Published online 9 November 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 6* ***1967–1979***
doi:10.1093/nar/gkq949

Ultraconserved cDNA segments in the human transcriptome exhibit resistance to folding and implicate function in translation and alternative splicing

J. Fah Sathirapongsasuti[1,2,3,*], Nuankanya Sathira[1], Yutaka Suzuki[1], Curtis Huttenhower[3] and Sumio Sugano[1]

[1]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan, [2]Department of Mathematical and Computational Science, Stanford University, 390 Serra Mall, Stanford, CA 94305-4065 and [3]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115, USA

Received June 14, 2010; Revised September 18, 2010; Accepted September 29, 2010

ABSTRACT

Ultraconservation, defined as perfect human-to-rodent sequence identity at least 200-bp long, is a strong indicator of evolutionary and functional importance and has been explored extensively at the genome level. However, it has not been investigated at the transcript level, where such extreme conservation might highlight loci with important post-transcriptional regulatory roles. We present 96 ultraconserved cDNA segments (UCSs), stretches of human mature mRNAs that match identically with orthologous regions in the mouse and rat genomes. UCSs can span multiple exons, a feature we leverage here to elucidate the role of ultraconservation in post-transcriptional regulation. UCS sites are implicated in functions at essentially every post-transcriptional stage: pre-mRNA splicing and degradation through alternative splicing and nonsense-mediated decay (AS-NMD), mature mRNA silencing by miRNA, fast mRNA decay rate and translational repression by upstream AUGs. We also found UCSs to exhibit resistance to formation of RNA secondary structure. These multiple layers of regulation underscore the importance of the UCS-containing genes as key global RNA processing regulators, including members of the serine/arginine-rich protein and heterogeneous nuclear ribonucleoprotein (hnRNP) families of essential splicing regulators. The discovery of UCSs shed new light on the multifaceted, fine-tuned and tight post-transcriptional regulation of gene families as conserved through the majority of the mammalian lineage.

INTRODUCTION

Evolutionary conservation is the hallmark of comparative genomics and has been regarded as an indication of function (1,2). Up to 5% of the human genome appears more conserved than expected and is believed to be under purifying selection (3). In 2004, Bejerano *et al.* (4) discovered 481 ultraconserved elements (UCEs) defined as genomic sequences longer than 200 bp, which are perfectly conserved (100% identity with no insertions or deletions) among the human, mouse and rat genomes. Although several other sets of conserved elements were identified, such as conserved non-genic sequences (CNSs) (2), long conserved non-coding sequences (LCNS) (5), multiple species conserved sequences (MCSs) (6) and highly conserved regions (HCRs) (7), UCEs are regarded as a class of genomic elements with the strongest level of conservation. Although originally only 225 (47%) of the UCEs were classified as transcribed (exonic) or possibly transcribed (possibly exonic) by overlapping with coding exons (4), an experimental survey found evidence for transcription of 325 (68%) UCEs and called them transcribed UCEs (T-UCEs) (8). While many of the T-UCEs overlap coding regions (i.e. exonic/possibly exonic UCEs), some are in intergenic regions or have short open reading frames (ORFs) and, hence, were believed to function as non-coding RNAs. T-UCEs show

[*]To whom correspondence should be addressed. Tel: +1 617 432 1056; Fax: +1 617 432 5619; Email: fsathira@fas.harvard.edu; jsathira@hsph.harvard.edu

tissue-specific expression and alteration of T-UCEs is associated with cancer, suggesting that UCEs may represent yet another class of functional non-coding RNAs (8).

Recent studies suggest that UCEs play important roles in genome regulatory mechanisms such as acting as distal enhancers (9,10), regulating alternative splicing (11,12) and serving as transcriptional coactivators (13). Non-transcribed UCEs are generally associated with transcriptional *cis*-regulation; for example, they may behave as long-range enhancers in gene deserts (8,14) and as tissue-specific enhancers (10). Transcribed or exonic UCEs are believed to play roles in post-transcriptional regulation such as alternative splicing and mRNA processing (4,11,12). For example, in every member of the serine/arginine-rich (SR) family of key splicing regulators, highly conserved or ultraconserved elements are alternatively spliced either as cassette exons containing premature in-frame stop codons or as alternative introns in the 3′-untranslated regions (UTRs) (11). The resulting alternatively spliced ultraconserved mRNAs are targeted for degradation by nonsense-mediated decay (NMD) (11). Moreover, a genome-scale survey of premature stop codon-containing exons suggested that UCEs are associated with homeostatic control of splicing regulators by means of alternative splicing and NMD (12).

Findings from these previous studies raise the question of whether the original definition by Bejerano *et al.* captures a complete class of 'transcribed' UCEs. First, in a splicing event, both the 5′- and 3′-ends of exons must be recognized. Considering ultraconservation at the genome level disallows conservation to span both ends of a splice site (which will be adjacent in the transcript but distal in the genome) and thus fails to fully capture ultraconservation regulating a splicing event. Furthermore, some UCEs are found to function as ncRNA, i.e. as mature mRNA (8,13). Such ncRNAs should also exhibit conservation at the mRNA level. There are other possible post-transcriptional regulation mechanisms in which ultraconservation may play critical roles; for example, translation suppression by upstream ORF (uORFs) and upstream AUGs (uAUGs), mRNA degradation and formation of mRNA secondary structures. The effect of ultraconservation on these regulatory mechanisms cannot be studied comprehensively without considering it at the transcript level.

As an initial effort along these lines, we present here an analysis of ultraconservation at the transcriptome level. We searched a 1.6 million 5′-expressed sequence tag (EST) sequence library (15,16) for 'ultraconserved cDNA segments' (UCSs) defined as stretches of at least 200 bp of cDNA transcripts that match identically (no in-dels) with the corresponding regions of the mouse and rat genomes. This definition is analogous to that of the UCEs originally defined for genomic sequences by Bejerano *et al.* (4). We report on 3096 ultraconserved transcripts discovered by this process. The transcripts are clustered into 96 non-overlapping UCSs and 19 of these are not included in the original set of 481 UCEs. The UCSs are enriched for RNA processing functions (alternative splicing events) and post-transcriptional regulatory mechanisms (uORF/uAUG, degradation by NMD, RNA stability and miRNA targeting). We hypothesize that UCSs are indicative of regulatory functions operating at the post-transcriptional level, which include alternative splicing, RNA destabilization and RNA degradation.

METHODS

UCS screening and clustering

We started with 1.6 million cDNA sequences from a full-length enriched and 5′-end enriched cDNA library (15,16) and filtered out those sequences with less than three supporting clones to minimize false positives [two clones are regarded as the same if they share the same transcription start sites (TSSs); 1.16 million sequences left after filtering]. We fetched human (hg18), mouse (mm8) and rat (rn4) Multiz alignments (17) of all cDNA sequences, using the option 'Stitch Gene blocks' in galaxy (18). The multiple alignments were scanned for UCS as defined above. 3096 UCSs were further clustered into 96 clusters. Two UCSs are in the same cluster if they overlap each other and the largest sequence is chosen as a representative UCS for each cluster (see Supplementary Methods for explanation and justification).

Evolutionary model

We assessed the likelihood that the UCSs occurred by chance by testing several null hypotheses. Under a simple model of neutral evolution (assuming independent substitutions at each site and the slowest neutral substitution rate observed for any 1-Mb coding region of the genome), the probability of finding just one UCS of at least 200 bp is $<10^{-20}$. One may argue that since the UCSs are mostly from protein-coding regions, they are subject to low non-synonymous substitution and are more likely to occur by chance. However, even using only the mean synonymous substitution rate, the expected length of a run with zero substitution is 27.6 bp and the probability of finding even one UCS of 200 bp is $<10^{-6}$.

Evolutionary depth

Multiz alignments (17) of 17 vertebrate species were obtained through option 'Stitch Gene blocks' in galaxy (18). Conservation level is defined by percentage of UCS sequences in human that are identical to those in respective species.

UCS expression confirmation

Cell culture and RNA extraction. Human embryonic kidney 293 cells (HEK293, American Type Culture Collection (ATCC) number: CRL-1573) were maintained in Dulbecco's modified Eagle's medium (DMEM) (Invitrogen) supplemented with 10% fetal calf serum and antibiotics at 37°C, 5% CO_2. RNA was extracted using the protocol outlined in the RNeasy kit (Qiagen). The qualitative assessment of all total RNA was done utilizing the Agilent 2100 Bioanalyzer.

qRT-PCR. Total RNA (8 ug) was reverse transcribed using an oligo(dT) and the Superscript II reverse transcriptase kit (Invitrogen). Negative control samples (no first-strand synthesis) were prepared by performing reverse transcription reactions in the absence of reverse transcriptase. Expression profile was performed using quantitative reverse transcription polymerase chain reaction (qRT-PCR). Gene-specific primer pairs were designed using Primer3 software, with an optimal primer size of 20 bases, amplification size of 100–500 bp and annealing temperature of 55°C. The primers were purchased from Operon. Quantitative real-time PCR was carried out with 100 pg of total RNA per test well using the Power SYBR Green PCR Master Mix (Applied Biosystem). The polymerase chain reaction (PCR) reactions were performed using an ABI 7900HT Fast Real-Time System (Applied Biosystems) using the following cycling protocols: 40 cycles of 15 s at 95°C and 60 s at 60°C. The threshold cycle (Ct) value was calculated from amplification plots, in which the fluorescence signal detected was plotted against the PCR cycle. We defined express sequence when the threshold cycle values (Ct) <37 in the presence of reverse transcriptase, but not when reverse transcriptase was absented. The product size was checked by 2% agarose gel electrophoresis and melting curves were analyzed to monitor the specificity of the PCR reactions.

Computational confirmation. By construction, each UCS is supported by no less than three cDNA clones. This is to minimize error arising from one-pass sequencing artifact. TSS tag data are obtained from Database of Transcriptional Start Sites (DBTSS) (19). We searched for a TSS within 1000 bp upstream (on the same strand) from each UCS and confirmed that there exists at least one TSS for each UCS. Additionally, to confirm splicing of the UCSs, we used splice junction information from Caltech RNA-seq (20) that includes seven cell lines and Burge Lab RNA-seq (21) that includes 14 tissues and cell lines. Both RNA-seq data sets were retrieved from UCSC Genome Browser (22).

Gene ontology and InterPro enrichment

We used DAVID Functional Annotation tool (23) to assess functional enrichment of RefSeq genes overlapping UCSs. Whole-genome annotation was used as background and DAVID default settings were used. Only Bonferroni-corrected *P*-values were used.

Alternative splicing event association

AltEvent track was obtained from UCSC Genome Browser (22). The association was assessed at two levels: gene level and transcript (EST) level. For gene level, we used the UCSC gene model because UCSC genes contained the greatest number of alternative spliced isoforms relative to other databases, 78.61%, which was closer to the reported number of 80% (24) than RefSeq and ENSEMBLE. First UCSC genes were overlapped with UCSs, resulting in 282 UCS-containing genes. Of these UCS-containing genes, 270 overlapped with AltEvent track and the other 8 have alternative isoforms that do not contain UCS (thus missed by the first screen). Together, these 278 genes constitute UCS-containing genes with alternative isoform. The same procedure is repeated with contiguous and cross-exon UCSs. At transcript level, ESTs from the cDNA library (16) were used instead of UCSC genes. The other steps remained the same. Hypergeometric test (one-sided Fisher's exact test) was used to assess statistical significance.

AS-NMD codon association

A list of mouse stop codon-containing exons was extracted from Ni *et al.* (12) and converted from mouse (mm5) to human (hg17) coordinates and then to hg18 coordinates. Five mouse exons could not be converted to hg17 and four hg17 exons could not be mapped to hg18 coordinates. In total, 66 exons remained after the conversions. Association is assessed by Fisher's exact test as described previously (12).

Enrichment for 5′-UTRs and coding regions

Two null models were tested: hypergeometric and binomial. In this context, the hypergeometric null model is defined over ESTs as we cannot compare UCSs with the genomic background directly. This test assesses the probability of observing at least (or at most) k 5′-UTR [3′-UTR, coding sequence (CDS)] ESTs among n UCS-overlapping ESTs given that there are m 5′-UTR ESTs in the library of N ESTs. The use of ESTs from the cDNA library also helps correct for the 5′-enrichment in the library.

In practice, because of high redundancy in the cDNA library, we did not use all of the 1.6 million ESTs in the library but clustered them into 21 272 non-overlapping clusters and picked the longest EST in each cluster as a representative. As such, one may suspect that the enrichment by hypergeometric test is biased by EST length. That is if an EST is very long, it is more likely to overlap both UCS and 5′-EST. To address this issue, we present a binomial test, which explicitly accounts for length variability of ESTs.

In the binomial model, a test assesses the probability of observing at least (or at most) k 5′-UTR (3′-UTR, CDS) ESTs among n UCS-overlapping ESTs, each of which falls in the 5′-UTR with probability P. The probability P is defined as the fraction of EST sequences that overlap 5′-UTR. As before, the use of EST sequences in defining P corrects for the 5′-enrichment in the library.

RNA secondary structure analysis

We used the program Mfold (25) to compare Gibbs free energy of the best folded structures of difference sets of sequences: exonic UCE, UCS, PhastCons conserved element, random EST, miRNA target and miRNA coding sequences. All fold analyses were performed using a default setting of Nucleic Acid Quickfold (http://dinamelt.bioinfo.rpi.edu/quikfold.php) and energy rule RNA (2.3). Exonic UCE and UCS sequences were fetched from UCSC Genome Browser (22) and input to Mfold directly. PhastCons sequences overlapping UCSC known genes were extracted from UCSC Genome Browser and filtered for sequences longer than 250 bp. 350 sequences of length 250 bp (250 bp being the average length of UCSs) were sampled and input into Mfold. 350 random sequences of length 250 were selected from spliced sequences of all cDNA positions in the library (26). miRNA target and miRNA-coding genes were extracted from miRBase (27). cDNA sequences surrounding the miRNA targets and miRNA-coding genes were sampled and 350 sequences of length 250 bp were used in fold analysis.

RESULTS

We identified 3096 ultraconserved transcript sequences among human, mouse and rat, which clustered into 96 non-overlapping UCSs for which all subsequent analyses were performed (see Supplementary Method for details and justification of the clustering method and see Online Supplementary Material of a BED (Browser Extensible Data) file of all 3096 ultraconserved transcripts). Each UCS is a subsequence of a mature transcript and can overlap multiple isoforms. Their sizes range from 200 to 551 bp (mean 250 bp). Thirteen UCSs are longer than 300 bp and three are longer than 400 bp. The list of the UCSs is provided in Supplementary Table S1 and also in BED format. The UCSs are highly unlikely to appear by chance ($P < 10^{-20}$, see 'Methods' section).

Almost all of the UCSs continue to exhibit very high conservation in other vertebrate genomes. There are 67 UCSs (69.8%) that align perfectly among orthologous regions in the chimp and macaque genomes and 93 UCSs align with 90% identity or higher. Sixty-six UCSs align completely with the dog genome and 92 align at 95% identity or higher. Strikingly, there are 13 UCSs that are completely conserved down to tenrec and as many as 93 UCSs have at least 90% identity with the opossum genome. One particular UCS, F-TESTI2015707 (Figure 1), is conserved with 100% identity in 13 of the available placental mammal genomes (human, chimp, macaque, mouse, rat, rabbit, shrew, dog, hourse, cow, armadillo, elephant and tenrec), with at least 95% identity in opossum, chicken and frog genomes and with at least 85% identity in fish genomes (zebrafish, tetraodon, fugu, stickleback and medaka). Interestingly, this UCS is a cross-exon UCS, which was not included in the UCEs. The gene in which this UCS appears is HNRNPH3, a member of ubiquitously expressed heterogeneous nuclear ribonucleoproteins (hnRNPs), which have been implicated in splicing repression (28) and, together with SR protein family, constitute key splicing regulators (24,29,30). Members of the hnRNP H group also promote alternative 5′-splice site selection, splicing inhibition and exon skipping (31).

qRT-PCR, upstream TSSs and RNA-seq confirm transcription and splicing of UCSs

To verify the existence of these UCSs, we performed real-time PCR in human embryonic kidney 293 (HEK293) cells. We used quantitative reverse transcriptase PCR (qRT-PCR) to validate the expression of 96 UCSs. Out of the 96 UCSs, 49 (51%) show significant expression. A reason why not all of the UCSs were expressed in HEK293 could lie in the tissue specificity of UCSs. Like T-UCEs (8), UCSs may express only in some tissues and cell states. Nonetheless, this provides overall confirmation of the expression of the UCSs. Moreover, in the full-length cDNA library, each UCS was found in more than three cDNA clones, additional evidence that the UCSs are not due to sequencing artifacts. The existence of TSSs in the

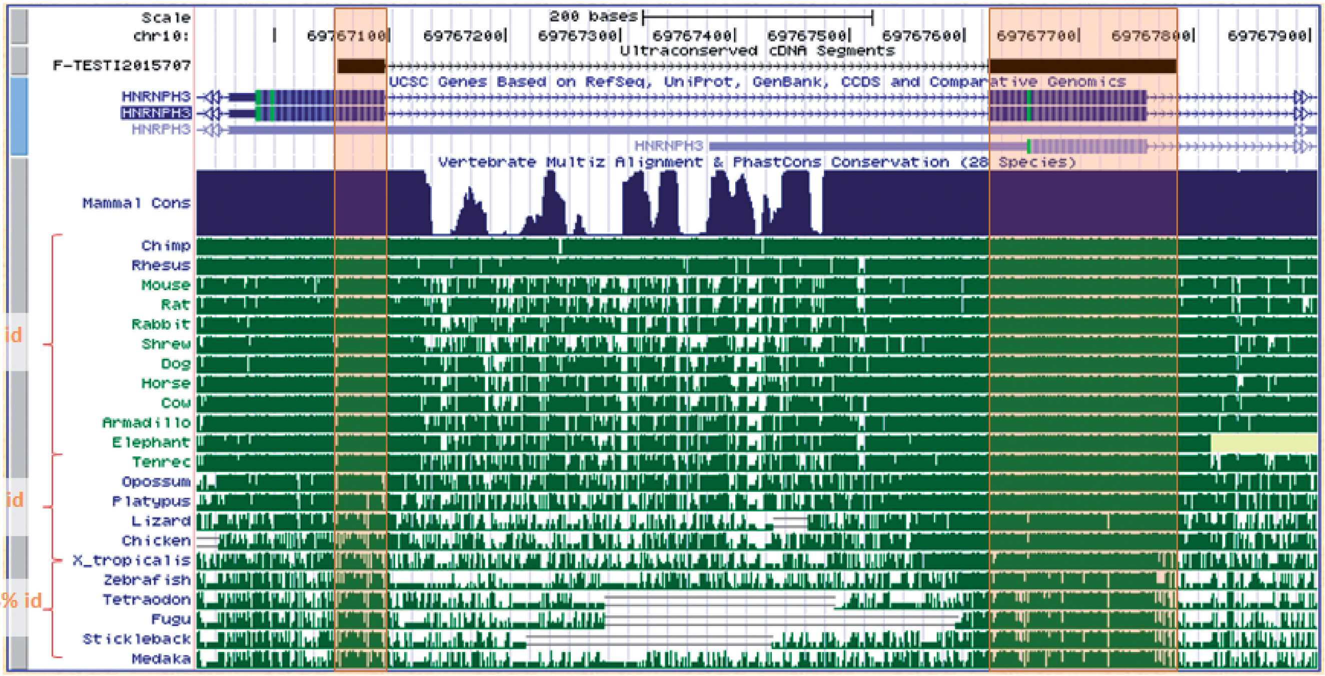

Figure 1. A cross-exon UCS, not previously found in UCE, exhibits extreme conservation across vertebrate genomes and overlaps the start codon of an alternatively spliced isoform.

DBTSS (19) within 1000 bp upstream of all of the UCSs also supports their existence.

To confirm splicing patterns of UCSs, we used RNA-seq data from two sources: Caltech RNA-seq (20), which includes splice junction information in seven cell lines and Burge Lab RNA-seq (21), which includes transcriptomes in nine tissues and five cell lines. All 45 of the cross-exon UCSs (defined and discussed below) have clear evidence of expression in one or more tissues/cell lines. These expression data show two (or more) halves of each cross-exon UCS in one transcript. Moreover, in all but two cross-exon UCSs in the Caltech data, we found splice junctions corresponding exactly to the UCS splice sites. Further investigation of the two cases where no splice junction was found (F-BRAMY2016664 and HPR06694) reveals that they are cassette exons and express in a tissue-specific manner; one was found to express only in the brain and the other in the heart. Taken together, RNA-seq data confirmed the expression and splicing pattern of each cross-exon UCS.

Novel identification of 'split' UCSs

As shown in Figure 1, these transcript-based UCSs differ from UCEs defined using genomic DNA alone in that they can be fragmented across two or more exons. We identified 58 such 'split' UCSs, termed 'cross-exon UCSs'. The remaining UCSs within single exons, termed 'contiguous UCSs', are by definition a subset of UCEs and not a novel class of conserved elements. As expected, contiguous UCSs share many features with UCEs.

Although some UCSs overlap with previously defined UCEs, 19 do not. These 19 UCSs are cross-exon UCSs whose 200 conserved bases come from the stitching of multiple short conserved exons, a phenomenon overlooked by the UCEs. Cross-exon UCSs, discussed in detailed below, represent a novel class of putative post-transcriptional regulators.

One might speculate that a cross-exon UCS may represent two ultraconserved but unrelated elements at neighboring splice junctions. While this is a possibility, it is unlikely based both on the statistical rarity of the event and on analysis of RNA-seq data. First, given the statistical rarity of the UCSs, it seems quite unlikely that the two halves of a UCS are two independent functioning elements. Given that only 96 UCSs occur among approximately 149 genes and 176 exons in the human genome, the chance of independent UCSs appearing in the same transcript by chance is <0.001. If the UCSs were not independent and functioned as, for example, a single element controlling splicing or RNA processing, they would effectively function as a single, albeit 'split', element and fit precisely our definition of a cross-exon UCS. It is important to note here that in order for a cross-exon UCS to function as one unit, the two halves do not necessary have to be together physically; they can function distally in orchestrating alternative splicing, for instance.

Conversely, if the function of cross-exon UCSs manifests at the mature mRNA level, i.e. when two halves are spliced together, we should observe enrichment of transcripts with two halves of the UCSs spliced together. This is indeed the case based on Burge Lab RNA-seq data (21), which provides transcriptomes of 14 tissues and cell lines. Out of 24 511 cassette exons, 25 contain cross-exon UCSs. Eighteen (72%) of these UCS cassette exons are found to be expressed in all 14 tissues and cell lines while only 8.8% of all cassette exons are expressed this ubiquitously (see Figure 2). The significant shift (one-tailed Wilcoxon test, $P = 1.0 \times 10^{-7}$) in the frequencies of expression patterns suggests that UCS cassette exons tend to be co-spliced with their neighboring constitutive exons and implicates the functions of cross-exon UCSs as one unit at the mature mRNA level.

Functional characterization

GO and InterPro enrichment tests suggest roles in RNA processing and RNA splicing. Enrichment test of Gene ontology (GO) and InterPro annotations reveals that genes-containing UCSs are enriched for many GO and InterPro terms related to post-transcriptional regulation, e.g. RNA binding (Fisher's exact test with Bonferroni correction, $P = 6.2 \times 10^{-13}$), RNA splicing ($P = 2.3 \times 10^{-12}$), mRNA processing ($P = 7.8 \times 10^{-12}$), mRNA metabolic process ($P = 8.5 \times 10^{-12}$), RNA processing ($P = 2.5 \times 10^{-9}$) and RNA recognition motif (RRM) ($P = 1.3 \times 10^{-8}$). This supports the findings of previous studies that transcribed UCEs, which share some characteristics of UCSs, are associated with mRNA processing (4,8,12). This also suggests that these RNA processing and RNA splicing factors are highly regulated at the RNA level.

It is important to note that these ontology term enrichments are indicative of entire proteins functioning as regulators and not necessarily the function of the UCS

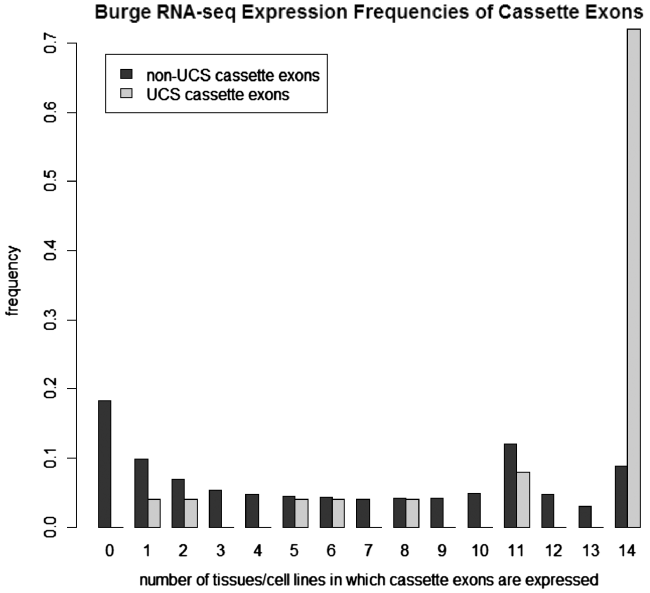

Figure 2. Expression patterns of cassette exons. Cassette exons containing cross-exon UCSs tend to be more ubiquitously expressed. This means UCS cassette exons are often spliced together with their neighboring UCS exons, forming one functional unit.

sites themselves. UCSs are conserved sites within coding regions and may act as recognition sites for regulatory factors such as the splisosome, ribosome or miRNAs; thus, UCSs could indicate tight regulation of UCS-containing transcripts. However, this does not imply that UCSs play a direct role in targeting or regulating RNA processing and splicing, but by providing sites to regulate the abundance of RNA processing/splicing proteins, the UCSs regulate RNA processing and splicing indirectly.

Strong association between cross-exon UCSs and alternative splicing events. Previous studies found some exonic UCEs to play roles in alternative splicing (11,12). Many exonic UCEs are observed to be in genes with clear evidence for alternative isoforms and many of them overlap cassette exons (4). However, these reports were primarily observational and no strong statistical association between exonic UCEs and alternative splicing was observed.

We assessed the statistical association between UCSs and alternative splicing events using the UCSC AltEvent database (22). This suggested different degrees of association between cross-exon and contiguous UCSs and alternative splicing events. Almost all (278 out of 282) of the UCSC genes containing UCSs have alternative spliced forms, while only 78.61% of the UCSC genes are alternatively spliced. The UCSC gene model was used because UCSC genes contain the greatest number of alternative spliced isoforms relative to other databases, 78.61%, which is closer to the reported number of 80% than RefSeq and ENSEMBLE. Ultraconserved transcripts (human EST that contain the UCSs) are also often alternatively spliced (hypergeometric, $P = 0.00068$). Since we expect that cross-exon UCSs are strongly associated with alternative splicing and the above association was observed over the whole collection of UCSs (both cross-exon and contiguous UCSs), we additionally investigated the two subtypes individually. As hypothesized, we found that the association between cross-exon UCSs and alternative splicing events are much stronger than that of contiguous UCSs. Out of 52 ESTs containing cross-exon UCSs, 33 are alternatively spliced. Contrasting with the background that only about 2.5 out of 8 million ESTs are alternatively spliced yields a P-value of 8.03×10^{-7}, while only 61 out of 133 ESTs containing contiguous UCSs are alternatively spliced ($P = 0.000304$).

This strong association between ultraconservation and splicing events can be partially explained by the presence of exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs). It is known that the presence of ESEs in coding sequence might impose purifying selection pressure on synonymous sites near splice sites (32). Since ESEs are predominantly located near splice junctions and the cross-exon UCSs are by definition located at splice junctions, the slow evolution rate in these UCSs can be attributed to splicing regulatory elements such as ESEs. This is also in agreement with the observation that alternative exons are highly conserved (33), reinforcing the evidence that the UCSs, particularly cross-exon UCSs, may be involved in homeostasis control of alternative splicing.

UCSs may function through coupling of alternative splicing and NMD. The mechanism through which ultraconservation regulates mRNA abundance is largely unknown. Nevertheless, one mechanism—coupling of alternative splicing and NSD (AS-NMD)—has been identified. AS-NMD involves the inclusion or exclusion of exon-containing in-frame premature stop codons by alternative splicing (Figure 3) (34,35). Previously, Ni *et al*. (12) performed a genome scale experimental screen to identify a class of 66 highly conserved stop codon-containing exons whose presence promotes transcript sensitivity to NMD. There are approximately 136 000 exons in the UCSC known genes and 96 UCSs.

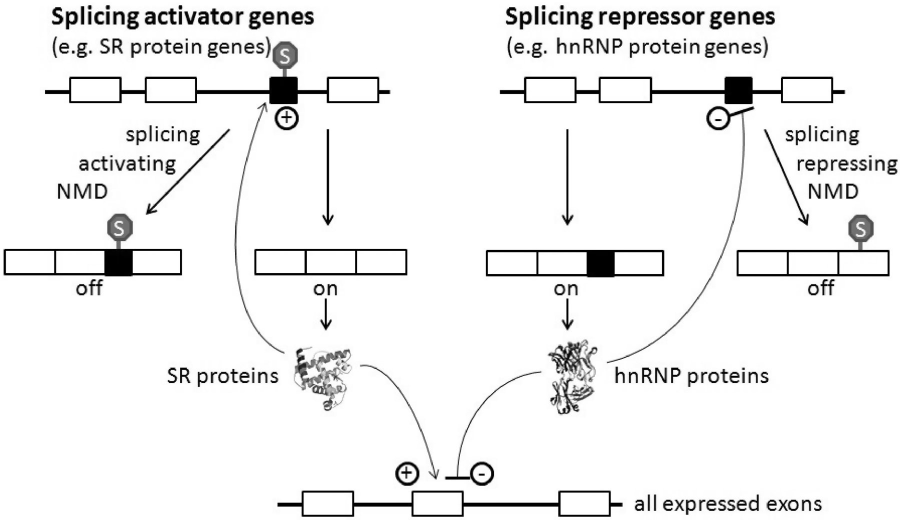

Figure 3. Mechanism through which stop codon-containing exons are alternatively spliced and degraded by means of NMD. This regulatory mechanism is prevalent in splicing regulators such as the SR and hnRNP protein genes. Figured inspired by (12).

Of the 66 conserved stop-codon containing exons, 7 overlap or are entirely contained within UCSs. This number suggests a highly significant association between UCSs and AS-NMD (Fisher's exact test, $P = 2.9 \times 10^{-14}$). This association is particularly strong in the cross-exon UCSs as all of the UCSs overlapping the stop codon-containing exons are cross-exon UCSs ($P = 1.1 \times 10^{-16}$). All of the seven exons are found in RNA splicing associated genes: RNPC2, TRA2A, SFRS3, SFRS6, SFRS7, SFRS10 and TIAL1. Combining the finding that UCSs are enriched in RNA splicing regulatory genes with the association between UCSs and AS-NMD, we conclude that UCSs are associated with homeostatic control of splicing regulators by AS-NMD.

Half of SR protein genes and many hnRNP protein genes may be regulated by UCSs through AS-NMD. SR and hnRNP protein genes are key splicing activators and repressors known to be regulated by AS-NMD. These two protein families act antagonistically by 5′-splice site selection (30) and there are nine genes in the SR family (36) and >20 genes in the hnRNP family (24) (see Supplementary Tables S2 and S3). Previously, highly conserved elements and UCEs within all members of the SR family and many of the hnRNP family were found to be alternatively spliced, either as alternative cassette exons containing early in-frame stop codons or as alternative introns in the 3′-UTR (11,12). Because the resulting alternatively spliced mRNAs are targeted for degradation by means of NMD (AS-NMD), it is believed that highly conserved and ultraconserved elements in the entire SR family play critical roles in translation regulation via Regulated Unproductive Splicing and Translation (RUST) (34).

Seven of the identified alternatively spliced, highly conserved and ultraconserved elements in the SR family are in fact included in the UCSs and five of them are cross-exon UCSs. We also observe alternative splicing events within all of the SR protein UCSs. Moreover, in SRP54, the UCE uc.28 is not located at an alternative splicing site while the UCS KDN02407 is (Figure 4). There are two different alternative splicing events observed around KDN02407: a retained intron and a bleeding exon. The retained intron is located in the middle of the two exons of KDN02407, while the bleeding exon (which also retains an intron) contains KDN02407 entirely. The close proximity of the two alternative splicing events suggests that UCS KDN02407, not UCE uc.28, may play a role in regulating alternative splicing of SRP54. Three of the five SR-protein UCSs (in SR proteins SRP20, SRP55 and 9G8) also contain Ni *et al.*'s stop codon-containing exons. This provides experimental evidence that these three UCSs indeed regulate homeostatic control of alternative splicing through regulation of key splicing regulators by means of AS-NMD.

Furthermore, eight members of hnRNP protein genes contain UCSs. It has been shown previously that some hnRNP genes are regulated by AS-NMD triggered through exon-skipping frame shift or 3′-UTR exon activation (12,37,38) (see Supplementary Table S3). In contrast with stop codon inclusion in SR family, NMD of the hnRNP family is triggered by exon skipping. The intricate antagonism between regulation of splicing activator SR proteins by stop codon exon inclusion and that of splicing repressor hnRNP proteins by exon skipping supports the model for auto- or cross-regulatory maintenance of splicing factor levels proposed previously (12,38) (Figure 3).

Unexpectedly high number of UCSs appear in 5′-UTRs and across start codons and contain upstream start codons (uAUGs). Of all 96 UCSs, 65 overlap coding sequences (CDS), 44 overlap 5′-untranslated regions (5′-UTRs) and 36 overlap 3′-UTRs. Enrichment tests under different null models suggest that it is very surprising to find as many as 44 UCSs in the relatively small 5′-UTR (binomial $P = 2 \times 10^{-7}$, hypergeometric $P = 2 \times 10^{-6}$). The UCSs are also enriched for overlapping with CDS (binomial $P = 0.005$, hypergeometric $P = 1 \times 10^{-8}$). Interestingly, there are only 23 UCSs that are completely contained within the CDS and UCSs are actually statistically depleted for being wholly contained within coding sequences (binomial $P = 2 \times 10^{-9}$). The contrasting enrichment for overlapping CDS and 5′-UTR and depletion for being completely contained within CDS implies that UCSs are often located at CDS–UTR junctions, especially between CDS and 5′-UTR.

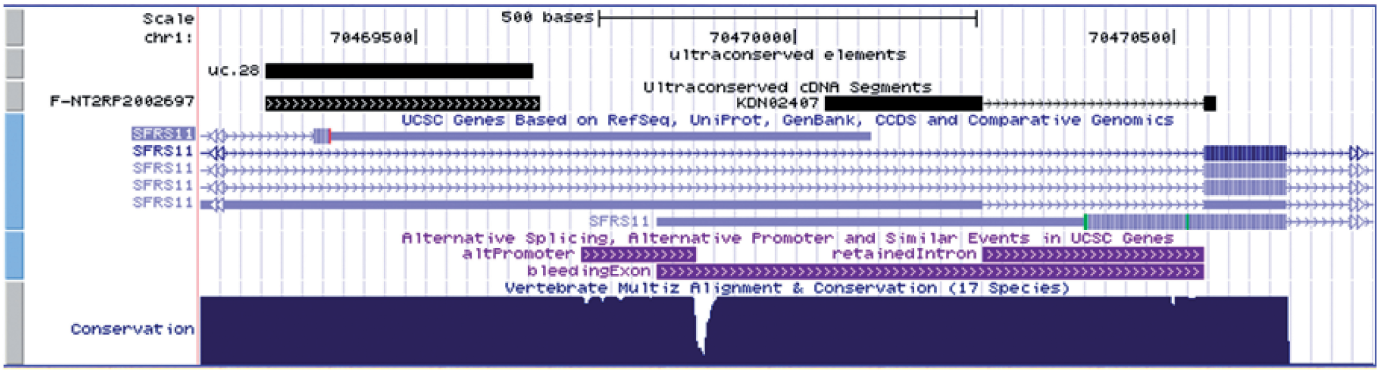
Scale
chr1:
500 bases
70469500
70470000
70470500
ultraconserved elements
uc.28
Ultraconserved cDNA Segments
F-NT2RP2002697
KDN02407
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
SFRS11
SFRS11
SFRS11
SFRS11
SFRS11
SFRS11
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
altPromoter
retainedIntron
bleedingExon
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation

Figure 4. SR protein gene SRP54 contains two UCSs, one of which is the same as UCE uc.28; the other is a cross-exon UCS KDN02407. Two alternative splicing events are associated with KDN02407, while none are found in uc.28. However, because of the upstream position of uc.28 from the alternative promoter, it may play role in regulating alternative transcription initiation instead.

One known regulatory mechanism in the 5′-UTR is translation regulation by upstream start codon (uAUGs) and uORFs. Presence of uAUGs and uORFs has been shown to affect translational efficiency of eukaryotic mRNAs, typically associated with translational repression in mammalian 5′-UTR (39,40). Also, mammalian uAUGs and uORFs have been demonstrated to be substantially conserved and thus functional (41,42). Because uAUG and uORF content affects translational efficiency, gene regulation can potentially be achieved through 5′-UTR diversity by altering the 5′-UTRs of mRNAs via events such as alternative splicing and alternative transcription initiation (42). Recent experimental studies showed that 5′UTR transcript diversity can itself be achieved during transcription (via alternative transcription initiation) and after transcription (via alternative splicing) (40,43–45).

A closer look at UCSs in 5′-UTRs revealed that an unexpectedly high number of UCSs contain start codons and uAUGs. We extracted start codon information from UCSC known genes (22) and found that as many as 24 UCSs contain start codons. Within these 24 start codon-containing UCSs, all but one have at least one uAUG. As uAUGs have been detected in only 29–48% of mammalian 5′-UTRs (26,46–48), the UCSs are strongly associated with uAUG. Moreover, 12 of these UCSs with uAUG have their 5′-UTRs alternatively spliced in such a way that the alternative isoforms contain different 5′-UTR sequences and uAUG content. The alternative splicing events include cassette exons, bleeding exons and retained introns. Each alters sequence composition and uAUG content through inclusion or exclusion of AUGs or start codons (Figure 5). Taken together, it seems that ultraconservation may also regulate gene expression via uAUGs, possibly by using alternative events to vary uAUG content and create 5′-UTR diversity.

Genes containing UCSs are also targets for miRNAs. As implied by enrichment of UCSs for post-transcriptional regulation, UCS genes may be under strong regulation. Because microRNAs (miRNAs) are a ubiquitous mechanism of post-transcriptional repression (49), we assessed the extent to which miRNAs regulate UCS genes by testing for enrichment of predicted miRNA target sites (50). Out of 186, 150 RefSeq genes-containing UCSs are targets of miRNAs, while only 36 864 out of 55 906 RefSeq genes in total are miRNA targets. This represents significant enrichment of miRNA targets relative to the genomic background (hypergeometric $P = 9 \times 10^{-10}$). However, the UCS loci themselves are not typically miRNA target sites (only 19 UCSs contain miRNA target sites), suggesting that UCS genes are generally subject to tight regulation by miRNA. In other words, UCSs in this case do not directly regulate the abundance of the mRNAs; rather, the existence of UCSs suggests important, possibly multiple,

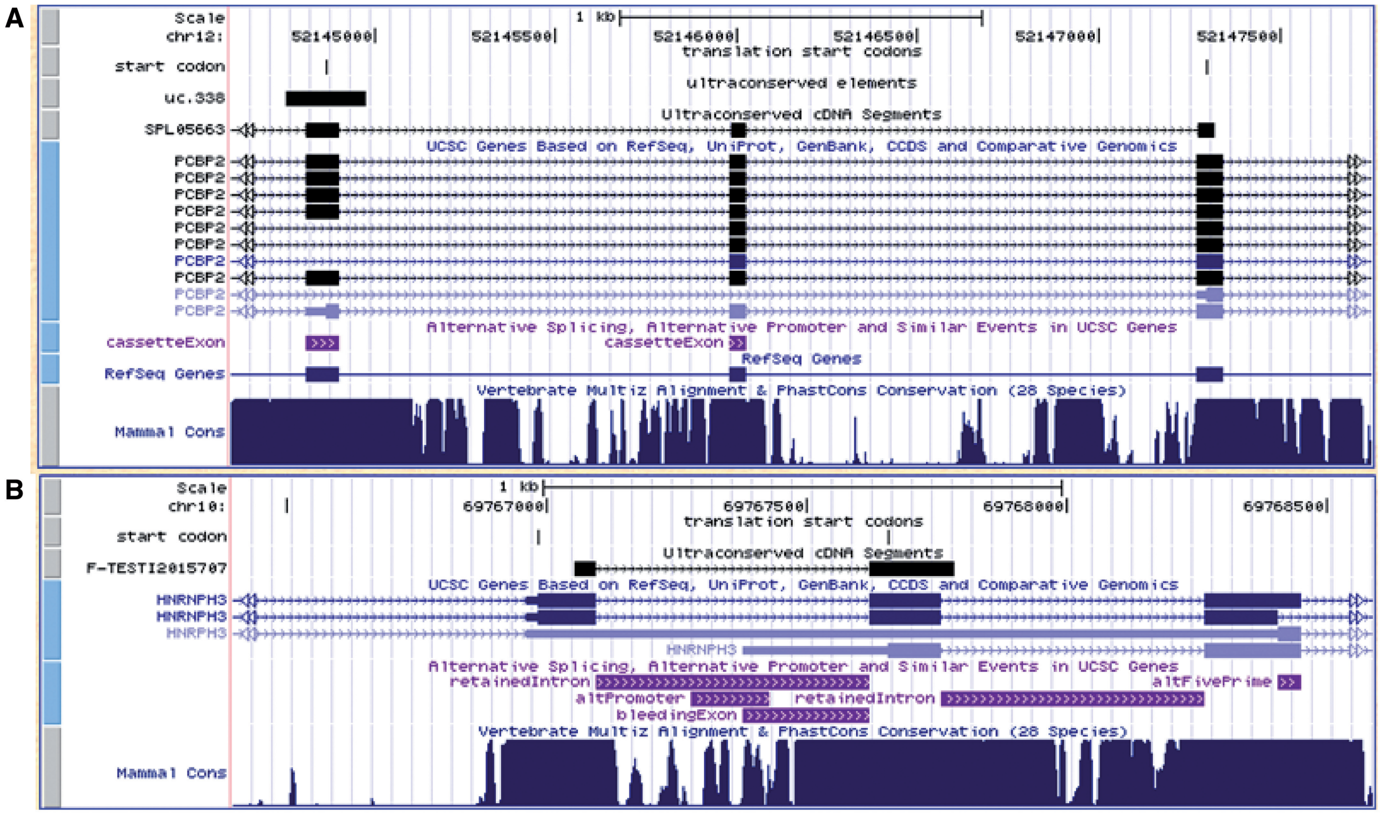
A
Scale
chr12:
1 kb
52145000|
52145500|
52146000|
52146500|
52147000|
52147500|
translation start codons
start codon
ultraconserved elements
uc.338
Ultraconserved cDNA Segments
SPL05663
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
PCBP2
PCBP2
PCBP2
PCBP2
PCBP2
PCBP2
PCBP2
PCBP2
PCBP2
PCBP2
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
cassetteExon
cassetteExon
RefSeq Genes
RefSeq Genes
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
B
Scale
chr10:
1 kb
69767000|
69767500|
69768000|
69768500|
translation start codons
start codon
Ultraconserved cDNA Segments
F-TESTI2015707
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
HNRNPH3
HNRNPH3
HNRPH3
HNRNPH3
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
retainedIntron
altFivePrime
altPromoter
retainedIntron
bleedingExon
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons

Figure 5. Examples of 5′-UTR diversity created through cassette exons, bleeding exons and retained introns. (**A**) 5′-UTR diversity (two light blue isoforms) is created through selective inclusion and exclusion of two ultraconserved cassette exons. The corresponding start codons are also contained within the UCS. (**B**) 5′-UTR diversity (two light blue isoforms) is created through retained introns and a bleeding exon. The start codon in one isoform becomes uAUG of the other. These exemplify how uAUG and alternative splicing might regulate translation efficiency.

biological functions of the containing genes requiring tight regulation.

UCSs resistance to folding reveals novel evolutionary constraint. Another potential explanation for conservation at the transcript level is structural conservation of base-pairing patterns and thermodynamic stability (51). In mammals, synonymous mutations are found to be under purifying selection so as to maintain mRNA secondary structure stability (52). Also, there is evidence that a UCE (uc.189) in arginine-/serine-rich splicing factor SFRS3 forms an important RNA secondary structure (4). We thus used the program Mfold (25) to assess the potential of UCSs to form an RNA secondary structure, comparing Gibbs free energy of the best folded structures to that of random cDNA sequences of the same length (Figure 6).

To our surprise, the predicted RNA secondary structures of the UCSs have significantly higher Gibbs free energy than those of randomly selected cDNA sequences of the same length (Wilcoxon test: $P = 0.005$), suggesting a resistance to the formation of RNA secondary structures. This observation is partially explained by the fact that UCSs are AT-rich (see Supplementary Materials). We hypothesize that this fold resistance is due to the need to keep UCS regions open for binding events to occur more deterministically. To support this, we predicted secondary structures of known miRNA binding sites and of miRNA coding sites from MiRBase (27). The hypothesis would imply fold resistance in miRNA binding sites but not in miRNA coding sites, which should exhibit stable structures. As expected, predicted folds of cDNA sequences containing miRNA binding sites also showed low stability (Wilcoxon test: $P = 0.0626$), while predicted folds of miRNA coding cDNA sequences had significantly higher stability than random (Wilcoxon test: $P < 2.2 \times 10^{-6}$).

This resistance to secondary structure formation is also observed in UCEs and in other highly conserved elements. Intriguingly, exonic UCEs exhibit even stronger resistance to folding (Wilcoxon $P = 1.67 \times 10^{-5}$) than UCSs and this was also the case for 17-species PhastCons conserved elements. We investigated four sets of PhastCons conserved elements with PhastCons scores $\geq$800 (most conserved), $\geq$650, $\geq$500 and $\geq$350 (least conserved). Only elements longer than 250 bp and overlapping CDS and UTR were considered. As shown in Figure 6, PhastCons conserved elements with scores $\geq$800 show the strongest fold resistance (Wilcoxon $P = 1.17 \times 10^{-9}$), stronger than UCSs and UCEs. The other three sets show weaker but significant fold resistance compared with random sequences (Wilcoxon $P = 0.019$ for score $\geq$650, $P = 0.003$ for score $\geq$500, $P = 0.045$ for score $\geq$350). Overall, this represents strong and consistent evidence for association between conservation and mRNA secondary structure instability.

In contrast to the previous observation (52), this suggests that mutations in coding sequences are under purifying selection incurring instability of secondary structures with more conserved elements possessing stronger fold resistance. As mentioned above, we believe that this resistance to folding is due to the need to keep mRNA open to allow regulatory factors to bind. Highly conserved genes are often highly regulated at transcription and therefore more likely to be bound by a larger number of factors (4,53). Extension of this observation to post-transcriptional regulation would explain the need for highly conserved mRNA to be more accessible.

UCSs are short lived. The evidence above suggests that UCS-containing genes are tightly regulated. To further support this, we examine mRNA degradation rates of the UCS genes, expecting an extreme rate in order to control transcript levels. Comparing the degradation rates and half-lives of UCS-containing mRNAs with those of 19 977 genes in mouse embryonic stem cells (54), we found that UCS-containing mRNAs tend to

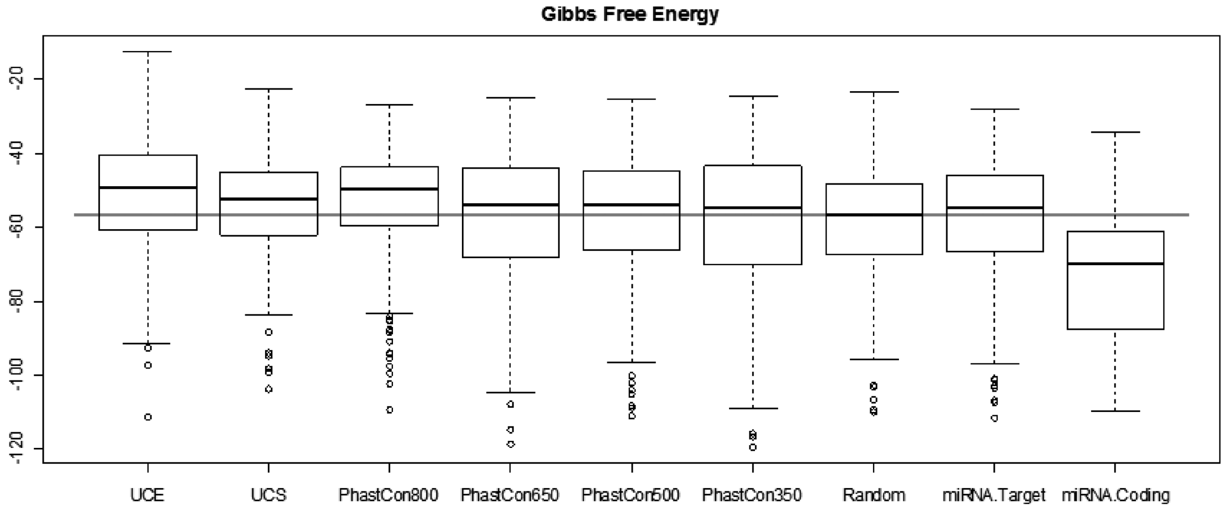

Figure 6. Gibbs free folding energy comparison shows significant fold resistance of ultra- and highly conserved elements as compared with random cDNA sequences. The gray horizontal line marks the mean of the fold energy of random sequences. miRNA coding sequences form significantly more stable structures while miRNA target sequences exhibit resistance to folding similar to conserved elements and suggesting an open structure for regulatory functions.

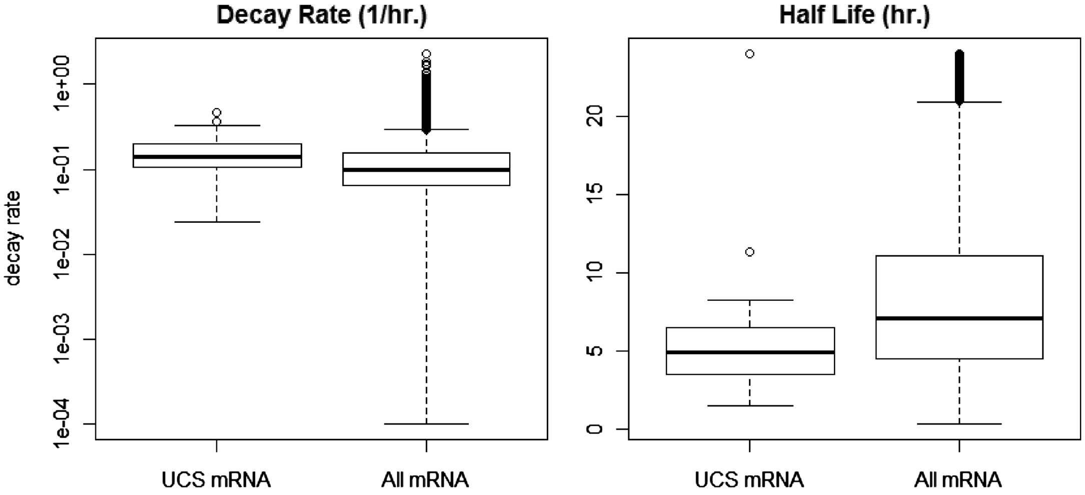

Figure 7. UCS-containing mRNAs degrade at faster rates and have shorter half-life than average mRNAs. Note that decay rate is presented in log scale.

have shorter half-life (Wilcoxon $P = 5.7 \times 10^{-8}$) and degrade at faster rates (Wilcoxon $P = 5.48 \times 10^{-8}$, Figure 7). This strengthens our hypothesis and suggests that the UCS genes are subject to strong repressive regulatory mechanisms.

Disease association. As expected, the UCSs exhibit very low variation in the human population. Only 12 out of 23 929 bases in the UCSs are at validated single nucleotide polymorphisms (SNPs) in the National Center for Biotechnology Information's SNP database (dbSNP built 129, June 2008). For an equivalent amount of DNA sequences, we would expect 53 validated sites by chance, so the validated SNPs are 4.4-fold underrepresented (hypergeometric test with Bonferroni correction: $P = 0.028$). When we include dbSNPs additional non-validated SNPs, the UCSs are even more highly depleted for SNPs ($P = 7.97 \times 10^{-23}$).

Five of the UCSs also overlap annotated regions in the Genetics Association Database (GAD) (55). One of them contains the start codon of NTRK3 gene, which is associated with panic disorder, and the other four are known miRNA target genes. One explicitly characterized UCS is found to be a miRNA target at the 3′-UTR of the thyroid hormone receptor alpha (THRA) gene, which is involved in thyroid cancer. In agreement with the mechanistic arguments made above, the gene THRA was also found to be regulated by AS-NMD (12), explaining the role of the UCS in this disease.

DISCUSSION

In conclusion, we have identified a novel class of putatively functional transcriptome elements, UCSs, based on a large experimental cDNA library and on their extreme conservation at the transcript level. UCSs differ from previously identified UCEs in that they can span multiple exons, a subtle distinction with implications in elucidating the role of ultraconservation in post-transcriptional regulation. Specifically, we find these UCS transcripts to be enriched for regulation at essentially every post-transcriptional stage: pre-mRNA splicing and degradation through AS-NMD, mature mRNA silencing by miRNA, fast mRNA decay rate and translational repression by uAUGs, all with the aid of fold resistance to keep the transcripts open and active. These multiple layers of regulation underscore the importance of these UCS genes as key global RNA processing regulators, for example in members of the SR protein family of essential splicing activators. Taken together, this evidence sheds new light on the multifaceted, fine-tuned and tight post-transcriptional regulation of gene families as conserved through the majority of the mammalian lineage.

A common theme of the mode of regulation of the UCS is downregulation; all of the regulatory mechanisms found to be associated with UCSs are repressive. One possible explanation of these multiple negative mRNA control mechanisms is to maintain a conserved defense against unwanted transcripts, since many parts of the genome may (56,57) or may not (58) be pervasively transcribed. Just as regulatory mechanisms of degrading unwanted peptides abound, the mRNA regulatory interactions characterized by UCS-containing families such as the SR and hnRNP proteins may be under conservation pressure to detect and control excess transcripts. This model of pervasive or baseline transcription coupled with degradation seems to fit the behavior of the UCS genes well.

Strictly defined UCSs may represent only a small subset of transcripts extremely conserved for functional reasons. A recent study indicated that UCEs themselves represent a subset of constrained developmental transcriptional enhancers (59) and as such, UCSs may similarly be a subset of a larger class of highly conserved post-transcriptional regulatory elements. Moreover, by retaining the published definition of ultraconservation, this study began with a small set of UCSs and might have missed additional important observations. A preliminary relaxation of the 200-bp length requirement found 301 UCSs of length 150 bp, 1132 UCSs of length 100 bp and 9767 UCSs of length 50 bp (see Supplementary Materials for lists of these shorter UCSs). Genes containing these shorter UCSs are enriched for similar GO and InterPro terms (e.g. RNA binding, RNA splicing, mRNA metabolic process, RNA processing, RRM) with even stronger *P*-values. This demonstrates that the arbitrary cutoff of 200 bp may not be necessary for ultraconservation and this is likely to be true of the 100% sequence identity requirement as well.

Thus, an interesting extension of this work will be to further relax the requirement for full ultraconservation and to assess the generalizability of the discovered properties of UCSs in this larger class of conserved elements. Other candidate methods include gradual lowering of PhastCons conservation score cutoff (60), Tseng and Tompa's algorithms for locating extremely conserved elements in multiple sequence alignments (61) and Gumby, a statistical approach with scoring parameters optimized through multiple genome-wide scans (62), see also (59). With various possible methods to discover UCSs, the expression 'ultraconserved' is perhaps best used as a descriptive term referring to a form of extreme conservation. Exploration of ultraconservation in other organisms and at different levels of conservation should also be encouraged and pursued.

Regardless of the definition of ultraconservation, the list of 96 UCSs presented here is by no mean exhaustive, as the cDNA library used is not yet complete. With the advancement of sequencing technologies, particularly RNA-seq, more whole transcriptome information will become available and the list of UCSs will grow accordingly. More UCSs may be found in transcripts expressed only in specific tissues or cell states and further categorization of UCSs by expression pattern may be possible.

Because UCSs are defined specifically in the context of mature mRNA transcripts, they would be more informative for studying post-transcriptional regulation than the traditionally used UCEs. As discussed above, one of the most deeply conserved UCSs, F-TESTI2015707, was not included in UCEs but appeared in a key splicing repressor HNRNPH3 and harbored multiple alternative splicing sites. Many previous studies used genome-level conservation (e.g. UCEs) as a guide to identify or understand post-transcriptional regulatory elements. This approach fails to find some important elements because of intervening introns, which with exception of pre-mRNA modification such as splicing, are irrelevant to post-transcriptional regulation. Here we promote the use of transcript-level conservation which allows discovery of elements that are functional only at mature mRNA level and could yield new insights into functions specific to post-transcriptional regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank Kenta Nakai, Riu Yamashita, Kohji Okamura, Yoshiaki Tanaka, Alexis Vandenbon and members of Nakai and Sugano Laboratories for discussion and guidance, Leah Segal for critically reading our manuscript and Etsuko Sekimori for technical support.

FUNDING

This project was made possible by the support of Stanford Center for Technology and Innovation. Funding for open access charge: Grants-in-Aid for Scientific Research, KAKENHI, Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Boffelli,D., Nobrega,M.A. and Rubin,E.M. (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.*, **5**, 456–465.
2. Dermitzakis,E.T., Reymond,A. and Antonarakis,S.E. (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, **6**, 151–157.
3. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
4. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
5. Sakuraba,Y., Kimura,T., Masuya,H., Noguchi,H., Sezutsu,H., Takahasi,K.R., Toyoda,A., Fukumura,R., Murata,T., Sakaki,Y. *et al.* (2008) Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mamm. Genome*, **19**, 703–712.
6. Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
7. Duret,L., Dorkeld,F. and Gautier,C. (1993) Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.*, **21**, 2315–2322.
8. Calin,G.A., Liu,C.G., Ferracin,M., Hyslop,T., Spizzo,R., Sevignani,C., Fabbri,M., Cimmino,A., Lee,E.J., Wojcik,S.E. *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, **12**, 215–229.
9. Paparidis,Z., Abbasi,A.A., Malik,S., Goode,D.K., Callaway,H., Elgar,G., deGraaff,E., Lopez-Rios,J., Zeller,R. and Grzeschik,K.H. (2007) Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. *Dev. Growth Differ.*, **49**, 543–553.
10. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
11. Lareau,L.F., Inada,M., Green,R.E., Wengrod,J.C. and Brenner,S.E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926–929.
12. Ni,J.Z., Grate,L., Donohue,J.P., Preston,C., Nobida,N., O'Brien,G., Shiue,L., Clark,T.A., Blume,J.E. and Ares,M. Jr (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.*, **21**, 708–718.
13. Feng,J., Bi,C., Clark,B.S., Mady,R., Shah,P. and Kohtz,J.D. (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.*, **20**, 1470–1484.

14. Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
15. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.*, **36**, 40–45.
16. Suzuki,Y. and Sugano,S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.*, **221**, 73–91.
17. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
18. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
19. Wakaguri,H., Yamashita,R., Suzuki,Y., Sugano,S. and Nakai,K. (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res.*, **36**, D97–D101.
20. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
21. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
22. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
23. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
24. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.
25. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
26. Suzuki,Y., Ishihara,D., Sasaki,M., Nakagawa,H., Hata,H., Tsunoda,T., Watanabe,M., Komatsu,T., Ota,T., Isogai,T. *et al.* (2000) Statistical analysis of the 5′ untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. *Genomics*, **64**, 286–297.
27. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
28. Smith,C.W. and Valcarcel,J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
29. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
30. Caceres,J.F. and Kornblihtt,A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.*, **18**, 186–193.
31. Martinez-Contreras,R., Cloutier,P., Shkreta,L., Fisette,J.F., Revil,T. and Chabot,B. (2007) hnRNP proteins and splicing control. *Adv. Exp. Med. Biol.*, **623**, 123–147.
32. Parmley,J.L., Chamary,J.V. and Hurst,L.D. (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.*, **23**, 301–309.
33. Kim,E., Goren,A. and Ast,G. (2008) Alternative splicing: current perspectives. *Bioessays*, **30**, 38–47.
34. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
35. Baek,D. and Green,P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl Acad. Sci. USA*, **102**, 12813–12818.
36. Shepard,P.J. and Hertel,K.J. (2009) The SR protein family. *Genome Biol.*, **10**, 242.
37. Rahman,L., Bliskovski,V., Kaye,F.J. and Zajac-Kaye,M. (2004) Evolutionary conservation of a 2-kb intronic sequence flanking a tissue-specific alternative exon in the PTBP2 gene. *Genomics*, **83**, 76–84.
38. Wollerton,M.C., Gooding,C., Wagner,E.J., Garcia-Blanco,M.A. and Smith,C.W. (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell*, **13**, 91–100.
39. Ji,H., Zhang,Y., Zheng,W., Wu,Z., Lee,S. and Sandberg,K. (2004) Translational regulation of angiotensin type 1a receptor expression and signaling by upstream AUGs in the 5′ leader sequence. *J. Biol. Chem.*, **279**, 45322–45328.
40. Song,K.Y., Hwang,C.K., Kim,C.S., Choi,H.S., Law,P.Y., Wei,L.N. and Loh,H.H. (2007) Translational repression of mouse mu opioid receptor expression via leaky scanning. *Nucleic Acids Res.*, **35**, 1501–1513.
41. Churbanov,A., Rogozin,I.B., Babenko,V.N., Ali,H. and Koonin,E.V. (2005) Evolutionary conservation suggests a regulatory function of AUG triplets in 5′-UTRs of eukaryotic genes. *Nucleic Acids Res.*, **33**, 5512–5520.
42. Resch,A.M., Ogurtsov,A.Y., Rogozin,I.B., Shabalina,S.A. and Koonin,E.V. (2009) Evolution of alternative and constitutive regions of mammalian 5′UTRs. *BMC Genomics*, **10**, 162.
43. Hughes,T.A. and Brady,H.J. (2005) Expression of axin2 is regulated by the alternative 5′-untranslated regions of its mRNA. *J. Biol. Chem.*, **280**, 8581–8588.
44. Newton,D.C., Bevan,S.C., Choi,S., Robb,G.B., Millar,A., Wang,Y. and Marsden,P.A. (2003) Translational regulation of human neuronal nitric-oxide synthase by an alternatively spliced 5′-untranslated region leader exon. *J. Biol. Chem.*, **278**, 636–644.
45. Pan,Y.X. (2005) Diversity and complexity of the mu opioid receptor gene: alternative pre-mRNA splicing and promoters. *DNA Cell Biol.*, **24**, 736–750.
46. Davuluri,R.V., Suzuki,Y., Sugano,S. and Zhang,M.Q. (2000) CART classification of human 5′ UTR sequences. *Genome Res.*, **10**, 1807–1816.
47. Iacono,M., Mignone,F. and Pesole,G. (2005) uAUG and uORFs in human and rodent 5′untranslated mRNAs. *Gene*, **349**, 97–105.
48. Pesole,G., Gissi,C., Grillo,G., Licciulli,F., Liuni,S. and Saccone,C. (2000) Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene*, **261**, 85–91.
49. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
50. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
51. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
52. Chamary,J.V. and Hurst,L.D. (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.*, **6**, R75.
53. Katzman,S., Kern,A.D., Bejerano,G., Fewell,G., Fulton,L., Wilson,R.K., Salama,S.R. and Haussler,D. (2007) Human genome ultraconserved elements are ultraselected. *Science*, **317**, 915.
54. Sharova,L.V., Sharov,A.A., Nedorezov,T., Piao,Y., Shaik,N. and Ko,M.S. (2009) Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res.*, **16**, 45–58.
55. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
56. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
57. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.

58. van Bakel,H., Nislow,C., Blencowe,B.J. and Hughes,T.R. (2010) Most "dark matter" transcripts are associated with known genes. *PLoS Biol*, **8**, e1000371.

59. Visel,A., Prabhakar,S., Akiyama,J.A., Shoukry,M., Lewis,K.D., Holt,A., Plajzer-Frick,I., Afzal,V., Rubin,E.M. and Pennacchio,L.A. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.*, **40**, 158–160.

60. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

61. Tseng,H.H. and Tompa,M. (2009) Algorithms for locating extremely conserved elements in multiple sequence alignments. *BMC Bioinformatics*, **10**, 432.

62. Prabhakar,S., Poulin,F., Shoukry,M., Afzal,V., Rubin,E.M., Couronne,O. and Pennacchio,L.A. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.*, **16**, 855–863.