

Intron loss and gene size

Maria D. Vibranovski,¹ Sandro José de Souza²

Keywords: intron loss, reverse transcription, gene length

1 Introduction.

Intron distribution is biased towards the 5' end of genes, especially in intron-poor species and intron-poor genes [1, 2]. This observed pattern could be a result of 3'-biased intron loss, 5'-biased intron gain or selective pressure for intron retention at the 5' ends. One possible mechanism for 3' biased intron loss is an homologous recombination between the genomic copy of a gene and its cDNA, product of the reverse transcription of its (intronless) mRNA. As the reverse transcription starts at the 3' poly (A) tract and not always reaches the 5' end of the gene, this process would remove more frequently 3' introns [1]. Mourier and Jeffares [1] argued that loss of 3' introns by recombination with RT-mRNAs would be the most plausible mechanism that could have generated such a bias because the other mechanisms would not be expected to cause such an enhanced tendency in intron-poor species and genes. The intron loss by recombination with RT-mRNAs would present two different features besides the 5'-biased distribution of introns. One of them, the concerted loss of introns, was recently tested by Roy and Gilbert [3]. They showed that intron loss distribution is biased to the 3' end of genes and adjacent introns were lost more frequently than what would be expected by chance. The second one would be the gene-size dependent 3' intron loss [1]. As the process of reverse transcription is not always complete due to a probable limitation of the number of nucleotides reversely transcribed in each round, it is expected that longer genes would present a less 5' biased distribution of lost introns. Here, we tested this second feature.

2 Methods.

In order to select lost introns, we first downloaded all genes from seven different species (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Anopheles gambiae*, *Caenorhabditis elegans*, *C. briggsae* and *Drosophila melanogaster*) from Ensembl and Flybase and annotated all intron positions and their respective phases in respect to the amino acid sequence encoded by each gene. We aligned (Blast [4]) all human gene products against all protein sequences from all other species to select orthologous genes to generate clusters of genes. Then, we aligned all sequences from each cluster with ClustalW [5]. Cluster alignments with more than 60% of gaps and with less than 40% of conserved amino acids (within ungapped regions) were discarded. Introns were considered for the analysis if they lied in reliable regions, i.e., if the 10 surrounding positions of the alignment presented no gap and at least 5 identical or conserved amino acids. In order to consider sliding, intron positions distant from each other by 5 amino acids or less were accounted as just one. Lost introns were defined by those absent in a lineage, but present in a sister group and at least one outgroup species [4]. The size of genes was calculated based on the CDS length. All analyses were done also summing the 3' UTR length (when available in the databases) to the gene size length.

¹ Ludwig Institute of Cancer Research, Sao Paulo Branch, Sao Paulo, Brazil and PhD. Program, Departamento de Bioquímica, Universidade de São Paulo, Sao Paulo, Brazil E-mail: maria@compbio.ludwig.org.br

² Ludwig Institute of Cancer Research, Sao Paulo Branch, Sao Paulo, Brazil E-mail: sandro@compbio.ludwig.org.br

3 Results.

Apparently, the relative position of 1,094 lost introns is not dependent of the length of genes. However, when only genes longer than 1.1 thousand amino acids were analyzed, a significant association between percentual intron positions (p) and gene length was observed ($P = 0.039$; linear regression of $\arcsin \sqrt{p}$ on log-transformed size: $b = 0.56$; $r^2 = 0.045$; $n = 94$). Moreover, when we separated genes by their lengths, the mean relative position of introns was 49.6% (SE = 0.75, $n = 1,080$) for genes < 2,000 amino acids and 67.5% (SE = 7.46, $n = 14$) for genes $\geq 2,000$ amino acids. The association between size and percentual introns positions is still significant when 3' UTR length was included in the calculation of gene size ($P = 0.024$; linear regression: $b = 0.48$; $r^2 = 0.0399$; $n = 127$, for genes longer than 3.3 thousand bps). Of course, some lost introns were not included in this analysis because they have been already lost in the sister group or in the outgroup species. However, this result corroborates the other evidences recently observed that suggest that the major mechanism of intron loss is the recombination of the genome with its reverse transcribed sequences [1, 2, 3].

References

- [4] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
- [1] Mourier, T. and Jeffares, D.C. 2003. Eukaryotic intron loss. *Science* 300: 1393.
- [3] Roy, S.W. and Gilbert, W. 2004. The pattern of intron loss. *Proc. Natl. Acad. Sci. USA* 102: 713-718.
- [2] Sverdlov, A.V., Babenko, V.N., Rogozin, I.B. and Koonin, E.V. 2004. Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* 338: 85-91.
- [5] Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

[Supported by FAPESP]