Genes with a large intronic burden show a higher evolutionary conservation on protein level

Ivan P. Gorlov[1]*, Alexei N. Fedorov[2], Christopher Logothetis[1], Olga Y. Gorlova[3], and

[4]Christopher Amos[3]

[1]Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA, [2]University of Toledo, Toledo, OH, USA [3]Department of Eidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA and [3]Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA

ABSTRACT

Exon/intron structure of eukaryotic genome may provide an evolutionary advantage by increasing protein diversity through alternative splicing and exon shuffling. On the other hand, removing intronic sequences from preRNA requires energy and may lead to errors. The goal of this study was to analyse relationships between intronic burden, described as a number of introns and total intronic size, and level of evolutionary conservation of the gene.

In an initial analysis we noted a positive correlation between size of the intronic region and the conservation index (CI) of the corresponding protein. In the multivariable model, both the number of introns and intronic size remained significant predictors of CI.

We found that RNA sequencing based expression levels negatively correlated with the number of introns suggesting that large multiple introns impede expression. Genes with multiple large introns carry significantly fewer missense and stop-gained (nonsense) mutations suggesting their stronger functional significance.

We hypothesized that the positive correlation between the number of introns and CI reflects the fact that only functionally important (and therefore more evolutionary conserved)

genes can tolerate multiple introns without risking that cost of having large multiple introns will exceed its benefits.

INTRODUCTION

One of the notable features of the eukaryotes is the intron-exon structure of their genes. The intron-exon structure is one of the hallmarks of eukaryotic evolution and is believed to be evolutionary beneficial because it allowed production of multiple proteins from the same gene through alternative splicing and accelerated the creation of novel genes through exon shuffling [1-3]. The review of the possible evolutionary advantages of spliceosomal introns can be found in the paper of Fedorova and Fedorov [4].

The origin and evolution of intron-exon structure remains an area of discussion [5-8]. The size of the intronic region of the gene has been shown to be correlated with important biological characteristics of the gene, e.g. it has been demonstrated that highly expressed genes tend to have shorter introns [9].

Despite of progress in understanding of origin, evolution, and biological significance of introns, little is known about forces that have shaped intron evolution [10]. The goal of this study is to identify biologically meaningful characteristics of genes correlated with intronic burden of the gene.

MATERIALS AND METHODS

The dataset included 16,194 human protein coding genes for which we were able to obtain estimates of the conservation index (CI). The presence of orthologs in the evolutionarily most distant lineage was used as conservation index (CI). CIs were esitmated based on the data from HomoloGene database http://www.ncbi.nlm.nih.gov/homologene . The CI changes

from 0 when a gene is unique to *Homo sapiens* to 9 when the human ortholog is detected in plants (Table 1). The approach we have used is similar to the phylostratigraphic approach used by Domazet-Loso and Tautz[11]. The only difference was that we used available alignments from the HomoloGene database,[12,13] and Domazet-Loso and Tautz performed their own protein alignments.

Table 1 Definition of conservation index (CI)

The most distant species with detectable human ortholog	Phylogenic group	CI
Homo sapiens	Unique to humans	0
Pan troglodites	Great apes	1
Macaca mullata	Primates	2
Ratus norvegicus	Rodents	3
Gallus gallus	Birds	4
Danio rerio	Fishes	5
Anopheles gambiae	Insects	6
Caenorhabditis elegans	Worms	7
Schizosaccharomyces pombe	Fungi	8
Oryza sativa	Plants	9

For the average gene expression level in normal tissues we used data from the RNA-Seq Atlas database [14]. The database provides RNA-sequence based estimates of the gene expression across 10 diverse human tissues: colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes. The estimates of the gene expression using next generation RNA sequencing are much more comprehensive and unbiased compared to the probe-based technologies. Z-scores were used as quantitative measures of the gene expression: see [14] for details.

To assess the association between the density of the potentially functional polymorphisms in coding regions and the number of introns (or size of the intronic region) we used the data from NCBI dbSNP database. Only sequence identified validated SNPs have been used for the analysis. Data on the number of synonymous, nonsynonymous and stop gained (those producing stop codon) variants were extracted from the database. To control for the differences in the gene size we used the number of SNPs per one codon as a measure of the SNP density.

RESULTS AND DISCUSSION

Distribution of the human genes by number of introns.

There is a vast variation between human genes in the number of introns. About 600 human genes are intronless [15]. On the other side of the distribution there is TTN gene (gene ID 5273) with more than 300 introns. The average number of introns per human gene is between 8 and 9 [10]. Figure 4 shows the distribution of the human genes by the number of introns in the sample we have used. The proportion of the genes with a small number of introns (0-2) is relatively low: 2, 4 and 6% correspondingly. Genes with 3 to 6 introns are most common: together they comprise more than 30% of the genes. Genes with a large number of introns are rare: genes with more than 30 introns comprise less than 5%.

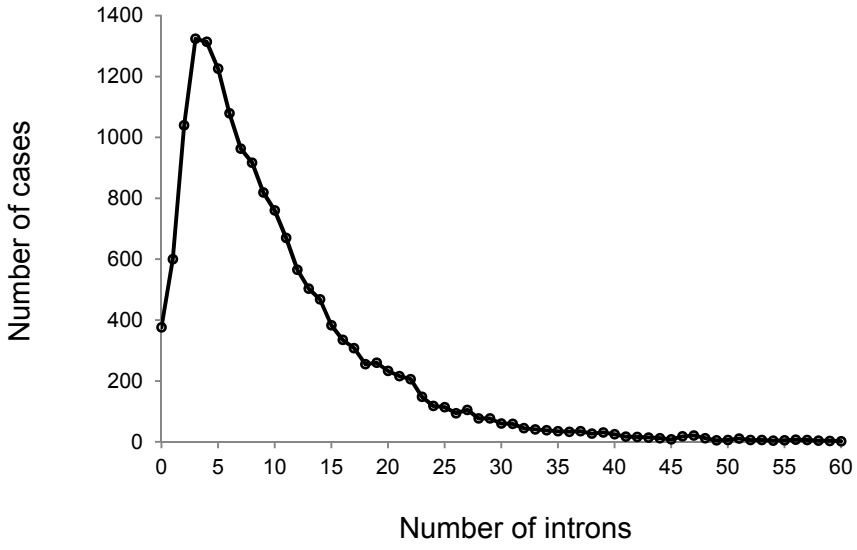

Figure 1. Distribution of human genes by the number of introns.

Dependence of conservation index (CI) on the number of introns and the size of intronic region

There is a significant positive correlation between CI and the size of the intronic region of the gene: r = 0.06, N = 16,194, $P < 10^{-6}$. Figure 2 (a, b) shows the dependence between the intronic size of the gene and its conservation index. The curve can be divided into 3 segments. Segment 1 encompasses intronic sizes from the smallest of 18 nucleotides (in the *GRP152* gene) to up to 3kb. The correlation between CI and intronic size for this segment is r = 0.23, N = 1,420, $P < 10^{-6}$. In the second segment (genes with intronic length of 3 to 30 kb) the correlation coefficient between CI and intronic size is r = 0.10, N = 6,642, $P < 10^{-6}$. For the third segment (genes with intronic length of >30 kb) the correlation between CI and intronic size is not significant: r = -0.02, N = 8,189, P < 0.08.

There is also a significant positive correlation between the number of introns and conservation index: r = 0.16, N = 16,249, $P < 10^{-6}$. Figure 2 (c) shows the relationship between the number of introns and CI. The curve has two segments: linear positive correlation from 0 to 10 introns and plateaued part from for the genes with the number of introns ≥ 11.

Because there is a positive correlation between the number of introns and the size of the intronic region ($r = 0.34$, $N = 16{,}249$, $P < 10^{-6}$), it is possible that the correlation between the intronic size and CI is driven by number of introns or vice versa. We applied multivariable linear regression model using genes in the segments with linear dependence to estimate independent effects of the intronic size and the number of introns as predictors of CI. For the segment 1 the relationship between the intronic size and CI after controlling for the number of introns considerably dropped but remained significant: $r = 0.11$, $N = 1{,}420$, $P < 0.00003$. The correlation between the number of introns and CI also remained significant after controlling for the intronic size: $r = 0.21$, $N = 1{,}420$, $P < 10^{-6}$. The results suggest that both the number of introns and the size of the intronic region independently contribute to CI.

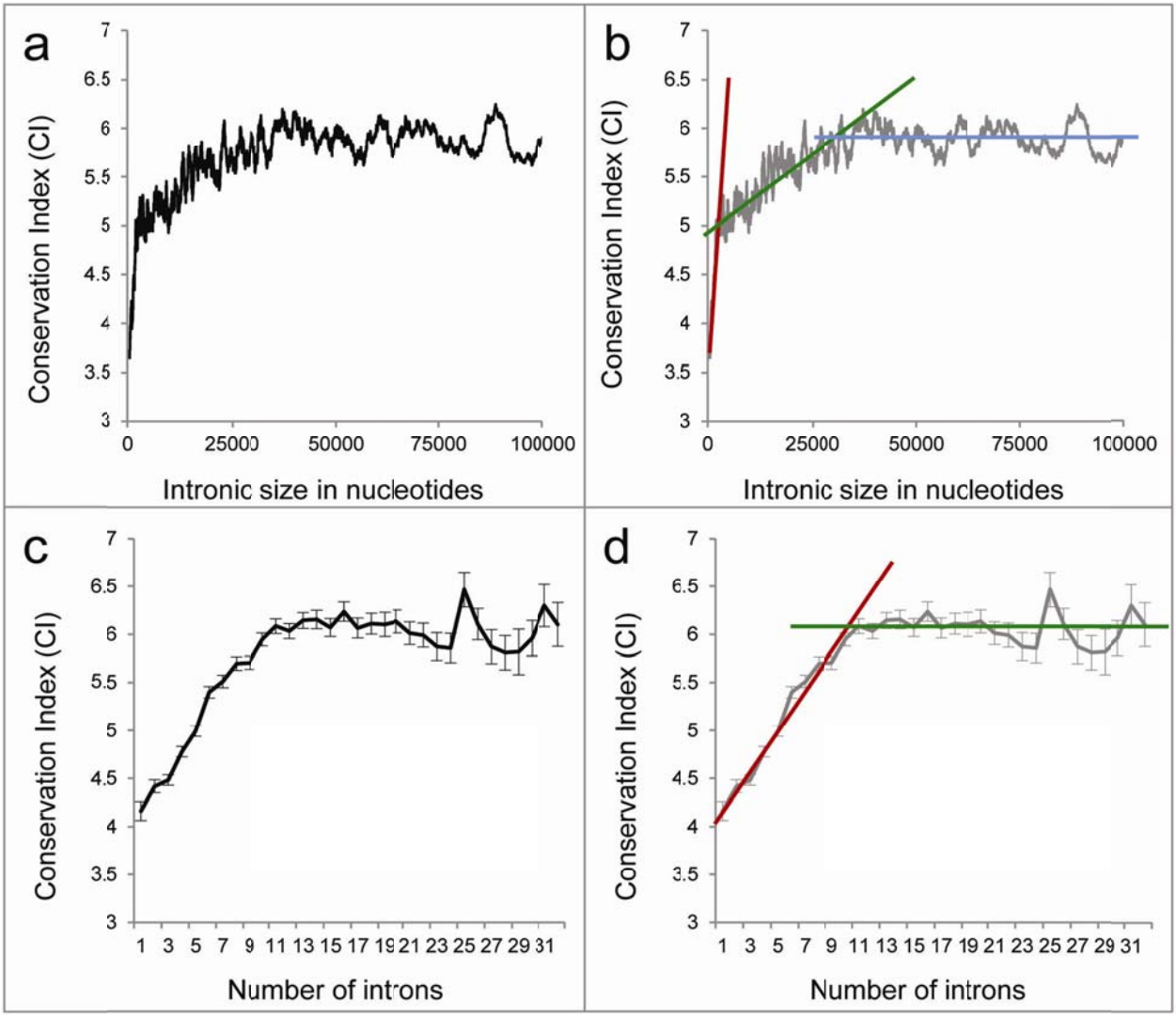

Figure 2. The dependence between conservation index (CI) and intronic size (figures a, b) and number of introns (figures c and d). Coloured lines mark different segments of the curves.

Intronic burden and the gene expression

Figure 3a shows dependence between the intronic size and average expression in 10 normal human tissues. The expression level of the genes with very small intronic size (less than 1kb) is relatively low and increases until the intronic size reaches ~5kb. For the genes with size of the intronic region of > 5kb there is a negative association between the intronic size and the expression level. A similar curve describes the relationship between the number of introns and

the mean expression level. Intronless genes and those with a single intron show relatively small average expression. Highest expression level was observed in the group of genes with 3 introns. Starting from this point the average expression decreases as the number of introns increases.

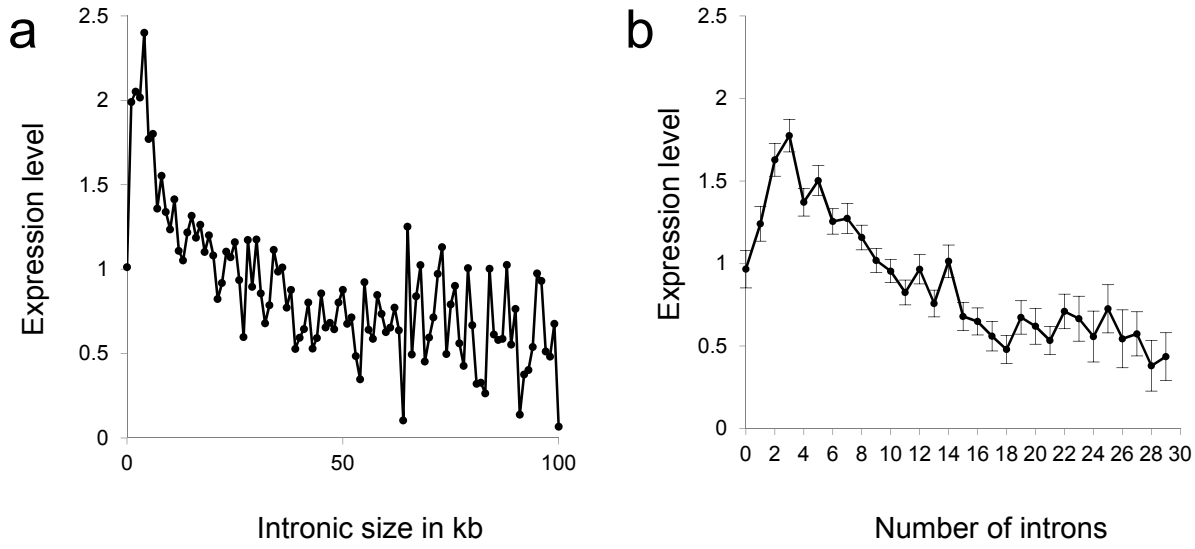

Figure 3. Dependence of the average gene expression in diverse normal human tissues on the intronic size (left panel) and the number of introns (right panel). Vertical lines on the right panel are standard errors of mean. Expression level is shown in Z-scores: see [14] for details.

By using multivariable linear regression model we found that both the intronic size and the number of introns are independent significant predictors of expression: F-statistics for number of introns was 92.1, $P < 10^{-6}$, and for the size of intronic region: F = 44.6, P = 0.0001.

The results of this analysis demonstrate that both the number of introns and the intronic size positively correlate with the level of evolutionary conservation of the corresponding protein. It is difficult to come up with any biological process that could provide a mechanistic link between the size of intronic region (or number of introns) and CI. Our working hypothesis is that

there is a biological cost associated with the size of intronic sequence and the number of introns and only functionally important genes can be "allowed" to have a large number of introns (or a large intronic size) without the cost of having large introns becoming too high. Indeed, transcription of large introns is associated with large energy expenditures [16] and the splicing out of the large number of introns also requires energy and may be associated with splicing errors [17]. The negative correlation between gene expression and the gene's intronic load supports the idea of the biological cost of introns.

The cost-benefit hypothesis of the association between the size/number of introns and the level of evolutionary conservation of the protein predicts that genes with large intronic burden should be more functionally important compared to the genes with small intronic burden. Unfortunately the level of evolutionary conservation reflects not only functional significance of the gene, but is also strongly influenced by evolutionary history, which is mostly unknown. To estimate the level of functional importance one can use data on the density of functional polymorphisms. Genes of low functional importance are expected to accumulate functional (protein-changing) polymorphisms more easily than functionally important genes. Using dbSNP data we found that the number of synonymous SNPs per codon does not correlate with CI: $r = 0.01$, $N = 16,194$, $P = 0.18$, while the number of non-synonymous SNPs and stop-gained SNPs negatively correlate with CI: $r = -0.1$, $n = 16,194$, $P < 10^{-6}$ and $r = -0.06$, $n = 16,194$, $P < 10^{-6}$ correspondingly which is consistent with the idea that density of potentially functional polymorphisms can be used as a measure of functional importance of a gene. Further we have used the ratio of nonsynonymous to synonymous – nonSYN/SYN and the ratio of stop-gained to synonymous SG/SYN to assess the strength of purifying selection on the gene. Genes of high functional significance are expected to have small ratios, while the genes of low functional significance are expected to have high ratios. We noted a significant negative correlation between nonSYN/SYN and the number of introns: $r = -0.05$, $n = 3,363$, $P = 0.006$. The

correlation between number of introns and SG/SYN was also significant: r = -0.26, n = 3,363, P < 10^{-6}. The correlations of nonSYN/SYN with the size of intronic region was negative but did not reach statistical significance: r = -0.03, n = 3,363, P = 0.11, while the correlation between SG/SYN and intronic size was significant: r = -0.13, n = 3,363, P < 10^{-6}. In brief, these results support the idea that genes with a large intronic size and/or those with multiple introns tend to be more functionally loaded, which results in stronger purifying selection manifested in a low proportion of nonSYN and SG SNPs.

The results of the conducted analyses also suggest that genes with a small number of introns (0-2) and a small intronic size (<3 kb) may be different from other genes. These smallest genes tend to have lowest conservation index and low expression level. These genes also show a higher density of non-synonymous and stop-gained SNPs compared to other genes: corresponding t-tests, 1.9 (P = 0.04) and 2.02 (P = 0.02), while synonymous SNPs do not show such differences: t-test = 0.2, df = 16,192, P = 0.87. Based on these observations we hypothesize that smallest genes may be enriched by young genes that did not take any important biological function yet and as a result cannot accumulate multiple introns. The alternative explanation can be that smallest genes are a graveyard of the genome containing dying genes that for some reason, e.g. changing environment, became less functional, lost their exons and started to accumulate non-synonymous and stop-gained mutations.

REFERENCES:

1 Lynch M, Conery JS: The origins of genome complexity. *Science* 2003; **302**: 1401-1404.

2 Koonin EV, Csuros M, Rogozin IB: Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdiscip Rev RNA* 2013; **4**: 93-105.

3 Rogozin IB, Carmel L, Csuros M, Koonin EV: Origin and evolution of spliceosomal introns. *Biol Direct* 2012; **7**: 11.

4 Fedorova L, Fedorov A: Introns in gene evolution. *Genetica* 2003; **118**: 123-131.

5 Koralewski TE, Krutovsky KV: Evolution of exon-intron structure and alternative splicing. *PLoS One* 2011; **6**: e18055.
6 Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV: Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform* 2005; **6**: 118-134.
7 Roy M, Kim N, Xing Y, Lee C: The effect of intron length on exon creation ratios during the evolution of mammalian genomes. *Rna* 2008; **14**: 2261-2273.
8 Zhang C, Li WH, Krainer AR, Zhang MQ: RNA landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci U S A* 2008; **105**: 5797-5802.
9 Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: Selection for short introns in highly expressed genes. *Nat Genet* 2002; **31**: 415-418.
10 Roy SW, Gilbert W: The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* 2006; **7**: 211-221.
11 Domazet-Loso T, Tautz D: An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol* 2008; **25**: 2699-2707.
12 Sayers EW, Barrett T, Benson DA *et al*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2012; **40**: D13-25.
13 Gibney G, Baxevanis AD: Searching NCBI Databases Using Entrez. *Curr Protoc Hum Genet* 2011; **Chapter 6**: Unit6 10.
14 Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A: RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* 2012; **28**: 1184-1185.
15 Louhichi A, Fourati A, Rebai A: IGD: a resource for intronless genes in the human genome. *Gene* 2011; **488**: 35-40.
16 Lehninger AL, Nelson, D.L. & Cox, M.M.: Principles of Biochemistry. New York, Worth, 1982.
17 Hsu SN, Hertel KJ: Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA Biol* 2009; **6**: 526-530.