Puzzles of the Human Genome: Why Do We Need Our Introns?

L. Fedorova[1] and A. Fedorov[1,2,*]

[1]*Department of Medicine and* [2]*Program in Bioinformatics and Proteomics/Genomics, Medical University of Ohio, Toledo, OH 43614, USA*

Abstract: Ninety five percent of human genomic DNA does not code for proteins or functional RNA molecules, and is frequently referred to as "junk" or "selfish" DNA. The vast majority of this noncoding DNA has no documented role in the cell. However, according to recent analyses, three quarters of the human genome is transcriptionally active. We discuss whether the expression of non-coding genomic sequences is valuable for the cell or if it is a second-hand "junk" because of the incompleteness in transcriptional machinery organization and functioning. Introns constitute a major fraction of the noncoding DNA, representing over 40% of mammalian genomes. They are ambivalent elements that cause several problems and at the same time bring benefits to their host cells. There is a strong correspondence between the average length of introns and the size of the genome. Here we review the latest summary statistics on human introns, the evolution of introns in mammals, and the distribution of genes that encode functional RNAs within introns. We also suggest that splicing is an important filter for organisms with large genomes, serving to distinguish between functional mRNAs and arbitrary RNA transcripts generated from random loci.

Received on: 26 October 2005 - Revised on: 10 November 2005 - Accepted on: 22 November 2005

Key Words: Gene, splicing, exon, evolution, genomics.

INTRODUCTION

The total length of all protein-coding mRNA sequences, non-coding RNA sequences with established functions, and regulatory elements that control their expression, comprises only about 5% of the mammalian genome. The remaining 95% of genomic DNA is frequently referred to as "junk" or "selfish" DNA. One could argue that, in theory, removing "junk" DNA from the genome would have no negative effects on the organism. This has in fact happened in one vertebrate species, the puffer fish *Takifugu rubripes*, whose genome shrank several times millions of years ago [1]. The general phenotype is essentially the same as that of closely-related genera, even though it has lost vast sections of its genome. "Junk" DNA has long been considered by many scientists as a playground for future evolution that provides the physical place for the origin of new genes and regulatory elements. Despite the fact that we cannot assign any valuable role to most of the "junk" DNA, the size of mammalian genomes seems evolutionarily stable and varies from 2.0 to 4.0 billion nucleotides for a majority of species from this taxon (see Animal Genome Size Database, [**2]). If we compare, for example, human and mouse, the size of their genomes differs only by 10%. This is an amazingly small difference considering that during the 70-90 million years after divergence of these two species, more than a million interspersed repetitive elements have been incorporated into their genomes (retroposons: *Alu* in *H. sapiens* and B1, B2, ERVs in *M. musculus*). In addition, thousands of independent deletions, chromosomal translocations, duplications, insertions and other rearrangements have occurred in the DNA of both organisms. These changes explain why 35% of the human genome does not align with the mouse genome using current computational methods [3]. Therefore, it appears that mammals follow a "golden proportion" between the number of their genes and the size of their genomes. This suggests a mechanism that has kept our genomes at this size for millions of years. The evolution of genome lengths was reviewed in Petrov [4] and recently by Vinogradov [*5] and Gregory [*6], while the possible involvement of small RNAs in the control of genome length was discussed by Hennig [7].

In this review, we concentrate on the structure and functions of mammalian introns, the ubiquitous elements of our genes that represent at least 40% of our genome and whose role is still poorly understood and appreciated. Introns are notoriously known for controversies in the interpretations of their origin and functions in cells. A range of recent studies has confirmed that introns already existed at the earliest stages of eukaryote evolution [8-11, **12]. The evolution of introns has been well reviewed from different perspectives by Lynch and Richardson [13], Collins and Penny [*14], and Rogozin *et al.* [**15]. There is still no consensus on the time of origin of these elements, or of their initial role. Following Gilbert [16] and Poole *et al.* [17], we recently suggested that introns were among the most ancient of genetic elements existing in the RNA world, where they governed the fate of different classes of RNA molecules [*18].

INTRON STATISTICS

According to the current dataset release (GenBank build 35, December 2004), the human genome contains 21,746 protein coding genes that possess introns and 1,760 intronless genes. Hence, only 8% of human protein coding genes are intron-free. Altogether, within protein-coding genes,

*Address correspondence to this author at the [1]Department of Medicine and [2]Program in Bioinformatics and Proteomics/Genomics, Medical University of Ohio, Toledo, OH 43614, USA; Tel: (419)-383-5270; Fax: (419)-383-3102; E-mail: afedorov@meduohio.edu

1389-2029/05 $50.00+.00 **©2005 Bentham Science Publishers Ltd.**

there are 190,159 introns, so the average segmented gene contains between 8 and 9 introns. The total length of introns is 1.1 billion nucleotides, representing 37% of the euchromatic part of the human genome. The average size of human introns is 5,667 bp, while the median is 1,504 bp. Many introns are extremely long. For instance, 1,234 introns are longer than 100 kb; 299 are longer than 200 kb; and 9 are longer than 500 kb (human Exon-Intron Database, www.meduohio.edu/bioinfo/eid/ [19, *20]). The longest human intron which spreads over 740,920 bp is found in heparan sulfate 6-O-sulfotransferase 3 gene (HS6ST3) on chromosome 13; for comparison, this is about 50% larger than the chromosome of the smallest free-living bacterium.

The correlation between genome size and the length of introns in different organisms was thoroughly reviewed by Vinogradov [21]. He showed that evolutionary alteration of genome size and intron size are tightly coupled processes. Yet some species have unique proportions of intron size relative to their genome [21]. An interesting example of fast evolutionary genome shrinkage was observed in *Takifugu* fish. In this case, the diminution of *Takifugu* intron lengths and the length of its intergenic regions were highly coordinated. Despite dramatic shortening of the *Takifugu* genome, the number of introns remains the same as other vertebrates [22]. The process of intron loss is extremely rare in vertebrates. In mammals, it happens only with short introns, whose lengths are less than 300 bp [23]. Since the median of human introns is 1,504 bp, most intronic sequences will be long companions of the human genome.

Intron length alteration during evolution is illustrated in the Figs. (**1**-**3**). Fig. (**1**) represents length comparison of 79,931 orthologous introns of human and mouse from Mammalian Orthologous Intron Database (MOID) [*20]. We define "orthologous introns" as introns from orthologous genes that also have the same position relative to the two coding sequences. As far as we know, there are no thoroughly characterized cases of intron gain and only solitary cases of intron loss in mammals [23]. (Few published examples of mammalian intron gain [24, 25] could have alternative explanations as described in the coming paper by Shepelev and Fedorov [26].) Hence, orthologous introns most likely descended from the corresponding intronic sequence of the last common ancestor of the two species.

As illustrated in Fig. (**1**), differences in length of orthologous introns from human and mouse can be considerable. Changes in intron length are presumably due to an imbalance between two opposite genomic processes. The first process, genome growth, occurs through the insertions of interspersed repetitive elements, duplications of genomic segments, and micro-insertions from one to several nucleotides. The second process genome contraction is caused by deletions (*via* DNA recombination between interspersed repeats or other rearrangements of genomic sequences) and by micro-deletions from one to several nucleotides. It is generally believed that genome expansion, to a great extent, involves retroposition of interspersed repetitive elements such as *Alu*-repeats in human and *B1*-, *B2*-, *ERV*- repeats in mouse [**27]. However, Fig. (**2**), which represents the distribution of human-

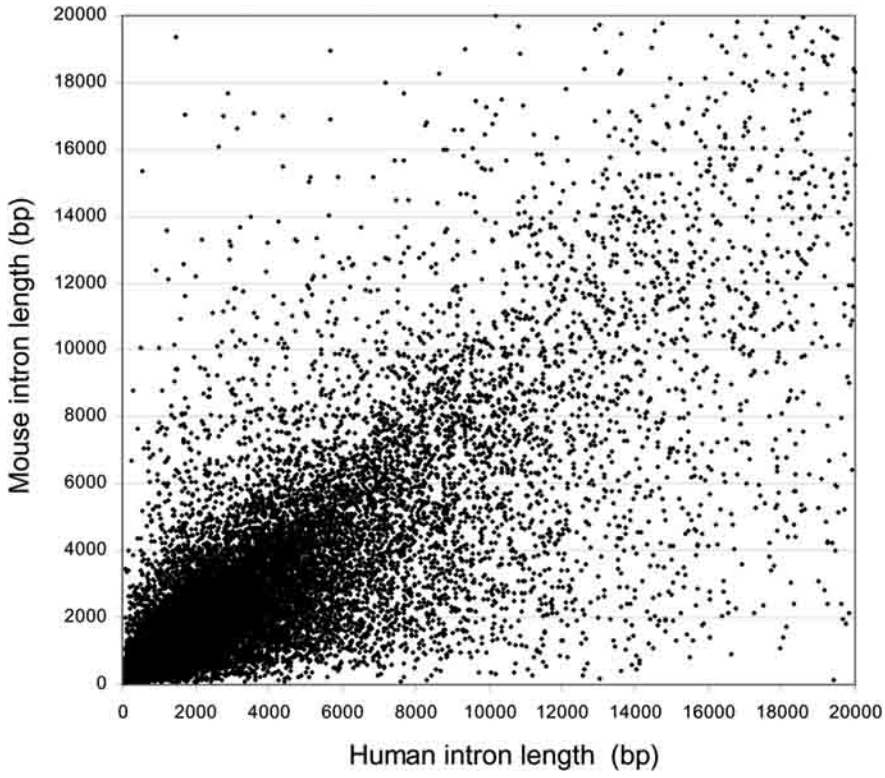

Fig. (1). Length distribution of human and mouse orthologous introns. Each dot represents a single mouse-human orthologous intron pair, where the horizontal axis shows the length of the human intron, and the vertical axis shows the length of the corresponding mouse ortholog. 79,931 orthologous intron pairs from the September 2005 release of Mammalian Orthologous Intron Database [20] are used for this figure. The graph presents all introns that are shorter than 20,000 bp (95% of all introns).

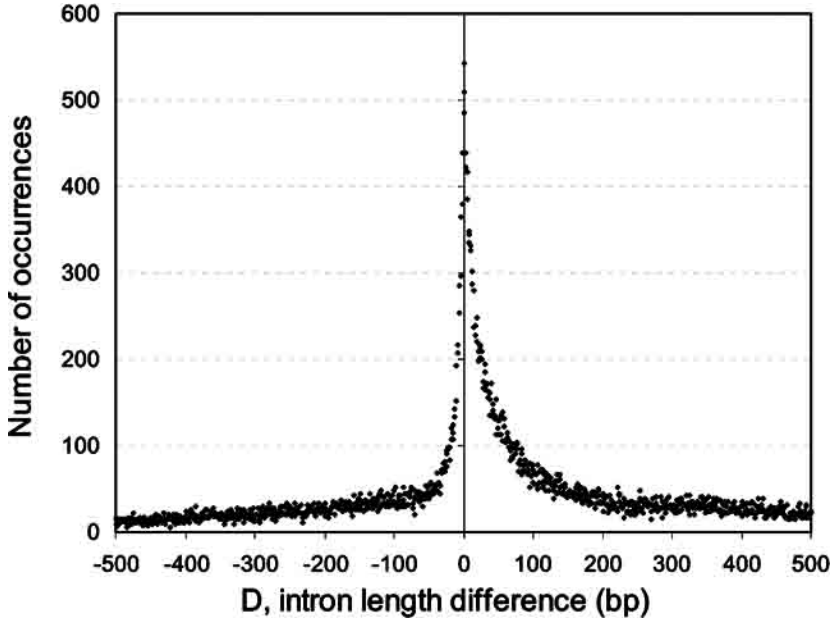

Fig. (2). Distribution of length difference between orthologous introns of human and mouse. The horizontal axis shows the difference in length between human and mouse introns ($D = L_{human} - L_{mouse}$). The vertical axis shows the number of orthologous intron pairs that have the length difference of D nucleotides. [NOTE: In our calculations we used best-hit approach applied to all large-scale computations of orthologous sequences [64]. For stringency, we studied only orthologous introns obtained for three species (human-mouse-rat 79,931 orthologous intron triplets from MOID [20]).

mouse orthologous intron length differences, shows no specific peak of either +300 bp which would correspond to the insertion of Alu-repeats inside human introns, or of a peak that would correspond to the insertion of B1- and B2- repeats into mouse introns (about -200 to -150 bp). These data suggest that retroposons have not played a dominant role in changing mammalian intron size.

A change of intron length is a complex multifactorial process where small insertions/deletions are among the major contributors in the regulation of intron length during evolution. Interestingly, micro deletions overwhelm micro insertions in human introns [28]. Our calculation of length difference between 79,931 human-mouse orthologous intron pairs show that, on average, human introns are 21% longer than those of mouse. The total length of human introns studied in our sample was 327,253,533 bp, while the total length of their orthologous counterparts in mice was shorter by 69,169,624 bp. Therefore, the human-mouse intron length difference is double the whole-genome length difference between these two species, which is ~10% (human genome 2.9 Gb [29] and mouse genome 2.6 Gb, [30]). This observation is readily explicable. Introns represent transcribed parts of the genome that are more susceptible to sequence change (due to relatively unpacked chromatin structures) than the more condensed chromatin regions often detected between genes. This year interesting details were described by Haddrill and co-authors that a negative correlation exists between intron length and sequence divergence in Drosophila [31]. Yet, our data on the relative intron length difference displayed in Fig. (**3**) does not support the same trend in mammals. Fig. (**3**) shows that short mammalian introns tend to be more resistant to length alterations than long introns.

By analogy, we calculated that, on average, human introns are 27% longer than rat introns, while mouse introns are 5% longer than their rat counterparts. The latter fact looks surprising, considering that the rat genome (2.75 Gb) exceeds that of mouse (2.6 Gb) [30]. However, large segmental duplications are more abundant in the rat genome compared to the mouse genome [30]. These duplications enlarge the genome size while having no effect on the interspecies comparisons of orthologous intron lengths.

We have discussed only protein-coding genes so far. However, during last two years about seventeen thousand non-protein coding genes have been characterized in humans [**32]. Thirty percent of them are spliced. We still know practically nothing about the introns of these non-protein coding genes. Their characterization in the nearest future should advance our knowledge in splicing mechanisms and in prediction of exon-intron gene structures. Taking into account all types of genes, the total length of human introns appears to comprise more than 40% of the genome.

INTRON FUNCTIONS

The enormous intron size in humans and other vertebrates creates several drawbacks, such as: 1) considerable waste of energy during gene expression which is "unwisely" spent on polymerizing extra-long intronic segments of pre-mRNA molecules; 2) delay in obtaining protein products (on average it takes about 45 min for RNA polymerase II to transcribe a 100,000 bp intron); 3) potential errors in normal splicing, since long introns contain numerous false splicing sites (so-called pseudo-exons [33]). Some benefits must be associated with introns to compensate for these disadvantages. We have already reviewed different constructive roles

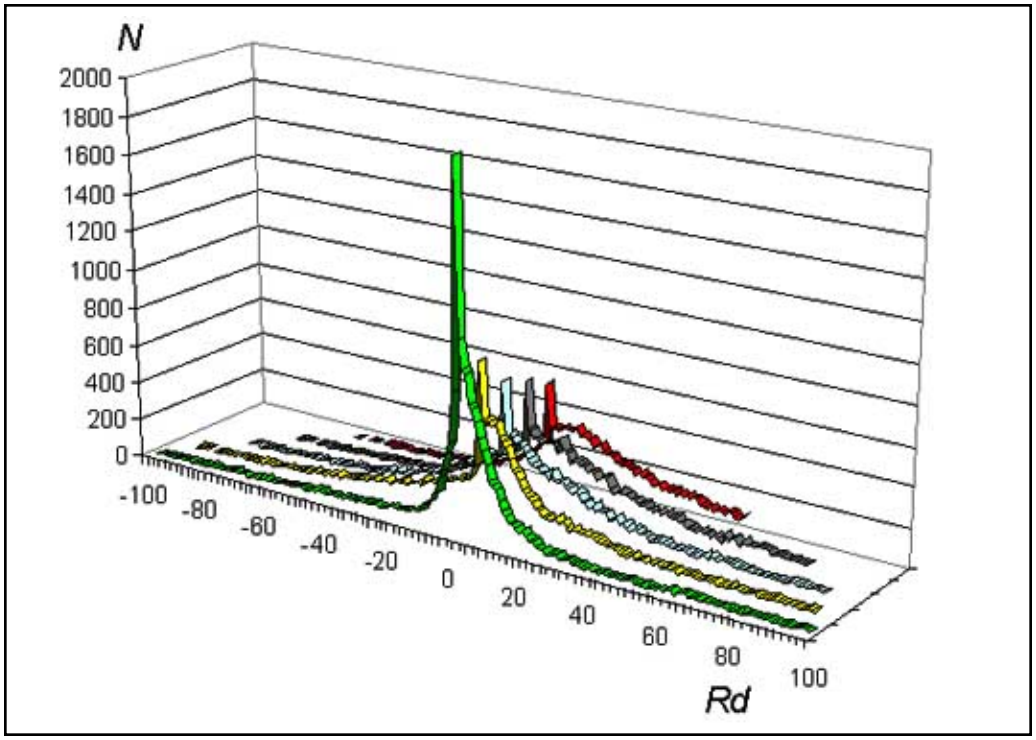

Fig. (3). Distribution of relative length difference between orthologous introns of human and mouse with respect to the size of human introns. Relative intron length difference, plotted on horizontal axis, was calculated as follows:

$Rd = [(L_{human} - L_{mouse})/(L_{human} + L_{mouse})] * 100\%$. Number of occurrences (*N*) of orthologous introns with the same Rd were calculated for each 1% interval (0-1%, 1-2%, …, 99-100%) and plotted on vertical axis. All human introns were divided into five groups according to their length: 1) short introns (less than 339 bp); 2) short-medium introns (339-969 bp); 3) medium introns (969-2000 bp); 4) medium-long introns (2000-4567 bp); and 5) long introns (longer than 4679 bp). All these groups are of equal size and each one contains 20% of all introns. First line (green) shows the distribution for short introns; the second for short-medium introns (yellow); the third for medium introns (blue); the fourth (grey) for medium-long introns; and the fifth (red) for long introns

for introns [34]. We update this discussion in light of novel findings coming from large-scale mammalian genome analyses. We do not touch upon alternative splicing here because this important topic has recently been reviewed elsewhere [35-39].

Non-Coding RNAs Inside Introns

Introns contain several types of non-coding but functional RNA sequences (ncRNAs). All known mammalian small nucleolar RNAs (snoRNAs) are located inside introns and are produced during post-splicing processing of intronic RNA [40, 41]. More than half of these snoRNAs are inside introns of protein-coding genes. Other snoRNAs are within introns of non-coding genes, whose exons likely do not have particular cellular roles and whose primary function may be simply to express the small functional RNAs [42]. Another type of ncRNA, microRNAs, are also found frequently inside introns [43]. According to the highest estimation, about three quarters of mammalian microRNAs are located within introns [44]. Currently, over 300 small ncRNAs have been reported within human introns and are available from the RNAdb database [45] and in a microRNA database [46]. Recently, the SNO.pl program was developed and used to reveal evolutionarily-conserved C/D box snoRNAs in mammalian introns [20]. When used to examine all human introns, around 1000 C/D snoRNA-like structures were found, each including the entire set of features found in true snoRNAs, yet these sequences are specific to human and do not have conserved counterparts in mouse or rat introns (our unpublished data). Therefore, in addition to hundreds of already-characterized small ncRNA, human introns may be used to create thousands of yet unknown ncRNAs specific to the species. In fact, some other types of ncRNA have been associated recently with introns [47]. Due to the presence of ancient ncRNAs within introns Poole and co-authors have hypothesized that introns are primordial genetic elements descended from the RNA-world [17].

Splicing and Arbitrary Transcription of the Mammalian Genome

Together, the exons and introns of protein-coding genes comprise 38% of the human genome. Since about $^3/_4$ of the human genome is transcriptionally active [48, 49, **27, **32], many intergenic regions and complementary strands of genes that were long considered to be transcriptionally silent, are now recognized as expressed genomic loci. It is not clear whether these loci have some function or instead represent the products of arbitrary transcription. Probably both answers are true. A number of ncRNA genes have been discovered within intergenic segments located between protein coding genes. These ncRNA genes represent all kinds of functional RNA molecules: from extra-long, (>100,000 bp)

like *Air*-RNA [50, 51], to extra-short, representing 22 bp long microRNAs [43].

Nevertheless, the organization of gene expression machinery implies that in mammalian genomes there should be a considerable amount of arbitrary transcription from random genomic segments. Indeed, one of the most important transcription activation elements is the TATA-box, which represents a short six nucleotide-long motif with the consensus "TATAAA" sequence. Some degeneration or sequence flexibility of adenines and thymines is allowed inside the TATA-box. For example, in the human beta-hemoglobin gene, this regulatory element is represented by an "ATAA AA" sequence. A known mutation of this motif, which converts its sole T into A, considerably reduces but does not abolish expression of the gene [52]. In rough approximation, a six nucleotide-long oligonucleotide should occur once in every four thousand nucleotides of genomic sequence (probability of 4^{-6} = 1/4096). (For more precise estimation someone should take into account the non-randomness of genomic sequences frequently referred to as "genomic signatures" [53]).

Taking into account the degeneracy of the TATA-box, it is not surprising that it is present several million times at random sites in the genome of humans and other vertebrates (TATAAA oligonucleotide itself appears 4,155,589 times in both strands of published human genome sequences). For an active transcription, a few additional, so-called upstream regulatory elements should be present within about 200 bp of the TATA-box [54, 55]. The key issue concerning transcription activation is the segmentation of regulatory sequences into several short motifs and the flexibility in their arrangement and distances relative to each other. Consider, for example, three six-nucleotide-long elements with fixed positions relative to each other. The chance of getting this particular pattern of 18 bases in the entire human genome is about 9% (4^{-18} multiplied by the entire length of both strands of the human genome = $5.8*10^9$). However, if we allow the same three elements to be in different arrangements within 200 bp from each other, we should find more than three thousand of these six-nucleotide-long regulatory triplets in the genome ($188 * 182 * 4^{-18} * 5.8*10^9$). There are dozens of different regulatory elements that activate transcription in particular tissues and at particular times [56, 57]. Therefore, the arbitrary transcription of random sites in the mammalian genome seems an inevitable consequence for the extra-long expanses of "junk" DNA. The genome could protect itself to some extent from this arbitrary transcription by specific sequences and DNA-modifications (such as methylation) that make many genomic loci inactive by converting them into compact heterochromatin. However, there exists a constant process of sequence change in the course of evolution. In addition to point mutations the genome undergoes insertions, duplications, translocations, and deletions of sequences of various sizes and compositions. All of these events should provide constant generation of novel random sites for transcription initiation, which was statistically confirmed by Dermitzakis and Clark [58]. Theoretically, the larger the genome the bigger is the proportion that such arbitrary random transcripts should form in the expressed pool of RNA molecules. Having very large genomes, mammals and vertebrates should be greatly affected by this problem. Splicing could be a valuable mechanism in the separation of functional mRNAs from random transcripts. Maniatis and Reed [59] emphasized in their review that during pre-mRNA processing dozens of different splicing factors bind tightly to the transcribed molecules. These serve as specific signals for exporting mature mRNA from the nucleus into the cytoplasm and for regulation of translation and degradation of mRNA. Because the majority (70%) of non-protein coding transcripts does not undergo splicing [**32], these molecules are not accompanied by splicing factors and, thus, cannot follow the pathway of mRNA leading to the production of proteins. The remaining 30% of non-protein coding genes that undergo splicing are still practically uncharacterized. They might include those ncRNA whose introns represent microRNA, snoRNA, or other types of functional ncRNAs. All in all, splicing seems ideally suited to serve as an important filter for organisms with extra-large genomes, helping to distinguish between functional mRNAs and random RNA transcripts. This intriguing property of splicing has been utilized in molecular biology for nearly two decades by incorporating short introns into various transfection vectors. Inclusion of an intron increases the protein expression from all cDNA inserts in the transgenic constructions [60, 61]. As demonstrated by Kurachi and others, the expression enhancing activity of introns in transgenic constructions is not due to specific enhancer elements within introns, but due to activation of splicing process itself [62].

CONCLUDING REMARKS

Some proteins play several unrelated roles in host cells, a phenomenon called "moonlighting" [63]. It is reasonable to expect that introns could acquire different roles in the cells during evolution in the same manner as proteins. In the compact genomes of flies, worms, and other invertebrates, introns are relatively small and have one set of cellular tasks as we described previously [34]. Mammalian genomes have grown upwards of 30 times the size of insect and nematode genomes, and so have the introns along with them. Mammalian genome complexity posed new evolutionarily challenges in regulating gene expression and in dealing with a great number of different classes of RNA molecules, including random transcripts. We propose that mammalian introns have adapted to these genome challenges and that splicing serves as an important filter for selection of mRNAs against random transcripts.

ACKNOWLEDGEMENTS

Support for this work was provided by the Medical University of Ohio Foundation and the Stranahan Foundation, through the Program in Bioinformatics and Proteomics/ Genomics. We would like to thank Robert Blumenthal and Peter Bazeley, Medical University of Ohio, for discussion and suggestions on our manuscript.

REFERENCES

[1] Venkatesh, B., Yap, W-H. Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *BioEssays* **2005**, *27*: 100-107.

[2]** Gregory, T.R. *Animal Genome Size Database*. **2005**, http://www.genomesize.com/. [A comprehensive database of genome sizes for four thousand species. Many interesting links are provided].

[3] Miller, W., Makova, K.D., Nekrutenko, A., Hardison, R.C. Comparative genomics. *Ann. Rev. Genomics Hum. Genet.* **2004**, *5*: 15-56.

[4] Petrov, D.A. Evolution of genome size: new approaches to an old problem. *Trends Genet.* **2001**, *17*: 23-28.

[5]* Vinogradov, A.E. Evolution of the genome size: multilevel selection, mutation bias or bynamical chaos? *Curr. Opin. Genet. Dev.* **2004**, *14*: 620-626. [A review of new data and hypotheses on the evolution of eukaryotic genome size].

[6]* Gregory T.R. Insertion-deletion biases and the evolution of genome size. *Gene* **2004**, 324:15-34. [The author reviews the modern approaches to C-value paradox – discordance between genome size and organism complexity].

[7] Hennig, W. The revolution of the biology of the genome. *Cell Res.* **2004**, *14*: 1-7.

[8] De Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S., Gilbert, W. Towards a resolution of the introns early/late debate. Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA* **1998**, *95*: 5094-5099.

[9] Long, M., de Souza, S.J., Gilbert, W. Relationship between "proto-splice sites" and intron phases: Evidence from Dicodon Analysis. *Proc. Natl. Acad. Sci. USA* **1998**, *95*: 219-223.

[10] Fedorov, A., Merican, A.F., Gilbert, W. Large-scale comparison of intron positions between plant, animal and fungal genes. *Proc. Natl. Acad. Sci. USA* **2002**, *99*: 16128-16133.

[11] Rogozin, I.B., Wolf, Y.I, Sorokin, A.V., Mirkin, B.G., Koonin, E.V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **2003**, *13*: 1512-1517.

[12]** Roy, S.W., Gilbert, W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci. USA* **2005**, *102*: 5773-5778. [The authors argue that the last common ancestor of eukaryotes had nearly as many introns per gene as in humans].

[13] Lynch, M., Richardson, A.O. The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **2002**, *12*: 701-710.

[14]* Collins, L., Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **2005**, *22*: 1053-1066. [Structure and evolution of the spliceosomal complex is investigated. The authors inferred properties of the splicing system in the last common ancestor of eukaryotes].

[15]** Rogozin, I.B., Sverdlov, A.V., Babenko, V.N., Koonin, E.V. Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform.* **2005**, *6*: 118-134. [Latest comprehensive review of exon-intron gene structure and evolution based on analyses of multiple complete eukaryotic genome sequences].

[16] Gilbert, W. The exon theory of genes. *Cold Spring Harbor Symp Quant Biol.* **1987**, *52*: 901-905.

[17] Poole, A.M., Jeffares, D.C., Penny, D. The path from the RNA world. *J. Molec. Evol.* **1998**, *46*: 1-17.

[18]* Fedorov, A., Fedorova, L. Introns: mighty elements from RNA world. *J. Molec. Evol.* **2004**, *59*: 718-721. [For ancient DNA-lacking cells, a crucial problem existed in distinguishing two distinct subsets of RNAs: those messenger molecules coding for proteins and those heritable genetic molecules complementary to messenger RNAs that propagate the genetic information through generations. The authors proposed that ancient introns could act as markers of RNA subsets, directing them to different functions].

[19] Saxonov, S., Daizadeh, I., Fedorov, A., Gilbert, W. EID: The Exon-Intron Database: An exhaustive database of protein-containing genes. *Nucl. Acids Res*. **2000**, *28*: 185-190.

[20]* Fedorov, A., Stombaugh, J., Harr, M.W., Yu, S., Nasalean, L., Shepelev, V. Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database. *Nucl. Acids Res.* **2005**, *33*: 4578-4583. [MOID database, which contains all known introns within the human, mouse, and rat genomes is presented for public usage. Applications for searching evolutionarily conserved motifs within mammalian introns are described].

[21] Vinogradov, A.E. Intron-genome size relationship on a large evolutionary scale. *J. Molec. Evol.* **1999**, *49*: 376-384.

[22] Elgar, G. Quality not quantity: the pufferfish genome. *Hum. Molec. Genet.* **1996**, *5*: 1437-1442.

[23] Roy, S.W., Fedorov, A., Gilbert, W. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. USA* **2003**, *100*: 7158-7162.

[24] Veeramachaneni, V., Makalowski, W. DED: Database of Evolutionary Distances. *Nucleic Acids Res.* **2005**, *33*(Database issue): D442-446.

[25] O'Neill, R.J., Brennan, F.E., Delbridge, M.L., Crozier, R.H., Graves, J.A. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc. Natl. Acad. Sci. USA* **1998**, *95*: 1653-1657.

[26] Shepelev, V., Fedorov, A. Application of Exon-Intron Database (EID) for gene structure analysis. *Briefings in Bioinformatics*, **2006**, Accepted.

[27]** Brosius, J. Disparity, adaptation, exaptation, bookkeeping and contingency at the genome level. *Paleobiology* **2005**, *31*: 1-16. [The author reviews the intricate course of eukaryote genome evolution and the versatile involvement of RNA in this process].

[28] Vinogradov, A.E. Growth and decline of introns. *Trends Genet.* **2002**, *18*: 232-236.

[29] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431*: 931-945.

[30] Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **2004**, *428*: 493-521.

[31] 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 Biol.* **2005**, *6*: R67.

[32]** Suzuki, M., Hayashizaki, Y. Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *BioEssays* **2004**, *26*: 833-843. [The authors summarize the data obtained from 60,770 mouse full-length cDNA clones and compare them with results from the human genome sequencing project. They provide exhaustive evidence that non-coding RNAs are a major component of the transcriptomes of higher organisms].

[33] Sun, H., Chasin, L.A. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **2000**, *20*: 6414-25.

[34] Fedorova, L., Fedorov, A. Introns in gene evolution. *Genetica* **2003**, 118: 123-131.

[35] Matlin, A.J., Clark, F., Smith, C.W. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*: 386-398.

[36] Lee, C., Wang, Q. Bioinformatics analysis of alternative splicing. *Brief Bioinform.* **2005**, *6*: 23-33.

[37] Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., Soreq, H. Function of alternative splicing. *Gene* **2005**, *344*: 1-20.

[38] Sorek, R., Shamir, R., Ast, G. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **2004**, *20*: 68-71.

[39] Lareau, L.F., Green, R.E., Bhatnagar, R.S., Brenner, S.E. The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.* **2004**, *14*: 273-282.

[40] Huttenhofer, A., Brosius, J., Bachellerie, J.P. RNomics: identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.* **2002**, *6*: 835-843.

[41] Hirose, T., Shu, M-D., Steitz, J.A. Splicing-dependent and -independent modes of assembly for intron-encoded box C/D snoRNA in mammalian cells. *Molec. Cell* **2003**, *12*: 113-123.

[42] Filipowicz, W., Pogacic, V. Biogenesis of small nucleolar ribonucleoproteins. *Curr. Opin. Cell Biol.* **2002**, *14*: 319-327.

[43] Bartel, D.P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **2004**, *116*: 281-297.

[44] Cullen, B.R. Transcription and processing of human microRNA precursors. *Mol. Cell* **2004**, *16*: 861-865.

[45] Pang, K.C., Stephen, S., Engstrom, P.G., Tajul-Arifin, K., Chen, W., Wahlestedt, C., Lenhard, B., Hayashizaki, Y., Mattick, J.S. RNAdb - a comprehensive mammalian noncoding RNA database. *Nucl. Acids Res. (Database Issue)* **2005**, *33*: D125-D130.

[46] Griffiths-Jones, S. The microRNA Registry. *Nucl. Acids Res.* (Database Issue) **2004**, *32*: D109-D111. http://microrna.sanger.ac.uk/sequences/index.shtml.

[47] Reis, E.M., Louro, R., Nakaya, H.I., Verjovski-Almeida, S. As antisense RNA gets intronic. *OMICS* **2005**, *9*: 2-12.

[48] Mattick, J.S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* **2003**, *25*: 930-939.

[49] Frith, M.C., Pheasant, M., Mattick, J.S. The amazing complexity of the human genome. *Eur. J. Hum. Genet.* **2005**, *13*: 894-897.

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

[51] Wutz, A. RNAs templating chromatin structure for dosage compensation in animals. *BioEssays* **2003**, *25*: 434-442.

[52] Fei, Y.J., Stoming, T.A., Efremov, G.D., Efremov, D.G., Battacharia, R., Gonzalez-Redondo, J.M., Altay, C., Gurgey, A., Huisman, T.H. Beta-thalassemia due to a T-A mutation within the TATA box. *Biochem. Biophys. Res. Commun.* **1988**, *153*: 741-747.

[53] Karlin, S., Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature *Trends Genet.* **1995**, *11*: 283–290

[54] Singh G.B., Singh H. Databases, models, and algorithms for functional genomics. A bioinformatics perspective. *Molec. Biotechnol.* **2005**, *29*: 165-183.

[55] Davidson, E.H. *In: Genomic Regulatory Systems.* **2001**, San Diego, Academic Press.

[56] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucl. Acids Res.* **2003**, *31*: 374-378.

[57] Fogel, G.B., Weekes, D.G., Varga, G., Dow, E.R., Craven, A.M., Harlow, H.B., Su, E.W., Onyia, J.E., Su, C. A statistical analysis of the TRANSFAC database. *Biosystems* **2005**, *81*: 137-154.

[58] Dermitzakis, E.T., Clark, A.G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **2002**, *19*: 1114-1121.

[59] Maniatis, T., Reed, R. An extensive network of coupling among gene expression machines. *Nature* **2002**, *416*: 499-506.

[60] Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., Palmiter, R.D. Introns increase transcriptional efficiency in transgenic mice. *Proc. Natl. Acad. Sci. USA* **1988**, *85*: 836-840.

[61] Choi, T., Huang, M., Gorman, C., Jaenisch, R. A generic intron increases gene expression in transgenic mice. *Mol. Cell. Biol.* **1991**, *11*: 3070-3074.

[62] Kurachi , S., Hitomi, Y., Furukawa, M., Kurachi, K. Role of Intron I in Expression of the Human Factor IX Gene. *J. Biol. Chem.* **1995**, *270*: 5276-5281.

[63] Jeffery, C.J. Moonlighting proteins: old proteins learning new tricks. *Trends Genet.* **2003**, *19*: 415-417.

[64] Tatusov, R.L., Koonin, E.V., Lipman, D.J. A genomic perspective on protein families. *Science* **1997**, *278*: 631-637.