PROTIST GENOMICS
Research Article • DOI: 10.2478/prge-2013-0001 • PRGE • 2013 • 19–24

VERSITA

Patterns of intron sequence conservation in the genus *Tetrahymena*

Yichen Zheng*,
Kristen L. Dimond,
Dan Graur,
Rebecca A. Zufall

Department of Biology and Biochemistry,
University of Houston, Houston, TX,
77204, USA

Abstract
Background:
Introns constitute a large fraction of eukaryotic genomes and were once considered neutrally evolving sequences. Recently, however, some introns have been found to harbor sequences that are involved in a variety of regulatory and other functions and show evidence of purifying selection.
Results:
We examine the pattern of sequence divergence among ciliates in the genus *Tetrahymena*. We find that on average introns are more highly conserved than four-fold degenerate sites. Among introns, we find a correlation between conservation strength and both position rank in the gene as well as size of the coding region; the most conserved introns are found closer to the 5' end of the largest genes.
Conclusion:
Our results indicate that *Tetrahymena* introns experience selective constraint, possibly due to harboring regulatory sequences. We advocate for further experimental study of possible intron functions in *Tetrahymena*.

Keywords
Intron • Evolution • Ciliate • Tetrahymena

Received 08 August 2012
Accepted 14 January 2013

Introduction

Introns, once considered exclusively junk DNA, are now known to contain various regulatory elements involved in a variety of aspects of mRNA processing [reviewed in 1]. In addition, introns may function to enhance meiotic crossing over, as signals for mRNA export from the nucleus and nonsense-mediated decay, and as sources of alternative splicing [reviewed in 2]. Thus, in contrast to the previous expectation that all introns should evolve neutrally, several studies have found evidence of selective constraint on intronic sequences [e.g., 3-6].

Previous studies of sequence conservation in introns have demonstrated that various aspects of intron structure and composition may affect the rate at which introns evolve. For example, in *Drosophila*, long introns evolve more slowly than shorter ones and introns positioned first within a gene tend to be longer than non-first introns [7,8]. In primates, long and first introns evolve more rapidly than short or non-first introns. However, short first introns have a higher GC content and are more conserved compared to long first introns [9]. In order to gain a fuller understanding of the evolutionary constraints on intronic sequences, it will be instructive to examine the factors that influence sequence conservation of introns in diverse lineages.

For a variety of reasons, ciliates in the genus *Tetrahymena* are good candidates to study the evolutionary pattern in introns. *T. thermophila* is a well-established model system in cellular and molecular biology [10] whose macronuclear genome has been sequenced and its genes and coding sequences (CDSs) annotated [11], making intron identification straightforward. The genomes of two additional species of *Tetrahyemena*, *T. malaccensis* and *T. elliotti*, have also been recently sequenced (*Tetrahymena* Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org)) allowing across-species comparisons.

Introns in *Tetrahymena* are, on average, small relative to those in animals, larger than those in *Paramecium* [12], and comparable in size to those in *Arabidopsis* [11,13]. Like other non-protein-coding sequences in the *Tetrahymena* genomes, introns are very AT-rich (16.3% GC) [11]. Notwithstanding the very famous group-I self-splicing intron described in *T. thermophila* [14], almost all nuclear introns in *Tetrahymena* are spliceosomal [11]. Here, we compiled homologous introns from the *T. thermophila, T. malaccensis*, and *T. elliotti* genomes and analyzed their rate of evolution with respect to various factors of location, composition, and size.

Data and methods

Data

The sequences of all *T. thermophila* genes (tta1_oct2008_finalrelease.gene.fsa) and CDSs (tta1_oct2008_finalrelease.cds.

* *E-mail: yzheng7@uh.edu*

fsa) were downloaded from http://ciliate.org/index.php/home/downloads [15]. The unannotated genomes of *T. malaccensis* and *T. elliotti* were downloaded as supercontigs from the Broad Institute (http://www.broadinstitute.org/annotation/genome/Tetrahymena/MultiDownloads.html).

The *T. malaccensis* and *T. elliotti* supercontigs were independently compared with *T. thermophila* genes using the default settings of BLASTN, except "-num_alignments 4000" to ensure capturing all results. Hits from different parts of the same gene were merged, and hits that covered less than 25% of a gene were removed. To prevent outparalogs from muddling the dataset and to ensure data quality, only genes with exactly one hit in the *T. malaccensis* genome and which included more than 70% of the full-length gene were used in later studies. Sequences with one or more unknown nucleotides ("Ns") were also removed.

Each *T. thermophila* gene was aligned to its *T. malaccensis* and *T. elliotti* orthologs by using MUSCLE [16]. To determine intron positions, gene sequences were aligned with *T. thermophila* CDSs. The alignment between gene and CDS was assessed for splice site "sliding" due to alignment error, and was corrected based on the observation that all but a few *T. thermophila* introns start with "GT" and end with "AG" [11]; we assumed this was also true for the other species.

To increase the confidence in the correct identification of introns, we further refined the data set. All first introns in genes whose 5' exon is less than 5 bp were removed, because the annotation of such small first exons is often not reliable. In order to account for the fact that introns in *T. malaccensis* and *T. elliotti* have not been annotated, introns with a neighboring exon whose interspecific alignment has less than 80% identity are removed. In order to account for exons that are present in one species but not the other, introns with neighboring exons with greater than 90% gaps were also removed. The remaining introns are referred to as "eligible introns." After these procedures, 55,162 introns from 11,594 genes were retained for the *T. thermophila-T. malaccensis* pair, and 22,583 introns from 5,355 genes were retained for *T. thermophila-T. elliotti* pair. There were 22,110 introns in 5,261 genes shared by all three species. We additionally created a second sub-set of introns including only those introns that are supported by EST evidence [17]. This resulted in 13,527 introns from 2840 genes for *T. thermophila-T. malaccensis* pair and 7070 introns in 1681 genes for *T. thermophila-T. elliotti*. The first and last five nucleotides in each intron were trimmed, in order to remove positions that are likely to be conserved due to splicing constraints.

Each intron was categorized based on four criteria: intron size, size of gene, or its coding sequence, in which the intron resides, GC content, and positional rank in the intron, defined as first, middle, last, or only.

Sequence divergence

The degree of conservation in introns was measured by the Jukes-Cantor (JC) distance between orthologs from *T. thermophila* and *T. malaccensis*. To account for the effects of rate heterogeneity across the genome, the distance at four-fold degenerate sites was also calculated. Distance ratios were calculated by dividing the JC intronic distances by the distances at four-fold degenerate sites. Distance ratios were log transformed in order to make the values comparable regardless of whether introns or four-fold degenerate sites have a larger JC distance. These Log Distance Ratios (LDR) were used to compare rates of evolution among introns.

Patterns of nucleotide substitution

Nucleotide substitution frequencies for the 12 different classes of possible mutations (A to C, A to G, etc.) were determined using the three species data set. The frequency of mutations was determined for *T. thermophila* and *T. malaccensis*, using *T. elliotti* as an outgroup to polarize the changes.

Results

Tetrahymena *introns*

Introns in *T. thermophila* and *T. malaccensis* range in length from 21-7880 nt, with a mean length of 132 nt (standard deviation, 157). The average number of introns per gene is 4.8 (4.3). The average GC content of introns is 15% (4.6%); the GC content of four-fold degenerate sites is 18% (5.5%) (using only genes with EST confirmed introns does not change these values).

Intron gain and loss

Because the *T. thermophila* genome is fairly well annotated, we are able to estimate the number of intron gains in this lineage (1) and the number of losses in the *T. malaccensis* lineage (4), however this is only in genes whose *T. elliotti* orthologs can be retrieved. We are not confident in assessing the alternative (i.e. losses in *T. thermophila* and gains in *T. malaccensis*) due to lower sequence and annotation quality of the *T. malaccensis* genome. Nonetheless, there appear to be very few intron gains and losses along these lineages.

Correlations among intron types

We found that first and only introns were both longer and more GC-rich than middle and last introns (Table 1). Only introns are more frequently found in the 5' end of the gene (Kolmogorov-Smirnov test, $p<0.05$), which may help explain their similarity to first introns. Larger introns were more GC-rich than smaller introns (correlation using all eligible introns, $R^2=0.039$, $p<0.0001$; using only EST confirmed introns, $R^2=0.065$, $p<0.0001$). Genes with larger CDS also tended to have more introns ($R^2=0.78$, EST confirmed $R^2=0.59$ $p<0.0001$), which were less GC-rich ($R^2=0.010$, $p<0.0001$; EST confirmed, $R^2=0.0062$, $p<0.0001$).

Intron sequence conservation

The total, concatenated, Jukes-Cantor distance between *T. thermophila* and *T. malaccensis* introns is 0.25 (0.24 for EST confirmed introns) substitutions per site. The distance for four-fold degenerate sites is 0.38 (0.35) substitutions per site. Thus,

Table 1. Mean length, GC content, and LDR of introns to four-fold degenerate sites categorized by intron positional rank.

Positional rank	Number of introns	Mean length** (± standard deviation)	Mean GC content** (± standard deviation)	LDR** (± standard deviation)
First	8,416, 2122*	157[a] (165) 165[a] (165)	0.17[a] (0.043) 0.172[a] (0.040)	-0.41[a] (0.61) -0.26[a] (0.61)
Middle	36,337, 8699	126[b] (154) 131[b] (164)	0.15[b] (0.044) 0.146[b] (0.041)	-0.33[b] (0.51) -0.22[b] (0.52)
Last	8,674, 2213*	122[b] (142) 126[b] (151)	0.14[c] (0.049) 0.136[c] (0.044)	-0.33[b] (0.59) -0.14[c] (0.59)
Only	1,735, 451	201[c] (219) 248[c] (241)	0.17[d] (0.052) 0.166[d] (0.046)	-0.39[a] (0.62) -0.23[ab] (0.59)

Numbers in black include all introns, numbers in grey include only EST confirmed introns.
**The inequality between the number of first and last introns is due to some first introns being removed if they flank a very short first exon.*
***Entries within a column with different letters are significantly different from one another (t-test with Bonferroni correction).*

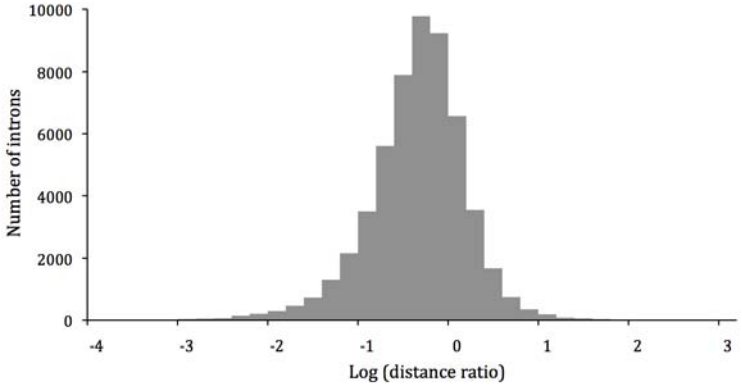

Figure 1. Distribution of the ratio of intron distances to four-fold degenerate site distances between T. thermophila and T. malaccensis. Log (distance ratio) (LDR) less than zero indicates introns that are more conserved than four-fold degenerate sites.

the log distance ratio (LDR) across the genome was -0.41 (-0.38), indicating that introns, on average, are more conserved than four-fold degenerate sites. When intron distances are compared only with four-fold degenerate sites from the same gene, the mean LDR is -0.34 (-0.22) (standard deviation, 0.54 (0.55)), which is significantly smaller than zero (z-test, $p < 0.0001$). This indicates that introns are more conserved than four-fold degenerate sites in the gene in which they are found. As seen in Figure 1, however, there is a huge variation in the degree of conservation in introns relative to third codon positions.

When we look at the different categories of introns, we find two patterns that help explain this variation: (1) First and only introns are more conserved than introns in other positions (Table 1). Likewise, the closer an intron is to the 5' end of the gene or the start codon, the more highly conserved it is (Figure 2A). And (2) introns in genes with larger CDS are more conserved than those in shorter genes (Figure 2B). Thus, the extent of sequence conservation in introns in *Tetrahyemena* is positively correlated with CDS size and proximity to the 5' end of the gene. However, these factors explain only a small fraction of the variance in the distance ratios between introns (bivariate regression, R^2=0.021).

Patterns of nucleotide substitution

The pattern of nucleotide substitutions in introns and four-fold degenerate sites of codons were determined using sequences of *T. elliotti* as an outgroup to polarize the differences between *T. thermophila* and *T. malaccensis*. We find significant differences in substitution patterns between species in both introns and four-fold degenerate sites (χ^2-test, $p < 0.0001$). However, the difference in substitution pattern between species is most pronounced in introns (Figure 3).

Based on these substitution patterns, the expected GC frequencies at equilibrium are 0.34 and 0.36 for introns in *T. thermophila* and *T. malaccensis*, respectively, and 0.35 and 0.42 for four-fold degenerate sites. These values are much larger than the current GC content (Table 1) indicating that CG content is not at equilibrium.

Discussion

In *Tetrahymena*, we find few instances of intron loss or gain, and also find that introns are more conserved than four-fold degenerate sites, suggesting the action of purifying selection on at least some intronic sequences. The *T. thermophila*

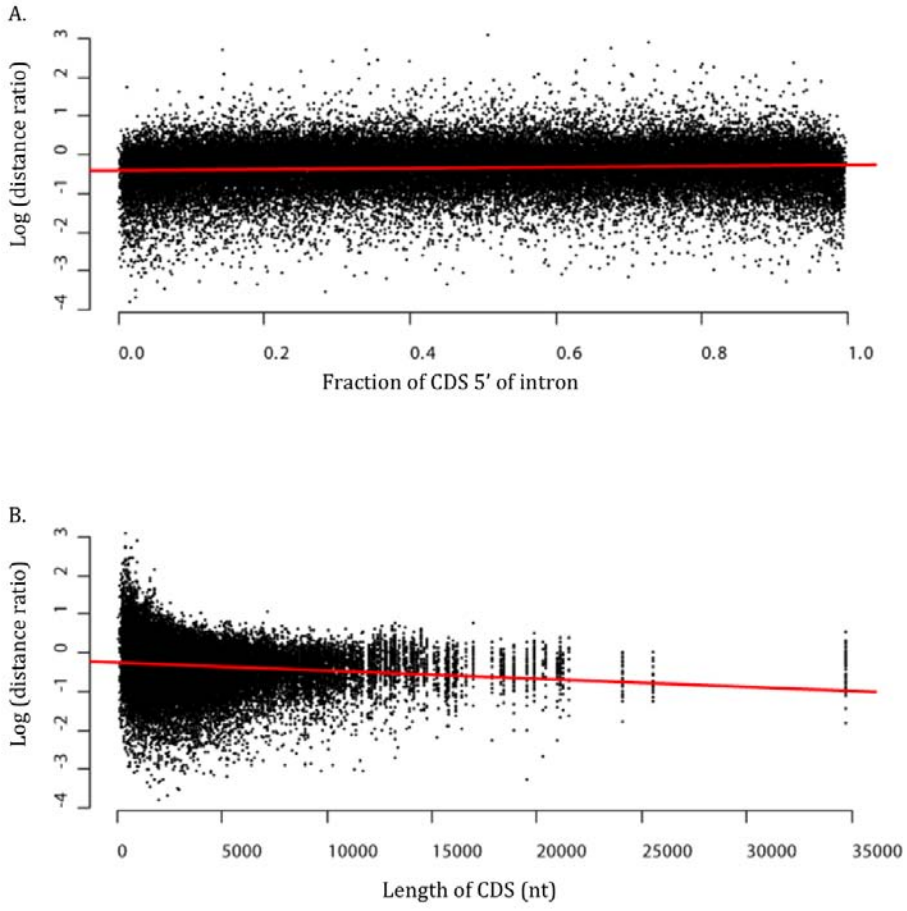

Figure 2. *The effect of (A) position of intron within a gene and (B) CDS size on intron sequence conservation. (A) Relationship between the fraction of CDS that is 5′ of an intron location and LDR (regression, R_2=0.0045 (using only EST confirmed introns R_2=0.0053) p<0.0001; using full gene instead of CDS, R_2=0.0055 (0.0064), p<0.0001). (B) Relationship between size of CDS of the gene in which an intron is found and LDR (regression, R_2=0.0165 (0.049), p<0.0001; using full gene size instead of CDS, R_2=0.0130 (0.041), p<0.0001).*

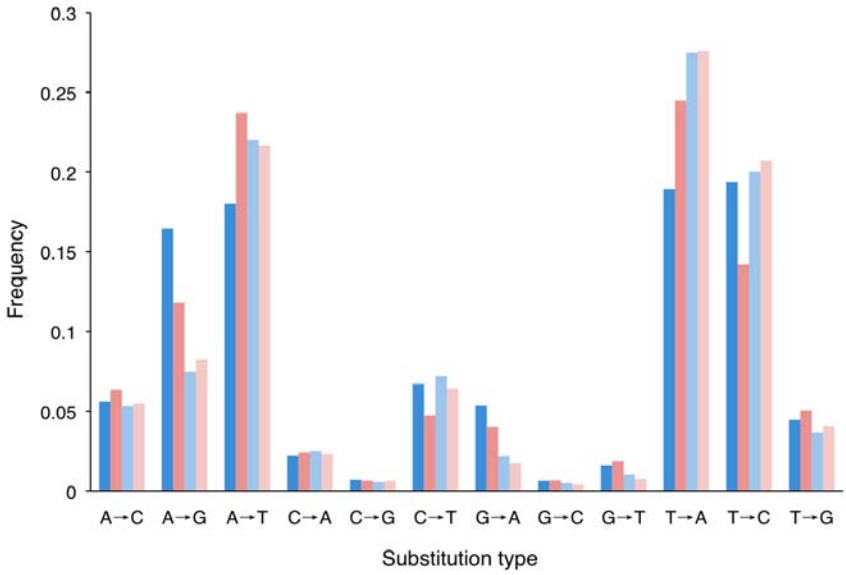

Figure 3. *The frequency of each of the 12 possible substitutions in T. thermophila (blue bars) and T. malaccensis (red) introns (dark) and third codon positions (light).*

genome has been previously identified as experiencing codon bias [11,18,19]; our results indicate that introns in *Tetrahymena* are experiencing even stronger selective constraint than that imposed by selection for preferred codons. However, it is currently unclear what is driving this constraint.

One possibility is that intron conservation in *Tetrahymena* is due to selection on regulatory sequences found in intronic sequences. Experimental data has implicated intronic sequences in transcriptional regulation in a variety of genes and organisms [20 and references therein]. High sequence conservation in introns is thus often used to suggest possible regulatory functions contained in introns [e.g., 1,4,6,21]. In *Tetrahymena*, first and only introns are the most highly conserved (Table 1), and this sequence conservation is highest in introns nearest the 5' end or start codon of genes (Figure 2). This pattern is consistent with similar patterns found in other species (e.g. rodents [4] and humans [22]) and may suggest the presence of regulatory sequences concentrated in first introns. Data from additional species support a higher abundance of regulatory elements in first introns. For example, Marais et al. [7] explain the relationship between first intron size and gene expression level in *Drosophila* by postulating a higher abundance of regulatory elements in first introns. And experimental studies have demonstrated the presence of regulatory elements in first introns of some genes [e.g., 23-28]. In *Tetrahymena*, first introns also have the highest GC content. Previous studies have suggested that high GC content, also found in first introns in humans, is due to the importance of CpG dinucleotides in transcriptional regulation [22]. However, note that the GC content of *Tetrahymena* introns is very low, substantially lower than that of humans, thus is unlikely to be as informative as in humans. The hypothesis that sequence evolution is constrained in *Tetrahymena* introns due to selection against mutations in regulatory elements, and insight into the precise nature of any such elements, await experimental testing.

Another factor that can result in elevated levels of sequence conservation is alternative splicing. 5.2% of genes in *T. thermophila* have been estimated to undergo alternative splicing [29]. Alternatively spliced introns in our data set will inflate the degree of intron conservation relative to four-fold degenerate sites, since these introns would be expected to be similar in conservation to exons [30]. Until data on alternatively spliced introns in *Tetrahymena* become available, the exact effect they have on our data set remain unclear. We suspect that some of these alternatively spliced introns are responsible for the elevated number of very short (<45 nt) 3n stopless introns seen in Figure 4a. Out of these 1,082 introns, 102 are completely identical between *T. thermophila* and *T. malaccensis*, and the remaining have a LDR of -1.03. This is significantly more conserved than both very short stopless introns of lengths that are not multiples of 3 (LDR = -0.67, $p < 0.0001$) and the very short stop-containing introns (LDR = -0.67, $p < 0.0001$). Thus, it is possible that some of these very short stopless codons of length 3n are alternatively spliced. If we remove these introns from our total data set of eligible introns, the mean LDR across all introns is slightly lower (-0.33 or -0.21 for EST confirmed introns) than when they are included (-0.34 or -0.22 for EST confirmed).

A previous study in another ciliate, *Paramecium tetraurelia*, found evidence of constraint on intron sequences due to selection favoring the maintenance of in-frame stop codons to promote nonsense-mediated decay in the event of improper splicing [12]. Introns in *P. tetraurelia* are over 5 times smaller than those in *T. thermophila*, with greater than 96% of introns smaller than 34 nucleotides [12]. In *T. thermophila* there are only 1364 (or 2.47% of) introns smaller than 34 nucleotides. Nonetheless, similar to the smallest introns in *Arabidopsis* and human [12], *T. thermophila* is deficient in small introns without stop codons in lengths that are multiples of three (Figure 4). This supports a role of selection in maintaining in-frame stop codons in these introns.

Our results demonstrate that *Tetrahymena* introns are evolving more slowly than expected under neutrality. However, it is currently unknown what selective force(s) result in this pattern. Thus, these results advocate for study of any regulatory or other function contained in *Tetrahymena* intron sequences, and for further study of patterns of evolution in intron sequences in more diverse eukaryotic lineages.

Ackowledgements

We are grateful to H. Long for his help in preparing the manuscript and to two anonymous reviewers for helpful advice and insight. This work was supported by the Texas Higher Education Coordinating Board (Norman Hackerman Advanced Research Program, 003652-0102-2009; RAZ).

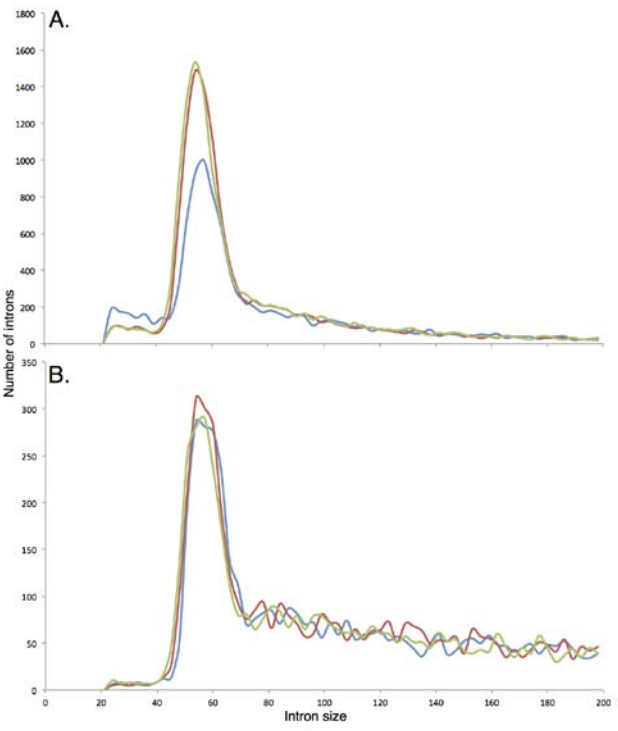

Figure 4. Size distribution of small introns (A) that contain no stop codons or (B) that do contain stop codons. Introns that are lengths multiples of 3 (3n) are shown in blue, 3n+1 in red, and 3n+2 in green.

References

[1] Chorev M., Carmel L., The function of introns, Front. Gene, 2012, 3, 55

[2] Barrett L.W., Fletcher S., Wilton S.D., Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements, Cell Mol Life Sci, 2012, 69, 3613-3634

[3] Parsch J., Selective constraints on intron evolution in Drosophila, Genetics, 2003, 165, 1843-1851

[4] Keightley P.D., Gaffney D.J., Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents, PNAS USA, 2003, 100, 13402-13406

[5] Andolfatto P., Adaptive evolution of non-coding DNA in Drosophila, Nature, 2005, 437, 1149-1152

[6] Halligan D.L., Keightley P.D., Ubiquitous selective constraints in the Drosophila genome revealed by genome-wide interspecies comparison, Genome Research, 2006, 16, 875-884

[7] Marais G., Nouvellet P., Keightley P.D., Charlesworth B., Intron size and exon evolution in Drosophila, Genetics, 2005, 170, 481-485

[8] Haddrill P.R., Charlesworth B., Halligan D.L., Andolfatto P. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content, Genome Biology, 2005, 6, R67

[9] Gazave E., Marqués-Bonet T., Fernando O., Charlesworth B., Navarro A., Patterns and rates of intron divergence between humans and chimpanzees, Genome Biol., 2007, 8, R21

[10] Collins K., Tetrahymena thermophila, In: Wilson L., Matsudaira P. (Eds.), Methods in Cell Biology, Academic Press, 2012

[11] Eisen J.A., Coyne R.S., Wu M., Thiagarajan M., Wortman J.R., Badger J.H., et al., Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote, Plos Biology, 2006, 4, 1620-1642

[12] Jaillon O., Bouhouche K., Gout J.F., Aury J.M., Noel B., Saudemont B., et al., Translational control of intron splicing in eukaryotes, Nature, 2008, 451, 359-362

[13] Hong X., Scofield D.G., Lynch M. Intron size, abundance, and distribution within untranslated regions of genes, Mol Biol Evol, 2006, 23, 2392-2404

[14] Kruger K., Grabowski P.J., Zaug A.J., Sands J., Gottschling D.E., Cech T.R., Self-splicing RNA - auto-excision and auto-cyclization of the ribosomal-RNA intervening sequence of Tetrahymena, Cell, 1982, 31, 147-157

[15] Stover N.A., Tetrahymena Genome Database (TGD): a new genomic resource for Tetrahymena thermophila research, Nucleic Acids Research, 2006, 34, D500-D503

[16] Edgar R.C., MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics, 2004, 5, 113

[17] Coyne R.S., Thiagarajan M., Jones K.M., Wortman J.R., Tallon L.J., Haas B.J., et al., Refined annotation and assembly of the Tetrahymena thermophila genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure, BMC Genomics, 2008, 9, 562

[18] Wuitschick J.D., Karrer K.M., Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila, Journal of Eukaryotic Microbiology, 1999, 46, 239-247

[19] Wuitschick J.D., Karrer K.M., Codon usage in Tetrahymena thermophila, Methods Cell Biol., 2000, 62, 565-568

[20] Moabbi A.M., Agarwal N., Kaderi B.E., Ansari A., Role of gene looping in intron-mediated enhancement of transcription. PNAS USA, 2012, 109(22), 8505-8510

[21] Irimia M., Maeso I., Burguera D., Hidalgo-Sánchez M., Puelles L., Roy S.W., et al., Contrasting 5' and 3' evolutionary histories and frequent evolutionary convergence in Meis/hth gene structures, Genome Biology, 2011, 3, 551-564

[22] Majewski J., Ott J., Distribution and characterization of regulatory elements in the human genome. Genome Research, 2002, 12, 1827-1836

[23] Jonsson J.J., Foresman M.D., Wilson N., McIvor R.S., Intron requirement for expression of the human purine nucleoside phosphorylase gene, Nucleic Acids Res, 1992, 20(12), 3191-3198

[24] Jonsson J.J., Converse A., McIvor R.S. An enhancer in the first intron of the human purine nucleoside phosphorylase-encoding gene, Gene, 1994, 140(2), 187-193

[25] Hadden T.J., Ryou C., Miller R.E., Elements in the distal 5'-flanking sequence and the first intron function cooperatively to regulate glutamine synthetase transcription during adipocyte differentiation, Nucleic Acids Res, 1997, 25 (19), 3930-3936

[26] Chen J., Hayes P., Roy K., Sirotnak F.M., Two promoters regulate transcription of the mouse folypolyglutamate synthetase gene three tightly clustered Sp1 sites within the first intron markedly enhance activity of promoter B, Gene, 2000, 242(1-2), 257-264

[27] Liu Y., Li H., Tanaka K., Tsumaki N., Yamada Y., Identification of an enhancer sequence within the first intron required for the cartilage-specific transcription of the alpha2(XI) collagen gene, J Biol Chem, 2000, 275(17), 12712-12718

[28] Charron M., Chern J.Y., Wright W.W., The Cathepsin L first intron stimulates gene expression in rat sertoli cells, Biol Reprod, 2007, 76, 813-824

[29] Xiong J., Lu X., Zhou Z., Chang Y., Yuan D., Tian M., et al., Transcriptome analysis of the model protozoan, Tetrahymena thermophila, using deep RNA sequencing, PLoS ONE, 2012, 7(2), e30630

[30] Kim E., Goren A., Ast G., Alternative splicing: current perspectives, Bioessays, 2007, 30, 38-47