCCAGGATCC
Rad9a       TATGGTGAGAATCCAGCTTTGACCTTTATTCAAGAGACCAGATGGGTTGCCCCAGGATCT
             **** **************************************************

RAD9A       GGCTGCCAGCCCTGAGGCCAAGCACGGCTGGAGACCCACGACCTGGCCTGCCGTTGCCCT
Rad9a       GGCTGCCAGCCAT--------GCAAGGCTGGAGACCCACAATCTGGTCTGCCATTGCCCT
            ********** *        *** ************** * **** ***** *******

RAD9A       GAGCTGCAGCCTCGGCCCCAGGATCCTGCTCACAGTCACCGCAGGTGCAGGCAGGAAGCA
Rad9a       GAGCTGCAGCCTTGGCCCCAGGATTCCACTTGCAGCCACCACAGGTGCTGGTGGGAGGGA
            ************ *********** *  **  *** **** ******* **  *** * *

RAD9A       GCCCTGGGGGACTGGACGCTGCTATTGATTCATTAAAAAAAGAAAAGAAAAATACAAAAA
Rad9a       GCCCTGGGAGATTAGATGCTGCTATTGATTC-TTTAAAAAATAAAGGAAAGACACTT---
            ******** ** * ** ************** ** ****** *** **** * **

RAD9A       AA
Rad9a       --
```




```
MUSCLE (3.6) multiple sequence alignment

SLC12A6     CCTACTCTGAATGACCGTGCTTGACCTGTTTTCTTAAAAGGCCTACGTCCTCCATGGAAG
Slc12a6     TCAGCTGTGGGAGATTCTGCGTGGTCCGACTTCCCTAAGA---CTCATCGTCCATGGAGA
             *  ** **   **   *** **  * *  ***   **       * ** ********
```

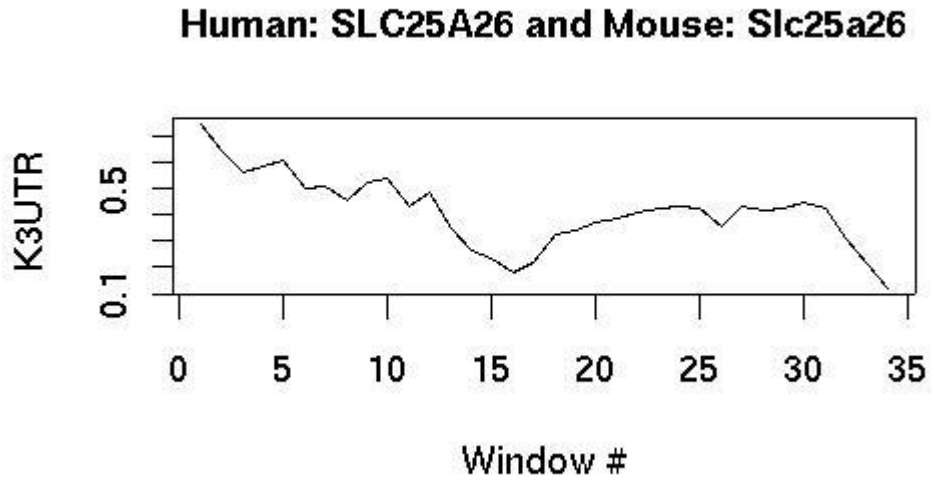
```
SLC12A6      TGCCAGCTCATTACTACCACTCCCACTCAACTAGAAGCCTGTGTTCTGTACACATCATAC
Slc12a6      TGATAGCTCTTTCCTACCACTCCCACTC--------------------------------
             **  ***** ** ***************

SLC12A6      TGAACTCTTGATGAGCTGAGCCTCAAGTA----CCTGTGTAAAAGAGCTCCCATCTGATC
Slc12a6      --TATTCTTGCAGAGCTGAGCCCCATCTATGCTCTTG---------GCTGCAACATGATC
                * *****  ********** **  **    * **         *** * *  *****

SLC12A6      TGCAGTCATTACAGAAAAAGCAAATATTCCCTCAACATCAGAACAATGCTCAAGTCTTTC
Slc12a6      TGC---CATCCCAGCAGAAACAAATAATCCCTCAA---CAGAAGAATGGTCAAGTCCTAA
             ***   ***  *** * ** ****** ********   ***** **** ******* *

SLC12A6      AAGCCACGTCTG-AGCAGTCAAAGGCAAATTAGAATTAACAAGCTGAGCCAATA------
Slc12a6      AAGCTATTTCTATGGCAGGAGAAGGCAAGTCAAAATTAACAAGCTAAGCCAAAGGGAACC
             **** *  ***   ****   ******* * * ************ ******

SLC12A6      ------------------AATGAATTGGTAAAAGGGATGCTAGAAATTCAACTGAAGAAA
Slc12a6      TTTTGGCACACAGAAGGTAATGAGTTAGTAAAACATATGCTAGAAATGTAACTGGCAACA
                           ***** ** ******   **********  *****   * *

SLC12A6      AAAAGCAAGTCAAGTACGTATTCAGCATTAAAGATGAATCTCAGAAGTCATGGTTCAATG
Slc12a6      GAAAGCAAT----GTTCATATCTAACATTCAGGACGAGTTTCAGAAGACACGGGTCAGTG
              *******     ** * ***  * **** * ** ** * ******* ** ** *** **

SLC12A6      TTGACACTGTGAGGATAACAACTAGAGACAGCTTCATCTTACTAAAGAATTTATGGTCAA
Slc12a6      --AGCACCAGAAAGTTAACAGGCAGTTA-AGCTTCACCTTAGTAAAAAATGAGTTCCCA-
                 ***    * * *****   **  * ******* **** **** ***   *   **

SLC12A6      GTATATTTGGACCTATTATCCTCGGCAAGCCAAGATGCAAACATTTTTTAGCTATATTTC
Slc12a6      --TTATAGAGAC-----AGCCTGCAGCAGCCAGGATGGAAAC------------------
                ***   ***     * ***     ***** **** ****

SLC12A6      TTTAGTATACCCACTGCTGTAATTTTATATTAGGATACTAACTTGAAACATGGCTGCAGC
Slc12a6      ---AGGATACTGA------------TATATAA-------AACCTGAGACATAGCTGCAGC
                ** ****  *            ***** *       *** *** **** ********

SLC12A6      CTCTACTTCTTCAAAAACATCCCCCCAAAATACCAGATTTAAATATCCAAATACTGTGTC
Slc12a6      CACTACTTCTTC-------------------CTCATGGTTTATTGTCT--GTACTGTACC
             * **********                    **   ** * * **    ******  *

SLC12A6      AGTGTATACTTCAACTGTTTTGTTTCTCTTAACTCAAAACCCATTCTGTCATTGTAACAC
Slc12a6      ACTGTACATTTTAACTGTCTTTGTT--------TTAAAACCCATTCTGTGGTTATAATGC
             * **** * ** ****** **  **        * **************  ** ***  *

SLC12A6      AG---GCAGTGTTACCATGGAAATTGTTTTTGGTAAGGTAAAAATTATTTCTTTCAGGTT
Slc12a6      AGACAGTAGTGTTACTATAGAAATGTTTTTTTGAAAGGCTAGAAATGTTT-TTTCAGGT-
             **   * ******** ** *****  ***** * ****  * ** * *** ********

SLC12A6      CCTTTCAAATGGCCGAACTAAGAGAACTCCAAAAGGAAGCTATTAACATTTACTGAAATA
Slc12a6      --TTTCAAAT-GCTGAACTAAGGGGATTCCAAAGAGTGCGTATTAAAATATATAGAAACA
               ******** ** ******** * * ******  *    ****** ** **  **** *

SLC12A6      GTCTGTGGTTCTCTATACAACATGGACAAAACATACAAGCTGTTGTC-TAAGAAATGCAG
Slc12a6      --CAGTGGTTAGCTATGCAACAT----AAAGCACACAAGCTATTATCTTAACAAATGTGG
               * ******  **** ******    *** ** ******* ** ** *** *****  *

SLC12A6      GTCATGGTGCACATTAATCAACAGGTTCTTCTCCCCCCGCCCCCACCCCCGCGCCTTTTT
Slc12a6      GTTGTTATGTA-GTTAACCAGAAGGGTCTCTTCTACTC-------------------TTA
             **  *  ** *  **** **  *** ***  **  * *                   **

SLC12A6      TTATGCCCTCTTATTAAGCACAGGTCAAA--AAATCT-TAAGGAACAAATACTAGGAAAA
Slc12a6      GTAGGCATTGGTCTTAAAAACCAAACAAACCAAACCTGTAAGGAAGAGACGCCAGG---G
              ** **  *  * ****  **    ****  *** ** ******* * *  * ***
```

```
SLC12A6      CTGGCTAAACATCCTTCCCTGAAAGCACTGGACCATCTTTTAAAGGGCACATCTAGCAAT
Slc12a6      CTGTGTG---ATGCCACCTCCAAGGCCCCAGG-----GTTCAGAGG-----TTCAGCAGT
             ***  *    ** *  **   ** ** *  *       ** * ***     *  **** *

SLC12A6      GGAATTGGAAGTTCTAGAGCCACACTCTTTCTTTAGTGCCAACGTTACTGAGCGGCTACT
Slc12a6      GGAACTGGGGGTGTGAAAGCCATACCCTTTCCACTGTGCCAA-------AATCGTCTGCT
             **** ***  **   * ***** ** *****    *******        * ** ** **

SLC12A6      CTCCTTACCTACTTTACTAAAATGGCATACAGATTGGC-ATTGACCTGTTGAAGAAA---
Slc12a6      C-------------------ACTGGCAGACAGGTTTGCAAGTGGCCTGACACAAGCACCT
             *                   * ***** **** ** ** * ** ****    *   *

SLC12A6      -TTTTACCACTAAGAA---ATCTTTTGCCTCATCTACAAAGCCATAGTTAAGACTTGAGG
Slc12a6      TTTCTTTCACAAGAAATCCATCTGTGGCCTCTTCACCAAAGCCATAGTGCAGATTTGAGG
              ** *  *** *  **   **** * ***** **  ************  *** ******

SLC12A6      CATTCAA-CTTAATGCCAAGTATAAGCATTTCCCCCTTTTTCTTTCAA---TTTTATTCT
Slc12a6      CACTCGACCTCCATGCCCAG----AGCTTTTCCC-------GTCTCAATATTATTATTAT
             ** ** * **  ***** **    *** ******        * ****   * ***** *

SLC12A6      TCATTCTGACTACAGAAACTTGGCCCTGAACTACGAGATTAGCCAAGCCAACAGTCTTCT
Slc12a6      TCATTCTGACTTCGGAAACTTGGCTCTGAAACAAATGGTTCACTAAGCTAATGATCTGCT
             ********** * ********** *****  *   * **  * **** **   *** **

SLC12A6      AATTTGTTTTCAGAAACTTCAGGCCGGGTGCCGTGGCTCACACCTGTAATCCCAGCACTT
Slc12a6      GTATTGGTTT----------------------------------------------------
                *** ***

SLC12A6      TGGGAGGCCAAGGCAGGTGAATCACTTGAGGTCAGGAGTTCGAGACCATCCTGGCCAACA
Slc12a6      ------------------------------------------------------------


SLC12A6      TGGTGAAAACCTGTCTATACTAAAAATACAAAAATTAGCTGGGCATGGTGGTGCATGCCT
Slc12a6      ------------------------------------------------------------


SLC12A6      GTGATCCCAGCAACACGGAAGGCTAAGGCAGGAGGATCATTTGAACCCAGGAGGCGGAGG
Slc12a6      ------------------------------------------------------------


SLC12A6      TTGCAGTGAGCCAGGATTGCACCTCTGCACTCCAGCCTGGGTAACAGAGTGATATGTTTC
Slc12a6      --------------------------TGC----------------------------TTTC
                                        ***                            ****

SLC12A6      AAAAAGAAAGTTCAGTGTGCCAAGGAAATACTTGTGTAAGCTCAG----CTTTTTATTTA
Slc12a6      AAAAAG--CTCTAACTGTGCCAAGGAAATACTTGTGTAAGATCAGATTTTTTTTTTTTTT
             ******     * * ************************ ****     ***** ***

SLC12A6      CTAATATACTGAT--CAAACATCTGGCTTACAAGCATAGGGAAACTACTTATGTGGCTGG
Slc12a6      TTAATTTACTGATCACAAACACTTGGCTTACCAGCATAGG----CAGCCTTTG----TGG
              **** *******  ******  ******** ********    *  * * **    ***

SLC12A6      CTGAGCAGTAATTTCAAGTATGTGAAATCTTCATTTCTTTTACAGAATAGTTATTTAGAG
Slc12a6      CTGAGCAGT-ATTCCCAGGGTGTGAAACCTCCATACCCTTTCCAAGATAGTTACTGA-AG
             ********* *** * **  ******* ** ***  * *** **  ******* * * **

SLC12A6      AAAC--TTTTCTTGCCTGTATTCATGGCTGTTTTTGAATTTCAATTTATTAGTCATAAAT
Slc12a6      AAACTGTCTTCTTGCTTAGATTCACGGCTGCTTTTACATTTCCTTTCACTGACAGTAAAT
             ****  * ******* *  ***** ***** ****  *****  ** * *     *****

SLC12A6      AGTGCCACTCTTCTTTGATTACTAAATAGGACATGGAGCAAGCTGACCTAGGTGTCACAA
Slc12a6      TGTGCCACTCTCC-----ATACCCTGTAGGACAATGAAC------ACCCAGA--------
              ********** *      ***    *******  ** *      *** **
```

```
SLC12A6      ATTATTCATTTGACCTGTAAGGACAG-AGAAAACCCTTCAATATAGCTTAAGAAATTGAG
Slc12a6      -------ATTCAACCC---AAGACAGAAGAAAACCCTCAAATATGGCTTAAGAAGCTGAG
                    ***  ***    * ***** **********  ***** *********  ****

SLC12A6      TAATTTTGGGCCGGGTGCAGTGACTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGG
Slc12a6      TCA-----------------------------------------------------------
             * *

SLC12A6      CAGGCGGATCACGAGGTCAGGAGATTTGAGACCATCCTGGCTAACACGGTGAAACCCTGT
Slc12a6      ------------------------------------------------------------


SLC12A6      CTCTACAAAAATACAAAAAAATTAGTCGGGCATGGTGGCGGGCGCCTGTAGTCCCAGCTA
Slc12a6      ------------------------------------------------------------


SLC12A6      CTTGGGAGGCTGAGGCGGGAGAATGGCGTGAACCCGGGAGGTGGAACTTGCAGTGAACCG
Slc12a6      ------------------------------------------------------------


SLC12A6      AGATTGCACCACTGCACTCCAGCCTGGGCGACAGAGGGAGACTCCATCTCAAAAAAAAAA
Slc12a6      ------------------------------------------------------------


SLC12A6      AAAAAGAAAGAAAGAAAGAAATTGAGTAATTTTGATTTGTTCCCTTTGCTTGCTTCAACC
Slc12a6      -------------------------------TATGGTTTGTTCTTCCTCCT----------
                                            * ** *******    * **

SLC12A6      TCCAGGCCCACTGAGGATGGCAAAGCGGATAGGATCTTAAGTCTTTTTTTCTGTTACTTC
Slc12a6      CCCAGGCCCACTT-GAGTAGCAA-------AGGATCC--AGCATGTTTTCCT-TTGCTTC
              ***********  *  * ****       ******   **  * **** ** ** ****

SLC12A6      CTACAGATGGTAGAAGCATTATTTACAACTCCCACTGCAACACTCCATTGCTTAGTTGAG
Slc12a6      CTGCAGAGGGTAGAAAC--TACTCGTAGTTCTCTTTCCAACACTCCATTGCATACTTGAG
             ** **** ******* *  ** *   *  ** *  * ************** ** *****

SLC12A6      TTTAACAGTAGTGAGTCATGTCCTGCAGGATCCAGTATCTTTTTTTTTTTTTTTCCGAGAC
Slc12a6      TATAA-------------------------------------------------------
             * ***

SLC12A6      GTAGTTTCGCTCTTGTTGCCCAGGCTGGAGTGAAATGGCACAATCTTGGCTCACCGCAAC
Slc12a6      ---------------TCCCCTAGGCT-----------------------------------
                            *  ** *****

SLC12A6      CTCTGCCTCCTGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGAATTACA
Slc12a6      ------------------------------------------------------------


SLC12A6      AGCATGTGCCACCACACCCAGCTAATTTTGTATTTTCAGTAGAGACGGGGTTTCTCTACT
Slc12a6      ------------------------------------------------------------


SLC12A6      GAAACTGATGTTGGTCAGGCTGGTCTCGAACTCCCAACCTCAGGTGATCCGCCCGCCTTG
Slc12a6      ------------------------------------------------------------


SLC12A6      GCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACTGTGCCCAACCAGGATCCAGTATC
Slc12a6      -----CCAAGGCATTGG-------------------------------------------
                  **** *   ***
```

```
SLC12A6      TTAAAGGCCCATACTGTAAAGGTCAGCCTGCCTAGGAATTCTGCCACATTTACACACACC
Slc12a6      ---AGCGCCCATTCTGTAAAAGCCAGCCTGCCCAGGAA--CTGCCTAA--CACATCCACC
                *  ****** ******* * ********* *****  *****  *   ***  ****

SLC12A6      CCTGGCCAAAGTAATATTTTTTTTTTCCATGTAGAACAAATTATGGGCAGAAAGTAAAGT
Slc12a6      CCTGACCAAAGTGACACATACTTTTTCCATATCAAACAAA-CATGGGAAGAGAATAAGGT
             **** ******* * *  *  ********* *  ******  ***** *** * *** **

SLC12A6      TAGTTTTAGAAGTGCTGTTTGTGGTTGGAGAATTGTAAAAATCTTTTAAAGGGATAAGGG
Slc12a6      TTCTTTAAGAAAGGCTGTTTGTGGTTAGAGCTACATAAGAATCTGTTAAAGGG-------
             *  *** ****  ************* ***     *** ***** ********

SLC12A6      AATTAGTGGCTTTTAATATAACCACTGCTTCAGCCCGAAACACAAAATCACTGTTACATG
Slc12a6      ---TAG--------AGTATAGTCTTCATTC---CCCAAGATA-AAAATCACTGTTACATG
                ***        * ****  *     *    *** * * * *****************

SLC12A6      GAAAACTGGTTATTAGGCCAGATTTATATTCATGTCTGTCTAGAGTATTTCATGTGTGTA
Slc12a6      GACAAATGGTTATTAGGCCAGATTTGTACTCATCTCTGTCTAGAACATTTCA--TGTCTA
             ** ** ******************* ** **** **********  ******  *** **

SLC12A6      TGTGTTGCTTACGTTGTCTGTCTCCAAAGTCTCTCATCTCTTTTAATCACTGCAAATAAA
Slc12a6      TGTGTTGCTTATGTTGTCTGTCTGCAAAGTGTCTTAGCTCTTTTAATCACTGCAAATAAA
             ********** ********** ****** *** * ************************

SLC12A6      GCAATACTGAAAACTTGAGAG----AATGCACCTTAGGGGAAGGGGC----TATTTCAAG
Slc12a6      GCAATACTGAAAACTTGAGAGAGAAAATGCACCTTAGGGGAAGGGGCCAAATGTTTCTAG
             ********************    **********************    * **** **

SLC12A6      ACAGAAAACAAAGGAAATA-CCTGCCTTTTGAATAAGATCCCGCTTTTGAGTCTTACCTA
Slc12a6      AGGGAAGAGGAAGACCTTCTCCTGCCTTTCTTATAAGACCCAGC--TTGTGTCTTACCCA
             *  *** *  ***    *  *********   ****** ** **  *** ******** *

SLC12A6      TGACTTTAC---CAGGGTAGATTAGAAATACACATCCTCCT---GCTGACCCTCTGCCTT
Slc12a6      TGAGTTTATTAGCAGGGCAGGCT-GTAACATTCGTTCTCCCCCAACTGGTCCTCAGCCTT
             *** ****    ***** **  * * ** *  * * ****     ***  **** *****

SLC12A6      CAATAGCTATCTATCTTAAAAGCTGAAGTTTGAACTGCAGCATCTGCTTGACAGGTGCCA
Slc12a6      CCGTAG----CTACCTTAAAAGCTGGAGTTTGAATTGCAGGGCCTGGTTGACAGGTGCCA
             *  ***    *** *********** ******** *****   *** **************

SLC12A6      GGTTCCTT-CGGCAGGGGGACGATCACT-CTATATATCTTCCGTTGCCTCAGATTCCTGT
Slc12a6      GCTTCCTCGTGGCAGAGCAAAAGCTGCTGCTTTTCATATTCTG-TGCCTCCTTTTCTTGT
             * *****   ***** *  *      ** ** *  ** *** * ******   *** ***

SLC12A6      GG-CCCAAGGATCCCACCAATCCTCGTTCCCCCTAAACATGCTAACA-AAAATCCCTTAT
Slc12a6      GGCCCCAAGGATCCCACCAATCCTCATTCCCCCTAA--ATGTTAAAAGAAAATTCCTTAT
             ** ********************** **********  *** *** * ***** ******

SLC12A6      CGTGGGTATTAAAATAATAACAGTTACACTGTATGCATATTTATGTGCTCCTTTT---GT
Slc12a6      TGTGGATATTA----------AGTTACACTGTAAGCATATTTACATGCTCTTTTTCCCCC
              **** *****          ************ *********  ***** ****

SLC12A6      CTGGTTTTTCTTTTCATCATGTATAAGCTGAATTCAGCATTAGTTTCTCACATCTTCCCC
Slc12a6      CTGGTTTTTCTTTTCATCATGTATAATTTGAATCCAGTGATAG--TCTCACATCTT----
             *************************  ***** ***   ***  ************

SLC12A6      CAGGTATCCCCAACAGAATTTTTATGTCCCAGCTTGTATTAAATAGAAGTGAAATATTAA
Slc12a6      ---------CCAAAAAAGTCT----------GCTTATGTGATATAGAAGAGAAATATTAA
                      **** * * * *          **** * * * ******* **********

SLC12A6      GGAAAAT----AAGGAACTTGTGCAACTTTTTTTATGCATTGTTCTCAACCATTTAATTT
Slc12a6      AGTAGACTGAAGGGGAACTTGTGAAAC----ATTAAGCATTGTTCTCAACCGTTTAATTT
              * * *       ********** ***     *** *************** ********
```

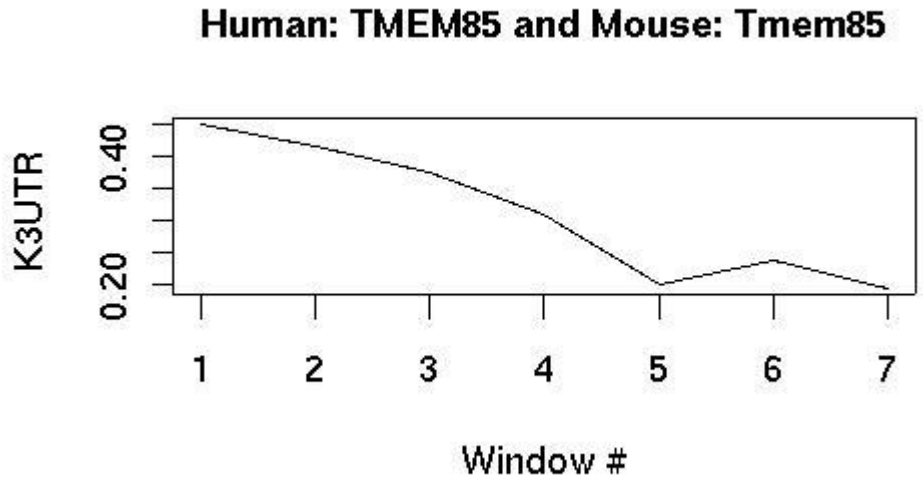
```
SLC12A6      ATTGAAAGGAGATGCTGCAACAGTTCTTGATTTAGCAGCAGTTATTCTCTTGTTTACATA
Slc12a6      ATTGAAAGGAGAAGCTGCTAC----------TGAGCAGCTGCTATTCTTTTGTTTACATA
             *********** ***** **          * ****** * ****** ***********

SLC12A6      GTTATGTTTTTTTGTTGTTGTTTTGCTCTGTACTGGAAACAAAAATAAAGTTTTCTACAT
Slc12a6      G---AGTCTGGTTTTGTTTGTTTTGCTCTGTGCTGGGAACATAAATAAAGTTTTCTACAT
             *    ** *  ** *  ************** **** **** ******************

SLC12A6      TATTTTCA----------------
Slc12a6      TATTTTCAAAAAAAAAAAAAAAAA
             ********
```



MUSCLE (3.6) multiple sequence alignment

```
SLC25A26     -------AGCAGAGACAAGCCTCACCTCCACTTCTGTCAAGAGAGGGGCCTGCAGTGCAA
Slc25a26     GCTTGCTGTTAGAGGTGGGC--------------------AGGAAGAGCCCATGAAGCAG
                       ****    **                      ** * ***      ***

SLC25A26     ACCCTCTTCCGCTGAGCAG--CTGTCTGAACTATAGGCCCCAGTG-------CTGAAGAC
Slc25a26     ACACTACATGTCCTTGCAGTCCTTCCTGTGCCTGGGGCTCCAGTGTGGCTCCCTGAAGA-
             ** **      *   ****  **  ***  *    *** ******       *******

SLC25A26     CAGTTGTGCTAAGATACCGGCATGGAGATTGTGCCATCCGTGGTATAGGCTGGCTGGTAT
Slc25a26     TGGCTGC----AGAGAGCAGCTTAATGAAGG------CCAAAGCACAGGCTGCC------
               * **     *** * * ** *   **  *      **   * * ****** *

SLC25A26     GAAGTCATTGGCCTGTATGCCAGAGAGCTAAGAGAAGAAAACGGGGTCTGTGGCGGTACT
Slc25a26     -------------------CCAGGAAACT----------------TTTCTTGGTGAAACT
                              ****  * **                *   *** *  ***

SLC25A26     CTGAACAATTTCCTCAGAACCTCTTAATAAATAAGTTTGGTAATGCTGAGGCCAGGCCTT
Slc25a26     -GGGGCCACCGTCTCAG-TCCTCTCAATAAATACTCCCGGTAAGGCCAAGGCCAGGCCCT
               *  * *    *****  ***** ********     ***** **  ********** *
```

```
SLC25A26      TTAGAGCTTTCATTTGATCTGTATCTG-ATCTTTCATTTCCTGCCACCTGATGGTGGATT
Slc25a26      TCAGAA--------TGCTTTTTATTAGCATCCTGTGCTTTCTGCTGTCTGTTGGAGTCAC
              * ***         ** * * ***  * *** *    ** ****   *** *** *

SLC25A26      CAGCAGAAGGCAAGATGGTTATAATTCTAAAAGAATAGCTTGTTTGTTTGTTTGGGAAAA
Slc25a26      CAG--GAAAGCAAGCTGGCTACAACTGC-AGTGAATGAC--------CAGACTGGGAAAA
              ***  *** ***** *** ** ** *   *  ****  *          *  ********

SLC25A26      GGAGACTTGGGGAAGAGTTGTGTATGTGGGTGTTTCTCCCCCTAGTTAATTCCTGTTGTG
Slc25a26      GGAAG--GGGGCAAGAACCGTGCGTGTGGCTGTTTCTCCCCTGGGCGGGTGCCTATTGTG
              ***     *** ****   ***  ***** **********   *    * *** *****

SLC25A26      TAAGGGTA-GGCTTTGTTGAAAAAGAAAGAAAGATTGAACTACAGGTGCATAGCAAGCAC
Slc25a26      TAAGATTATAGCTTTGTTTAAAAAGCA--------TGAGCTACAGGTGCATAGCATGCAT
              ****  **  ******** ****** *        *** ***************** ***

SLC25A26      TCTTTCTGGGTAACTAGGCTGCTGGTTTTAATTACCCTCAGATTTCACCCATAAAAACGC
Slc25a26      TCTTTCTGG----TGAGGATGCTGTTTTTAATTA-CCTTGTATTTCTCCTATAAAAACAC
              *********      *** ***** ********* ***   ***** ** ******** *

SLC25A26      ACAATTGTATTATTTTACAGAGATGTGTCCAGCGCCCCCTGTGGTGTGTGAGAGAAAG--
Slc25a26      ATGATTATATTACTTTACAGCAAAGTGCCCCAAAGTCCCTGGAGTGTGTGACAGACAGCA
              *  *** ***** *******  * *** **      *****  ******** *** **

SLC25A26      --CAGCTGCAACTCAAGTGACTAGGTGGGCCCAGCTGGCTTCGTGCAGGAGGGCACGGTG
Slc25a26      GACAGCTGCAGCGCCAGTGAGTGGGTGGGCCCAAGTGACTTCAGTTAGGA--ACATGTTG
                ******** * * ***** * **********  ** ****    ****   ** * **

SLC25A26      GGTGAG-------CCATTCTCGCCATTCTCATGTCAGACTGAAAGGAGGGCCTGGGCCAG
Slc25a26      AGGGAAGCAAAGCCCATGCCCACCACTGTC-CGTC--CCTGGAGGCCAG--CTGGAGCAG
               * **        **** * * *** * **  ***   *** * *   *  ****  ***

SLC25A26      ---CTTTG----------AAAAGGCAGGATGAAATGGAAAGGTCACCACACTTAGGGATT
Slc25a26      CTCCTTTGTCCTCATGGTATATGCCAGCAGGAGGTGGGAGGGTCAGCATACTTGGGGAGC
                 *****          * * * *** * **  *** * ***** ** **** ****

SLC25A26      TTAGACCTTGACTAACAAGCTCCAGGTGTAGAAAAATTCAAAACAAAATGTCAGGAATCT
Slc25a26      ATAAGGCTTGACTAACCAACTCCATGTGGAAAAAAA---AAAGCTATCTG--AGGAATGC
               **   ********** * ***** *** * *****   *** * *  **  ******

SLC25A26      AGCAGTGTTGTCTGCCCTGGAGCAAACAAACAGTATGTGATTT----TGCTTCGCCTATT
Slc25a26      AGTGCCACTGTCT-TTCTGGGGC-----AACAGTGTGTAGTCTATGCTGCTTCGTCTAAG
              **      *****   **** **     ****** ***  * *    ******* ***

SLC25A26      TTTTTTTTCTTTTTTGGGGGAAGATAATTAAAGGCAGAATGACTGCGTTTGTAAAAGAAG
Slc25a26      TATTCTCCTTTT------GGAGAAACATTAATAGCAGAATAA----------AAGAGCAG
              * ** *   ***      ***  *  *****  ******* *          ** ** **

SLC25A26      -GACCACCAACTATACTGA-CATTTATAAATGAACCTTTATTAAAGACACTTCAATGCCA
Slc25a26      TGACCGCCGCACACACTTAGTATTTATAAATGAACCTTTATTAAAGACACTTT-ATGCCA
               **** **    * *** *  ******************************  ******

SLC25A26      TTTGTTAGACACTTCAATATTTTACATGGTTTTCAATGTACACTGTACCAAAATTTCTAT
Slc25a26      TTTGTTAGACACTTCAATATTTTACAT-GTTTTCAATGTACACTGTACCAAAATTTCTAT
              ************************** ********************************

SLC25A26      AAATAAATAACTTTGTACATAAAAGT---------------AAAAAAAAAAAAAAA
Slc25a26      AAATAAATAACTCTGTACATAAAAGTAATACTTCCTCTTTCAAAAAAAAAAAAAAA
              ************ *************               ***************
```

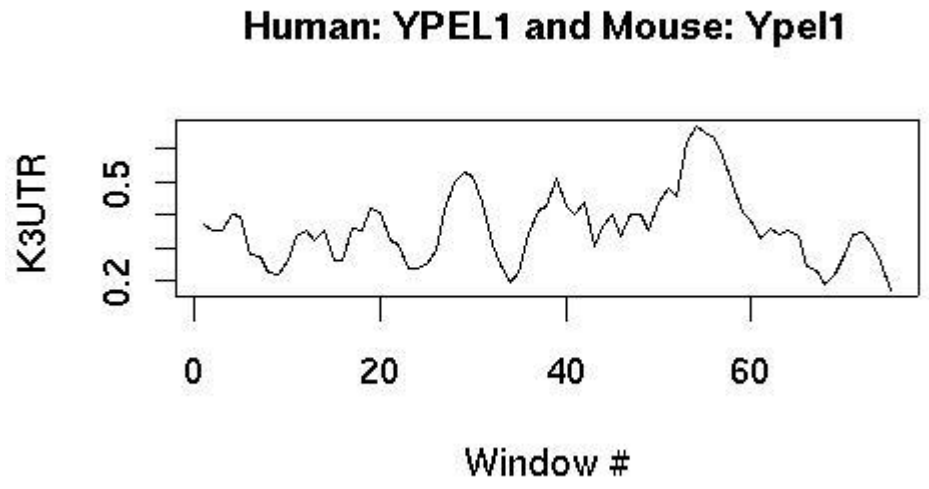


MUSCLE (3.6) multiple sequence alignment

```
TMEM85      A-CATGAGAAAGCAGCGCCTGGTCCCTATGTATTTGGGTCTTATTTACATCCTTCTTTAA
Tmem85      ACCCCCAGAAAGCA-----------CCATGCATATGGACCTCATCTGCAT-----TTTAA
            * *   ********           * *** ** ***  ** ** * ***     *****

TMEM85      GCCCAGTGGCTCCTCAGCATACT--CTTAAACTAATCACTTATGTTAAAAAGAACCAAAA
Tmem85      G-CTGTTGACTCCTAAATATCCTCCCTTAAAACAGTCATGCATGTGGAGAAGAGCC----
            * *   ** ***** *  ** **  ******  * ***   ****  * **** **

TMEM85      GACTCTTTTCTCCATGGTGGGGTGACAGGTCCTAGAAGGACAATGT---GCATATTACGA
Tmem85      ---------TCCTGTGGTG-----------------AGGCCAGTGTGCAGCATGTTACAG
                       *  *****                 *** ** ***   **** ****

TMEM85      CAAACACAAAGAAACTATACCATAACCCAAGGCTGAAAATAATGTAGAAAACTTTATTTT
Tmem85      CAGA---AAAGAAACTACAGCACAACTTATGACTGAAAATAATGTAGAAAACTTTATTTA
            ** *   ********** * ** ***  * * ****************************

TMEM85      TGTTTCCAGTACAGAGCAAAACAACAACAAAAAAACATAACTATGTAAACAAGAGAATAA
Tmem85      TGTTCCCAGCACAGAGCAAAACAAACAAAACCAGAC---TCTATGTAAACAAAAGAATAG
            **** **** **************  * **  * **    ************ ******

TMEM85      CTGCTGCTAAATCAAGAACTGTTGCAGCATCTCCTTTCAATAAATTAAATGGTTGAGAAC
Tmem85      CAGCTGCT----------CAGTAGCAGCTTCTCCTTTCAATAAATTAAACGGTTGAGAAC
            * ******          * ** ***** ******************** **********

TMEM85      AATGCATAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Tmem85      AATG--------------------------------
            ****
```

<Human: TPSG1, Mouse: Tpsg1; no graph, UTR too short for sliding window>

```
MUSCLE (3.6) multiple sequence alignment

TPSG1       --------TGGCAG-GAATCCAAGTGCATTTCTTAAATAAGTTACTATTTATTC-----C
Tpsg1       ATGAGGTGTAGCAACCAACCCAAGTGTCTTTCTTAAATAAGTTAGTGTTTATTCAGCTTT
                    * ***   ** *******  *************** * *******

TPSG1       GCTCCGCCCCCTCCCTCTCC----------CTTGAGAAGCTGAGTCTTCTGCATCAGATT
Tpsg1       GCTTTGCCCCTCCCCTCCCCTTAGCTTTGACTTAGGAAGCCAAGTTTTCTGCATCAGATT
            ***  *****  ***** **          ***  *****  *** **************

TPSG1       ATTGCAACATTTAACCTGAATTTAAA------AGCACACGAAACAAA------------A
Tpsg1       ATTGCAACATTTAACCTGAATTTGTAGAACGGATGACATAAAGCAAATGGATGTCAAATG
            **********************  *      *  ***  ** ****

TPSG1       AAAAAAAAAAAAAAAAAA
Tpsg1       AAAAAAAAAAAAAAAAAA
            ******************
```




```
MUSCLE (3.6) multiple sequence alignment

YPEL1       -TGTGCGAACTTTCCCTTCTCCTTTGAATGCTGTTTTGTGAAAGAAACTGTGAATGTAAT
Ypel1       AAGTG-AAGCGGTCCACTCCTGTTTGAATCCTATTTTATGAAAGC-----TGAGTGTACA
              ***  * *  ***  **   ******* ** **** ******      *** ****

YPEL1       GGAAACGTAGGAGCATC-TGGTGACAGCCTTTCTTGCCCTCTGACCTCAAAGGCTAGCTG
Ypel1       GAAGACCT---AGCTTCTTGGTGACAATGCTTCTTGAACT-TAACCTCACAGGCCAGCCC
            * * ** *   *** ** ********    ******  ** * ****** **** ***

YPEL1       CG-CATAGCTCTTGA------CACTCTCGGCCATCTCTGTGGGTAAGGTGTCCCTCGGAT
Ypel1       AGTCTCAACACTTGAAGGTGTCTCTCTAGGCTATC-CACTTGGTAAGGTGCCCATTTGA-
             * *  * * *****      * **** *** *** *  * ********* ** *  **
```

```
YPEL1      CTGTCCTCTTCGTGTACACAGTTGTTTCTGAAAATTTTCAATGAG--CTTTTTCTAACTT
Ypel1      --ATCCCTGTCGTGTAAATG--CGTTTTTGGAAGTC-----TGAGTTCTTTTTCTAACAT
              ***   ******* *     **** ** ** *      ****  *********** *

YPEL1      CTCAAGTTCTAGAGAAAGAATTAACCAACTGATGACTTACCTGCCTAGTTAATATCTTCC
Ypel1      CTTAAGTTCTAGAGAAAGAATTGACCAGCTGGTGACTCACCTGCCTAGTGAATACTTTCT
           ** ****************** **** *** ***** ********** ****  ***

YPEL1      TTTCACCTTTGTCTTCAATATAGTTGGGCTCTGCTTTTTTAAGGTTCAGTTGAAAACCAA
Ypel1      TTTACCCTACGCCTGCAAT-----------------------------------------
           ***  ***  * ** ****

YPEL1      ACTGGGGCCGGGTGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGATGG
Ypel1      ------GCCTGG---------------CTGTGTTCTCACCA-------AGCCCAAG----
                 *** **               ****  ** ** **       ** *****

YPEL1      GTGGATCACCTGAGGTCAGGAGTTCTAGATCAGCCTGGCCAACATGGTGAAACCCCATCT
Ypel1      ------------------------------------------------------------TCC
                                                                        **

YPEL1      CTACTAAAAATACGAAAATTAGCCGGGCATGGTGGCGAGTGCCTGTAATCTTAGCTACTC
Ypel1      CT------------AAAAGCAGAGAAGCAAGGT---------------------CTCCTC
           **            ****  **    *** ***                     ** ***

YPEL1      AGGAGGCTAAGGCAGGAGAATCACTTGAACCTGGGACACGGAGGTTGCAGTGAGCTAAGA
Ypel1      A-CAAGCAAAGGCAT--------------------------AGTTAC------------
           *  * ** ******                            *** *

YPEL1      TCATGCCATCGCACTCCAGCCTGGGGGACAAAAGTGAGACATCGTCTCAAAAAAAAAAAA
Ypel1      ----------------------------CAGAAGTAAAGCAT-GTCTCCAAAAACTGAAA
                                     ** **** *  *** ***** *****   ***

YPEL1      AAAAGCTGGGTATGGTGGCGCATGCCTTTAATCCCAGCTACTCGGGAGGCTGAGGCACGA
Ypel1      ---------------------------------CCAGCCACTGG----------------
                                           ***** *** *

YPEL1      GAATCACTTGAACCCAGGAGGCGGAGGTTGCAGTGAGCCAAGATCGCGCCACTGCACTCC
Ypel1      ------------------------------------------------------------


YPEL1      AGCCTGGCAATAGGGCGAGACTCCGTCTCAATTTAAAACAAAAGAGAACCAGACTGAGTC
Ypel1      ---------------------------------TAACATGGAAGAGAAGC----------
                                           *** *   ******* *

YPEL1      TCTGAAGACCACAGGGACAGGGTCTCTTTAGATAGCAAGTCTCACCATTCCCTTTTTTAG
Ypel1      ----------------------------TAGGAAGCAAG------CATT-----------
                                      ***  ******      ****

YPEL1      AGAAAAGGTATTGTAGCCCACCCTCCACCCCGCTGTTTTTCTTAAATTTGCAGAACTTCA
Ypel1      ------------------------------------------------------------


YPEL1      AATTGGCTATTCCTCTTGCAAATGAACCCTTAAAGTACAGTGTTATTTAAGAATCTTCCA
Ypel1      ---------------------------------------CAGTGT-----------------
                                                 ******

YPEL1      GAGGCAGTCAACAGACTTATACACTAAGGGCATTTTTGGTTTTTAGCTTGTTCAAAAACA
Ypel1      ------------------------------------------------------------


YPEL1      GAGGCCAGCACAGATGACATTTTAGATACACTCTAAATTGAGAATGGTGTCTAGTGGAAC
Ypel1      ----------------ACATCTTAAGCA--------------AATAG-------------
                          **** ***   *              *** *
```

```
YPEL1      ATGTTTATTTAAGCCAGTAGATTCCTTATCTAGAAAGCAGGTGAGCTAGCCCTTAGAGAA
Ypel1      ----------------GGAGGTTCCTGATGTAGAAGGCGAGTGACTTACTGGTGGCTGAA
                          * ** ***** ** ***** **  ****  **    *    ***

YPEL1      GG--CTGTCCCGGGGCC-CGCAGAGGTGCCCTTACTGAGGTGACAGCCTCACAGGGTCTG
Ypel1      CGCTCTGTCCCAGGGCTACCCAAGTCTGCC---ACTGAGCTGACTCCCT---------TG
            *  ******* ****  * **    ****   ****** ****  ***         **

YPEL1      GTACCAGGGGTTGTGCCCTCAGCAGTGACAGCAGCTTAGGTGTCAGGCAGTTGCTGAGTG
Ypel1      GTGCTGCGG----------CAGC----ACAGCGGCTCAGA----ATGTAG---------G
           ** *   **          ****    ***** *** **     * * **         *

YPEL1      GCTGGTCCATGTCTATAGAGTAACACACTGGACCGAGGAAAAGTCAGATTTCATTTTCTA
Ypel1      GCTGGTCCCTG-CCACAGAGCAACAGGCTG----------------GGTTTCACTTTATA
           ******** ** * * **** ****  ***                * ***** *** **

YPEL1      CCCTGGATGTACTTGAAGAAAAAGAATTATTTTTGCATATGAAAGAGGCCAGAACCCACA
Ypel1      CCCTGGATGTACTTGAAA---------TAGTTATGCATAC---AGAGGCCAGAGCCAGTA
           **********          ** ** ******    ********** **   *

YPEL1      GGAAAAACTTCAAAACTTGACATTTGCCAGAATGTTTAAAATTTGTTCAGAAAAGGTTAA
Ypel1      CGGGAAACTGCAAAACATG---------AGAGTTTTCCAAG---GTAAAGGAAGAGTTAG
            *  ***** ****** **         *** * **  **    **  ** **  ****

YPEL1      AGCAACAAGTTTAGCCTTTGTGCATGAAGACGCCTGGCCTGCTAGACGCGTTGCCCGTCC
Ypel1      ---------------CTTAGTATGGGA-------------GCTA----------CCTTTT
                         *** **    **             ****          ** *

YPEL1      CTGCGTGGTGCTGTCCCATGTCACTTGAACTGATA--GAGGGGCCTGTGCAATCTCCTAA
Ypel1      CTGCCTGGAATCACCTCA-------TGACCTGACAAGGGAGGATCTGGGCCATCCCCCAA
           **** ***      * **       *** **** *  *  **  *** ** *** ** **

YPEL1      GGCCTGTGTTTCTGCCATATATTTTATTATAAATTACAATCCACTCATCCACCTGCCCTC
Ypel1      GCCCAGGCTCTCAG-----------ATTTTAAGTCACAACACA-TCAT------------
           * ** *  * ** *           *** *** * ****  ** ****

YPEL1      CACCAGGAGTGGGCACCCCATAAGGGTTTAGGCCACTTTGC--AGAGGATGGAGGTCAAA
Ypel1      --------------ACTCTATAAGGAGT--GGCCAATTTGCAGAGAGGATGGAGGCC-AG
                         ** * ******  *  ***** *****  ************ * *

YPEL1      AACCACTCCCAGATAAGTTTGGTTTTCAACATTTAGTAACTTGTCTCAGGGCAG-AGGGC
Ypel1      AACCATTCCCAGATAAG------TTTCATTGTTTTGTACCCT-CCCCAGGGCAGAAGGGC
           ***** ***********      *****   *** *** * *  * ******** *****

YPEL1      AGGCAG---------GGGGACCGAGGGGC----AGCAGATAGGAGAGCACTGAGCCCGGA
Ypel1      AGGCAGTAGGCCTGTGAAGGCTGAGAGCCCTTGAGCTGCAAAGGCAGCATTA---CAGGA
           ******         *  * * *** * *    *** *  * *  **** *    * ***

YPEL1      TAGTTCTCAGCCTGGCAAGTGGCTCTGAAGCTGCCTTCAGACAAGGCTAGTCTAGGGGCA
Ypel1      AAGCGCTTAGC---ACCGGCTGCTCTAGAGCTGCCT------------------------
            **  ** ***    *  *  *****  ********

YPEL1      AGAGTGCGAGCTGGCTGACAATAAGAACGTGGCCACCTGCCCAGCTTCACACCTC----C
Ypel1      -GGGTG-----TGGTAGTCAAAAAGCA-GTTTCTATCT----------AGACCTTTGGGC
            * ***     ***  * *** *** * **  * * **          * ****     *

YPEL1      CCCGACTTCAGCCCTTCCTAAACCCAGACCTGCGGTCCAG-GCAGGCACTGGGCTGTGCC
Ypel1      TCCAGCCCCAGACCTG-CTAAACCCACATCCTCTGGCCAGCACAGGCACTCG-------C
            **  *  *** ***  ********* * *  * * ****  ******* *       *

YPEL1      CACTCGAGCTCACTGCCCACACACAGCATGCCTTTGGGTGCCATCTCTTTGCCCAAGCCT
Ypel1      TACTCCTCTTCACTT------CAAAACCTACATTTTGGTGCCATTTCTTTGCCCAAGATT
            ****    *****       ** * * * * *** ******** ************  *
```

```
YPEL1      GGAAGCCTTGGCAGGTGGGAAATGCCGCTGCCCTGGTGGGCATGGCACTGAGATGCATCC
Ypel1      GGAAGGCTGGTAATACAGGAAATGCCTGCGCAATGCTGGTCCTAGCCC------------
           ***** ** *  *    *********   **  ** *** * * ** *

YPEL1      ACTCAGCAGGAGTGACAGAGGCAGAAGTTCCTTTAAAGCACATCTTCCACTTAGGAAAGG
Ypel1      -------------------------AAGCTGCTTT----------TTCC------------
                                  *** * ****          ****

YPEL1      AAGGAAATCTTTGTACTGTCTTGGAAGCCTCCACATCCGGCTATGGCCCTGCAAGCTGCT
Ypel1      ----------------------------------------------------CCAGCTACT
                                                              * **** **

YPEL1      TTATCCCTGCGCTAGTCTCCCCCGAGGGTTTAGGCTGGCCCAGCACATCCTGTCCTCCTG
Ypel1      T----------CCAGTTTCCCTCTA-GACTAAGGCTGGCAAAGCAAA-------------
           *          * *** **** * * *  * ********  **** *

YPEL1      AGCTCGCGTGCAGCCACCCAGAGCGCAGGGGTCACTGCACGCTGCAGGGCTCTTGCTGCC
Ypel1      -------------------------------GTGACTCTAC-CAGTAAGGTAAGTGCT---
                                        ** ***  ** * * * **    ****

YPEL1      ATGGTCTCAAGCCTGAAGAGGCTCCGCCCACAAGCTGGCCCATGAAGTTAGCAATGCCTG
Ypel1      -----------------GAAACTCAG--CACAAGCAGCCGTGTGACCTTGGCA-------
                         **  *** *  ******* * *   ***  ** ***

YPEL1      TGGCTTCAGTCAATTGTCTTGAGACTGTGAAGAGGCTGAAAGACACCTTCCCGGGTGGAA
Ypel1      ---------------------------------AGGCTCAGGGCCATCCTT----ATCTAG
                                           ***** *  * ** * *      *  *

YPEL1      GAAGGAGTTCACTG---AAAACTTATCTTAAACTGACCCTTCCCTTTGAGTGAGTCTTCA
Ypel1      GAATGAA--CACTGGTCAACAGTCTGCTCAAACTGGACATGCCCGCCTGGCTAGTTTTTA
           *** **   *****   ** * *   ** ******  * * ***     *  *** ** *

YPEL1      TTCCTCTCCCATGTGGGAACCCAGCCTCCGATGCCCCGGGGACTAGGGGAAACAGTTGGA
Ypel1      ATACTTTTCAGAGTGGGAACTTGGTGGCTGAGGCATGGGCG--------------------
            * ** * *   ********   *   * ** **   ** *

YPEL1      GGTTCGTGCCGTCCCCAGCCTGCCACGGGTGCGAGGACAGCCAAGTCCTGAGTGACTCA-
Ypel1      -------GCTTTCCCCAG--TGGGAAGGCGGCAAGGGAGAC--AGACCCGAGTACGTCCT
                  **  *******  **  * **  ** ***    *  ** ** ****   **

YPEL1      TGGCTTCAGTCAATTGTCTTGAGACTGTGAAGAGGCTGAAAGACACCTTCCCGGGTGGAA
Ypel1      ---------------------------------AGGCTCAGGGCCATCCTT----ATCTAG
                                           ***** *  * ** * *      *  *

YPEL1      GAAGGAGTTCACTG---AAAACTTATCTTAAACTGACCCTTCCCTTTGAGTGAGTCTTCA
Ypel1      GAATGAA--CACTGGTCAACAGTCTGCTCAAACTGGACATGCCCGCCTGGCTAGTTTTTA
           *** **   *****   ** * *   ** ******  * * ***     *  *** ** *

YPEL1      TTCCTCTCCCATGTGGGAACCCAGCCTCCGATGCCCCGGGGACTAGGGGAAACAGTTGGA
Ypel1      ATACTTTTCAGAGTGGGAACTTGGTGGCTGAGGCATGGGCG--------------------
            * ** * *   ********   *   * ** **   ** *

YPEL1      GGTTCGTGCCGTCCCCAGCCTGCCACGGGTGCGAGGACAGCCAAGTCCTGAGTGACTCA-
Ypel1      -------GCTTTCCCCAG--TGGGAAGGCGGCAAGGGAGAC--AGACCCGAGTACGTCCT
                  **  *******  **  * **  ** ***    *  ** ** ****   **

YPEL1      ---AGATGCTTCACTTACATGGAAGAAACTTCTAAAACTCTACCGAGTGGTTTTTGTATA
Ypel1      TATAGATGCCTCAGTAAAGCTGAAGAGTCTTGGAGAAGTC-----------TTCTGTATA
              ****** *** * *    *****  ***  * ** **           ** ******
```

```
YPEL1      TACTAAAGTTCTATTTAGAGCTTTTCTGTTTTGGGCAAGTTCGCTGCTCCTTCTATTTGG
Ypel1      TACTAAGATTATATTTAGA--CTTTATGCTCTGGGCATGCTATCTGCCCCT--CAACTGA
           ******  ** ********   *** ** * ****** * *  **** ***   *  **

YPEL1      GCACTTTGGTTTTTGTACTGTCTTTTGTGACGGCATTGATTGAACATTTTTTACTAGTAG
Ypel1      GCAC-------TCTGTATTATCTTTTGTGAGAGCTCTGACTGGAACGTTTTTGCTAGTAC
           ****       * **** * **********  **  *** ** *   ***** ******

YPEL1      TCT-TATGACTT--TTGTATTTTTTTTTTTTTTTGTAATTTATACCAACAACACTTTTAT
Ypel1      TCTATATGACTTACCTGTA--------TTGGGGAGTAATGGATATAAACAAACCATTTAT
           *** ********   ****         **     *****  ***  *****  * *****

YPEL1      CACTTTTTTTTT---------GTTGGGCTTCTGCAAAATACAAGCTCATTTTTAAACCAA
Ypel1      CAAGGTTTTTTTGGATTTTGCGTTGGCCTTCTGTGAAATA------------TAAACCAA
           **   *******          ***** ******  *****             ********

YPEL1      ATGAACAGACCATGAGCTGGCTTCAGGGGAAGTGCTATTCACAGGACCATATCCA-----
Ypel1      AAGAACTGACCATGAGCTGGCTTCAGGGTAAGTGAT----GCAAGACCACAACCAGGCAT
           * **** ********************** ***** *     ** ***** * ***

YPEL1      ----------------------CCACCCTCTTAAATTCCTAAACAATATCA-TCTAGGACT
Ypel1      GCTGGTAAACCACTGAATGTCTCACCTCCTTACATTTGTAAACAGCACCATTCTGTGAC-
                              ****  **** ***  ******  * ** ***  ***

YPEL1      TCTATTTAAGTTATTTAAAATAAATCTTCCTTGAGAGCCTTGGGAGGTGATGTCAGGGTT
Ypel1      -----------TATTTTAAAGAGG--TTTCCTCAGAACAGTGGGAGGTGATACCA-GGTC
                     ***** *** *    ** * * *** *  **********  ** ***

YPEL1      ATAAATGGCACAGTGCATTTGCTGTAGGAATGTGGTTTG----------GCATTGTT-TT
Ypel1      ATGAACGCCAC------TTTGCTTTAGAAATGTGGTTTATTTCCTTGTTGTCTTGTTACT
           ** ** * ***      ****** *** **********           *  *****  *

YPEL1      ATACACACAGTATTTTTTATACCTTAATGCTTATTCTTGATGGCATCTGTCAGATATTAG
Ypel1      GTATACATAGTA-TTTTTATACCTTAATGCTTATTCTTGATGGCATCTGTCAGGTATTTG
            ** *** **** ************************************** **** *

YPEL1      AATTGAAAATAAGAATCTTCCCAAAATCCTTTAATTTACCTGATGCCCTCATCAGGTCGT
Ypel1      AAGAGA------------------------------------------------------
           **  **

YPEL1      TAAAAATTCAAATGGTTTTAATAGCTAAAAAACTACAAATTAAGCTCTAAAACAAACAAA
Ypel1      ------------------------------------------------------------


YPEL1      CTACAGAAATGTAAACCTTCATTTGCCAAAGGTCCTTGGTGGCCTGTCCCCTGCCCTGGG
Ypel1      --TCAGAAATG------------------AGATTTTT-----------------------
              ********                  ** *  **

YPEL1      AGCAGATGGCCCTGAAGCCCTTCCCTCACTGTGCAGGCCACCGGGTGAGGCTGGACGGTC
Ypel1      ------------------TTTCCCCTCG---------------------------------
                            * *****

YPEL1      ACCCATGGTGGCTTCACTGCAAGGAGCAGGACTGCCGAGCTCAAGCACGGGGCCTTCAGC
Ypel1      ------------------------------------------------------------


YPEL1      TTCCCCTGTCCTCTGGCCACACCGCCAGCCCTTGGTCCTTATCTGTGTGAGGTTTACAAA
Ypel1      ------------------------------------------------------------


YPEL1      TAAAGCTTCTGATGTCAAATGTTTAAAAAAAAAAAAAAAAAAA
Ypel1      -------------------------------------------
```

VITA

Name: Charles Michael Dickens

Address: Department of Animal Science
Kleberg , Mail Stop 2471
Texas A&M University
College Station, TX 77843

Education: Ph.D., Genetics, Texas A&M University, 2009
A.A.S., Computer Information Systems, San Jacinto College, 2004
M.S., Agronomy, University of Arkansas, 2001
B.S., Agronomy, Texas A&M University, 1991