

First exons and introns - a survey of GC content and gene structure in the human genome

Krishna R. Kalari^{1,2,3}, Melanie Casavant⁴, Thomas B. Bair^{1,2}, Henry L. Keen^{1,2,5}, Josep M. Comeron⁶, Thomas L. Casavant^{1,2,3,7} and Todd E. Scheetz^{1,2,3,8*}

The University of Iowa, Iowa City, Iowa - 52242, USA

¹ Center for Bioinformatics and Computational Biology

² Coordinated Laboratory for Computational Genomics

³ Department of Biomedical Engineering

⁴ Department of Biochemistry

⁵ Department of Internal Medicine

⁶ Department of Biological Sciences

⁷ Department of Electrical and Computer Engineering

⁸ Department of Ophthalmology

* Corresponding author

Email: todd-scheetz@uiowa.edu

Phone: +1-319-335 6054; Fax: +1-319-384 0944

Edited by E. Wingender; received January 26, 2006; revised April 06, 2006; accepted April 09, 2006; published April 14, 2006

Abstract

Most transcriptional regulatory elements are located in non-coding DNA. In particular, some first introns play a vital role in transcriptional control and splicing. The length and GC-content of first exons and introns in complex organisms suggests that these structural units are likely to be important functional elements in large genomes. Hence, in this paper we perform a systematic comparison of exon-intron structure and GC content on all known genes in the human genome. Our in-silico analysis found that the GC content of introns and exons varies significantly depending on their length. On average, the first intron of a gene is significantly longer than other introns in the same gene. Our results also show that first introns and exons are more GC rich than last and internal. This study provides insight into the structure of eukaryotic genes. These results confirm and expand the previously identified regulatory potential of first exons and introns.

Keywords: gene structure, GC-content, intron length

Introduction

The purpose of much of the non-coding DNA in eukaryotic genomes is not well understood and has been termed "junk DNA" by some investigators [1]. However, recent research indicates that much of this so-called junk DNA may be functional [2, 3]. Specifically, the portion of the gene structure near the start of transcription (promoter, first exon and first intron) is known to harbor many of the elements that regulate transcription [3].

The characteristics of exons and introns have been studied many times [4-14]. These studies include several that investigated intron-exon structures and intron-genome size relationships in various model organisms [9, 13]. However, such studies typically focused on a limited portion of a single genome [11] or on summary statistics (such as total exon length) [12, 14]. Chen et al. concluded that first introns are longer than other introns based on analysis of chromosomes 21 and 22 [11]. Other studies have suggested this might be due to the presence of elements regulating transcriptional initiation and gene processing [15]. Therefore, it is suspected that the lengths of introns are important and first introns play a vital role in transcription control.

To date, no systematic analysis on gene structure and composition has been performed across the entire human genome. This analysis examined the length and GC content of first, internal and last introns and exons separately for the entire human genome. This analysis provides new data that will aid in the elucidation of factors responsible for organization of the human genome and to enhance computational methods of gene structure prediction. It will also help in identifying those transcribed portions that are likely to harbor regulatory elements.

Methods

To avoid duplicate identification of exons and introns, this analysis focused on genes with a single transcript. Single transcript genes and gene structures were identified and retrieved from Ensembl (<http://www.ensembl.org>; build 34). Ensembl's Perl modules were used to obtain exon and intron lengths for all single transcript genes in the human genome. All introns were evaluated individually and required to be at least 20 bp long to ensure sufficient splicing signals were present [16] and to avoid systematic errors based upon automated annotation methods. To further increase our confidence in annotated gene structure, only introns with canonical splice donor and acceptor sites were utilized in the analysis. A local MySQL database was built to store all relevant information including the lengths of first, internal and last introns and exons for each gene. The sequences for each feature were retrieved using Ensembl's Perl modules and stored locally. Due to the large sample size and unknown distributions nonparametric tests were used to assess significance in the intron and exon lengths. The R statistical analysis package was used to perform the statistical analyses. Only p -values of less than 0.05 were considered significant.

To assess GC content, introns and exons were first partitioned into three categories: short, intermediate and long. For all classes evaluated (e.g., first introns, first exons, etc.) the categories

were defined such that the longest third of all elements in the class were classified as long. The same procedure was used to define the short categories, with the remaining elements classified as intermediate. Multiple thresholds were explored using this method, all of which yielded nearly identical results. Further partitioning was performed on all three categories of first exons to allow comparisons of the untranslated and coding portions separately.

Results and discussion

Previously published results provided a general view about intron-exon structures, but limitations in the datasets analyzed including quality, redundant transcripts, and lack of coverage of the entire genome make accurate interpretation of their findings difficult [4, 9, 11, 12, 13, 14, 21]. Although several resources are available that annotate and display genome-scale data [17, 18, 19], the Ensembl gene model was utilized in this analysis because of the high specificity of gene annotations [17]. Although Ensembl's gene models exhibit high specificity, the accuracy of *bona fide* first exon annotation is almost certainly not perfect. However, based upon a single method of first-exon identification developed by Davuluri *et al.* [20], they were able to identify first exons with 88% accuracy. Hence it is expected that Ensembl's gene model using multiple sources of evidence to evaluate first exons, can out perform this.

The results reported in this manuscript are based upon all single transcript human genes annotated in Ensembl's database. These analyses were also performed using the complete complement of genes (i.e., those with single and multiple transcripts) with similar results (data not shown). There are a total of 23,531 genes in the Ensembl database, of which 18,154 are comprised of a single transcript.

Intron lengths

To avoid ambiguity in assessing differences among first, internal and last introns in the study, only 9,499 single transcript genes with at least three introns were utilized. The length of introns at different positions (first, last) were compared to the median intron length for each gene. Significant results were found for the first intron, indicating that on average, first introns are longer than other introns. These results agree with previously published analyses focused on chromosomes 21 and 22 [11].

Tabulated results for each human chromosome are shown in [Tab. 1](#). Based upon the sign-test and Fisher's exact test, first introns were found to be longer than median length introns more often than expected at random ($p < 0.05$). The only exceptions were chromosomes 21 and Y. Although these two chromosomes have more long first introns than short first introns, the sample size is insufficient for statistical significance. Unlike the analysis of first introns, no statistically significant difference was found in the lengths of other intron classes (second, last; data not shown). The relative difference in length between first introns and the median intron length is 8.9. Our study also observed that intron length decreased with respect to their position in most genes (data not shown), in agreement with previously published results [14]. No specific reason

was identified that explained the increased length observed in first introns. In particular, the relative prevalence of repeated elements (LINES, SINES, etc.) was the same in all intron classes.

Table 1: Comparison of first intron length to median intron length.

| Chromosome | N (Genes) | First intron length < median intron length | First intron length = median intron length | First intron length > median intron length |
|--------------|-------------|--|--|--|
| Chr1 | 979 | 268 | 0 | 711 |
| Chr2 | 647 | 201 | 0 | 446 |
| Chr3 | 511 | 157 | 0 | 354 |
| Chr4 | 404 | 110 | 0 | 294 |
| Chr5 | 444 | 149 | 0 | 295 |
| Chr6 | 469 | 126 | 0 | 343 |
| Chr7 | 426 | 124 | 1 | 301 |
| Chr8 | 332 | 97 | 0 | 235 |
| Chr9 | 345 | 100 | 1 | 244 |
| Chr10 | 343 | 108 | 0 | 235 |
| Chr11 | 595 | 168 | 1 | 426 |
| Chr12 | 529 | 137 | 0 | 392 |
| Chr13 | 137 | 48 | 0 | 89 |
| Chr14 | 268 | 74 | 0 | 194 |
| Chr15 | 332 | 95 | 0 | 237 |
| Chr16 | 433 | 143 | 0 | 290 |
| Chr17 | 588 | 163 | 0 | 425 |
| Chr18 | 137 | 39 | 0 | 98 |
| Chr19 | 651 | 210 | 0 | 441 |
| Chr20 | 259 | 66 | 1 | 192 |
| Chr21 | 94 | 36 | 0 | 58 |
| Chr22 | 191 | 56 | 0 | 135 |
| ChrX | 342 | 101 | 0 | 241 |
| ChrY | 43 | 17 | 0 | 26 |
| Total | 9499 | 2793 | 4 | 6702 |

The third, fourth and fifth columns present the number of genes observed with first intron length shorter than, equal to or longer than the median intron length.

However, we hypothesize that long first introns facilitate increased regulatory complexity. Specifically, we present two separate rationales for how long introns may increase regulatory complexity. The first hypothesis is that longer first introns can harbor (and evolve) additional transcription factor binding sites. This hypothesis is supported by the identification of regulatory elements in the first intron [15]. The second rationale for increased regulatory complexity is that longer first introns present the ability for development of new transcription initiation sites - additional promoters and first exons that could be utilized under different conditions. Such new promoters could then add further levels of regulatory control. This rationale matches well with the evolutionary advantage of introns in general.

Exon lengths

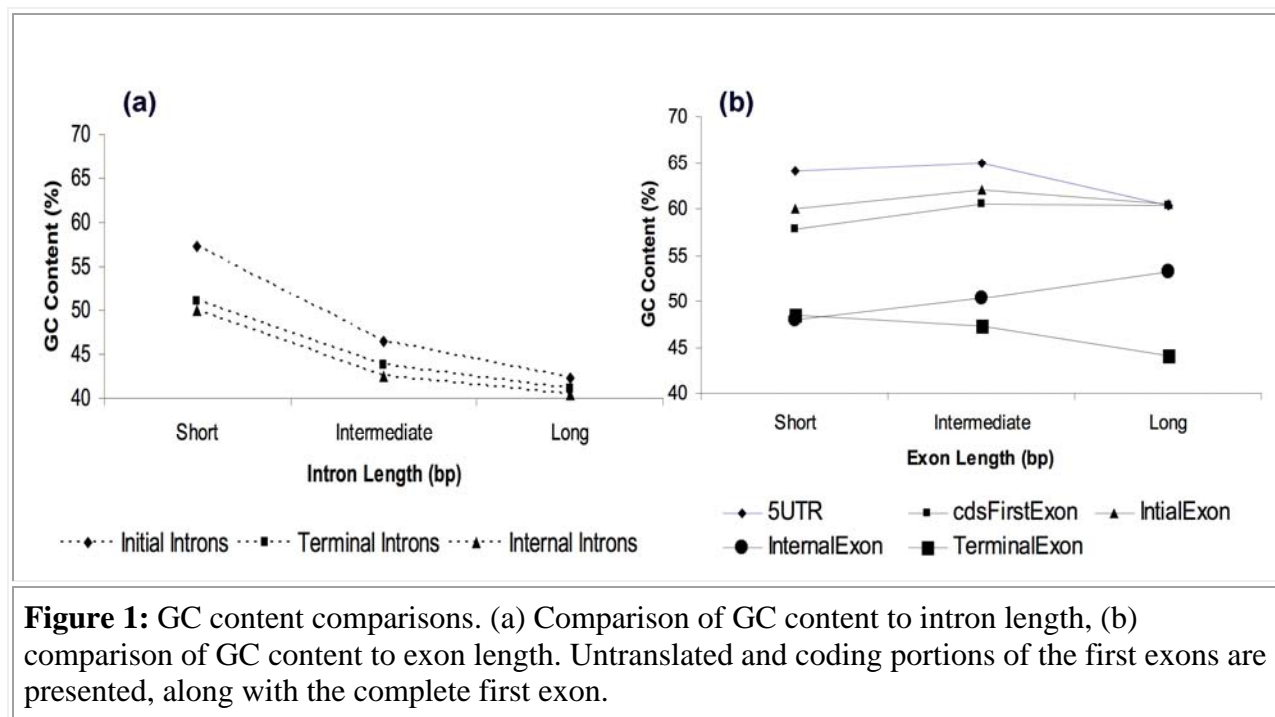
A similar analysis was performed looking for length bias in first and last exons. To avoid ambiguity in classification only the 12,224 single-transcript genes with at least 3 exons were used in this analysis. Both the first and last exon length was compared to the median exon length. No significant bias in first exon length was observed. However, statistically significant results were observed for a difference in exon length when comparing last exons to the median exons. This is likely related to the presence of long 3' UTR sequences.

Relationship between first, internal and last introns and GC content

GC content was classified with respect to intron length by partitioning introns into categories by length: short, intermediate and long. Virtually no difference was observed between internal and last introns. Long introns were observed to contain lower GC content across first, internal and last introns, and vice-versa (Fig. 1a). In fact, internal and last introns exhibited nearly identical GC content across all three length categories. However, first introns tended to have a higher GC content - particularly in short first introns.

Assessment of GC content in first, internal and last exons

Fig. 1b shows the GC content of first, internal and last exons with respect to exon length. First exons were also sub-classified into non-coding (5' UTR) and coding (CDS) regions. Exons were divided into short, intermediate and long exons depending on their lengths as described in the methods.



Higher GC content was observed in the first exons than in either the internal or last exons. This is in agreement with a previous study by Majewski and Ott in which they observed an overrepresentation of the CpG dinucleotide in both the promoter and first exon [15]. Even more striking, the average GC content in the 5' untranslated region (UTR) of first exons averaged over 60% (significantly higher than the coding portion of the first exons, or indeed of the GC content of internal and last exons) as indicated in Fig. 1b. Although the last and internal exons exhibit similar GC content, they displayed opposing trends with GC content increasing with exon length for internal exons, but decreasing with exon length for last exons. The last exon result is not surprising, as long last exons are likely to be dominated by AT rich 3' UTR. However, the significance of increased GC in long internal exons is unknown.

Conclusion

The results presented confirm earlier indications that gene structures near the start of transcription are compositionally distinct. Specifically, that first introns are typically longer than other introns and that first exons - particularly the non-coding portions - are more GC rich than other exons. These results offer intriguing hints into the regulatory potential contained in first exons and introns in addition to the canonical promoter.

Acknowledgements

TES was supported with a Career Development Award from Research to Prevent Blindness. TBB was supported by an NRSA post-doctoral fellowship (1F32HG002881).

References

1. Ohno, S. (1972). So much "junk" DNA in our genome. *In: In Evolution of Genetic Systems*, Smith, H. H. (ed.), Gordon and Breach, New York, pp. 366-370.
2. Comeron, J. M. (2001). What controls the length of noncoding DNA? *Curr. Opin. Genet. Dev.* 11, 652-659.
3. Klamut, H. J., Bosnoyan-Collins, L. O., Worton, R. G., Ray, P. N. and Davis, H. L. (1996). Identification of a transcriptional enhancer within muscle intron 1 of the human dystrophin gene. *Hum. Mol. Genet.* 5, 1599-1606.
4. Hawkins, J. D. (1988). A survey on intron and exon lengths. *Nucleic Acids Res.* 16, 9893-9908.
5. Dorit, R. L., Schoenbach, L. and Gilbert, W. (1990). How big is the universe of exons? *Science* 250, 1377-1382.
6. Palmer, J. D. and Logsdon, J. M., jr. (1991). The recent origins of introns. *Curr. Opin. Genet. Dev.* 1, 470-477.
7. Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. and Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20, 4255-4262.
8. Oliver, J. L. and Marin, A. (1996). A relationship between GC content and coding-sequence length. *J. Mol. Evol.* 43, 216-223.
9. Deutsch, M. and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 27, 3219-3228.
10. Kriventseva, E. V. and Gelfand, M. S. (1999). Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J. Biomol. Struct. Dyn.* 17, 281-288.

11. Chen, C., Gentles, A. J., Jurka, J. and Karlin, S. (2002). Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99, 2930-2935.
12. Sakharkar, M. K., Chow, V. T. and Kanguene, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biol.* 4, 387-393.
13. Vinogradov, A. E. (1999). Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* 49, 376-384.
14. Comeron, J. M. (2004). Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* 167, 1293-1304.
15. Majewski, J. and Ott, J. (2002). Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12, 1827-1836.
16. Bell, M. V., Cowper, A. E., Lefranc, M. P., Bell, J. I. and Sreaton, G. R. (1998). Influence of intron length on alternative splicing of CD44. *Mol. Cell. Biol.* 18, 5930-5941.
17. Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., Curwen, V., Cutts, T., Down, T., Durbin, R., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Kahari, A., Jekosch, K., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., Mcvicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, C., Clamp, M. and Hubbard, T. (2004). Ensembl 2004. *Nucleic Acids Res.* 32, D468-D470.
18. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32, D493-D496.
19. Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A. and Wagner, L. (2004). Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* 32, D35-D40.
20. Davuluri, R. V., Grosse, I. and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nat. Genet.* 29, 412-417.
21. Sakharkar, M. K., Perumal, B. S., Sakharkar, K. R. and Kanguene, P. (2005). An analysis on gene architecture in human and mouse genomes. *In Silico Biol.* 5, 1814-1817.