In Silico Biology 4 (2004) 387–393
IOS Press

Short Communication

Distributions of Exons and Introns in the Human Genome

Meena Kishore Sakharkar[a,*], Vincent T.K. Chow[b] and Pandjassarame Kangueane[a]
[a]*Nanyang Centre for Supercomputing and Visualisation, N3-2c-113b, 50 Nanyang Avenue, Nanyang technological University, Singapore*
Tel.: +65 6 790 5836; Fax: +65 6 791 1859
[b]*Human Genome Laboratory, Department of Microbiology, Faculty of Medicine, National University of Singapore, Kent Ridge, Singapore*
E-mail: mmeena@ntu.edu.sg

Edited by E. Wingender; received 13 April 2004; revised and accepted 27 May 2004; published 16 June 2004

ABSTRACT: The human genome is revisited using exon and intron distribution profiles. The 26,564 annotated genes in the human genome (build October, 2003) contain 233,785 exons and 207,344 introns. On average, there are 8.8 exons and 7.8 introns per gene. About 80% of the exons on each chromosome are $<$ 200 bp in length. $<$ 0.01% of the introns are $<$ 20 bp in length and $<$ 10% of introns are more than 11,000 bp in length. These results suggest constraints on the splicing machinery to splice out very long or very short introns and provide insight to optimal intron length selection. Interestingly, the total length in introns and intergenic DNA on each chromosome is significantly correlated to the determined chromosome size with a coefficient of correlation $r = 0.95$ and $r = 0.97$, respectively. These results suggest their implication in genome design.

KEYWORDS: Exon, intron, length, distributions, human, genome, architecture, profile, chromosome, correlation, size, non-coding DNA, gene, average, genomics, gene evolution, genome evolution, DNA, gene structure

INTRODUCTION

The availability of complete genome sequence of many eukaryotic organisms continues to contribute towards better understanding of their genome design and evolution. An average vertebrate gene consists of multiple small exons separated by introns that are 10 or 100 times longer [Hawkins, 1988]. In order to understand the structure and evolution of eukaryotic genomes, it is important to know the general statistical characteristics of the exons and introns. Many authors have published the analysis of some characteristics of nuclear introns [Dorit et al., 1990; Palmer *et al*., 1991; Mount *et al*.,1992; Fedorov *et al*., 1992]. Deutsch *et al*. reported intron-exon structures from eukaryotic model organisms and analysed the statistical distribution of spliceosomal introns (splicing of these introns requires the participation of a specific set of protein-RNA particles) and exons of nuclear genes in 10 model organisms from GenBank

*Corresponding author.

Electronic publication can be found in *In Silico Biol.* **4**, 0032 <http://www.bioinfo.de/isb/2004/04/0032/>, 16 June 2004.

[Deutsch and Long, 1999]. The analysis provides a general picture of intron-exon structure of eukaryotic genes. The data though valuable and informative, has caveats associated with the source, redundancy and quality of GenBank data and are not representative of the genome as a whole. The availability of complete genome sequence of many eukaryotes provides a podium for understanding the distributions of introns and exons at genome level. This provides insight to their role in shaping and structuring of the genome. In this report we provide a detailed analysis on exon and intron distributions in the human genome [Venter *et al*., 2001; Lander *et al*., 2001]. Using genome data for exon-intron distributions circumvents the errors due to sampling bias and redundancy during purging and allows for intron-exon distribution studies in a concerted manner.

Here, we examine the distribution of genes, exons and introns on the 24 human chromosomes and discern correlations between them. This analysis is fundamental for a quantitative view of human genome organization. These findings could help improve gene structure prediction by computational methods by providing better understanding of factors that govern genome design and architecture.

MATERIALS AND METHODS

The Human genome data was downloaded from the National Center for Biotechnology Information (NCBI) (Oct 2003, build) at ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/. The data was processed for extraction of exons and introns based on the CDS feature table annotation [Sakharkar *et al*., 2002]. Starting with 26,564 genes, we filtered out 233,785 exons and 207,344 introns from the human genome. The results of exon (exons in the coding region) and intron (introns between the coding exons) distributions were tabulated for further analysis.

Results and discussion

It is well known that human chromosomes are very different among themselves [Venter *et al*., 2001]. Putting aside the obvious differences in size, there are also divergences in the density and spatial location of genes, the types of genes, organization of Alu repeats [Grover *et al*., 2003] and the distribution of CpG islands [Chen *et al*., 2002]. This fact suggests a unique mechanism of structural and architectural evolution of the human genome. We revisited the human genome using exon and intron distribution profiles and studied correlations among them. The results of our observations are summarized below.

Gene Distributions and Chromosome size

The total determined chromosome size (genome size) is 3,017,700,646 basepair (bp). The distributions of genes on different chromosomes based on CDS feature are shown in Table 1. The smallest chromosome is Y with 98 annotated genes. The largest chromosome is chromosome 1 with 2,514 genes. The number of genes on each chromosome is marginally correlated to chromosome size ($r = 0.73$). This weak correlation may suggest a limited causal relationship between number of genes and chromosome size. It also suggests that other factors besides number of genes also affect chromosome size. The longest annotated gene is DMD (Dystrophin Dp140bc isoform) 2,217,347 bp (79 exons) found on chromosome X [Nishio *et al*. 1994].

Table 1

Exon – intron distributions for human genome

Chr #	Total # genes	Total # exons	Total # introns	Max # exons/ gene	Chromosome size (determined)	Avg # of exons/ gene	Avg length (bp)		Std dev.		Total length (bp)		Shortest (bp)			Longest (bp)		
							exon	intron	exon	intron	exon	intron	exon	intron	gene	exon	intron	gene
1	2514	22345	19831	107	226828929	8.89	167.01	4736.52	229.37	14268.19	3731870	93929919	2	1	78	8449	476158	980961
2	1354	12506	11152	148	238349289	9.24	163.98	5883.23	226.88	17012.24	2050855	65609873	2	1	90	7572	483412	1897544
3	1394	13517	12123	118	195073306	9.70	164.06	6375.63	224.21	21019.22	2217700	77291760	2	1	150	6654	497816	990999
4	926	8299	7373	85	187239983	8.96	174.78	7168.94	266.64	19497.08	1450541	52856617	2	53	132	6255	494708	1467842
5	1186	9946	8760	90	177696509	8.39	189.50	7277.28	332.86	21277.20	1884777	63748970	2	1	150	6574	370360	930401
6	1306	11406	10100	145	169212327	8.73	173.62	5961.61	253.56	18967.75	1980397	60212251	2	31	159	7152	469892	1377570
7	2508	23045	20537	82	310210944	9.19	167.87	6703.87	271.88	20177.41	3868769	137677396	2	1	14	11923	458139	1641567
8	908	7823	6915	86	143297300	8.62	171.16	7354.15	258.43	21384.09	1339052	50853964	2	54	84	7308	453268	2055833
9	1033	8941	7908	72	117790386	8.66	170.66	5351.68	253.19	14121.26	1525926	42321074	2	33	105	6598	276306	865661
10	1017	10273	9256	69	132016990	10.10	153.79	6412.91	219.97	20271.48	1579898	59357955	2	52	105	7812	482575	1727184
11	1567	12459	10892	87	130908954	7.95	177.66	4341.42	237.03	15362.46	2213526	47286795	3	1	87	6183	437543	1463302
12	1299	12399	11100	89	129826379	9.55	158.07	4570.21	192.23	12979.23	1959945	50729293	2	30	81	6324	328545	1248678
13	426	3784	3358	83	95749578	8.88	183.47	7351.75	396.79	19082.4	694268	24687182	2	37	279	11555	317646	1175762
14	854	6837	6106	114	87191216	8.01	176.24	5653.70	276.66	19076.38	1204982	33826109	2	51	51	11304	479079	1210740
15	843	8106	7263	104	81992482	9.62	169.79	4660.70	271.38	11542.05	1376321	33850721	2	1	168	9527	207178	620362
16	1093	9986	8893	62	79932432	9.14	166.96	3661.25	242.60	13092.99	1667340	32559472	2	1	75	8607	466049	1167938
17	1459	13179	11720	74	79376966	9.03	165.08	3193.16	215.89	9875.72	2175698	37423835	2	30	63	4786	283762	712668
18	367	3333	2966	75	74658403	9.08	174.9	7905.40	256.53	19377.24	583054	23447419	3	67	225	4721	411175	1189866
19	1609	12169	10560	106	55878340	7.56	187.31	2032.87	279.92	4741.54	2279436	21467122	2	1	81	5059	170796	298909
20	775	6492	5717	80	59424990	8.38	160.34	4403.10	215.29	13613.39	1040952	25172558	3	54	135	3738	303713	1108855
21	309	2539	2230	47	33924367	8.22	168.59	5086.89	306.51	16098.67	428056	11343761	3	74	102	5916	323563	833627
22	671	5173	4502	54	34352072	7.71	171.14	3924.83	281.85	12999.39	885356	17669584	3	42	38	6762	447252	492969
X	1048	8568	7520	79	152118949	8.18	185.33	7627.85	299.66	23527.35	1587926	57361443	2	54	129	6102	493512	2217347
Y	98	660	562	44	24649555	6.73	173.74	5288.54	255.05	19676.46	114670	2972162	3	67	228	2493	400349	681119

Gene Distributions and Chromosome size

The average number of exons in human genes is about 8–10 and the mean value of 8.8 exons per gene. Exon lengths are distributed much more tightly (S.D. = 192.23 – 396.79) than introns on each chromosome (Table 1). The average exon length is about 170 bp. About 80–85% exons on each chromosome were found to be less than 200 bp in length. It is well established that most protein coding sequences are strongly constrained that is, they are under high selection pressures and most amino acid altering mutations are deleterious and become selectively eliminated. This is consistent with previous observations.

Conversely, the average intron size is about 5419 bp. However, the standard deviation (S.D.) about the mean intron size on 24 chromosomes is in the range of 4741.54 – 23527.35 (Table 1). The greater standard deviations about the mean intron length suggests for their being under lesser selection pressures resulting in the tendency of large-scale changes which is reflected in their length distributions (Table 1). It is interesting to see that though, an intron can be thousands of base pairs in size (Table 1), very large introns make up only a small proportion of total introns in the genome. About 5.24% of introns are more than 200,000 bp and less than 10% of introns are more than 11,000 bp in length. Also, $<$ 0.01% of the introns are $<$ 20 bp in length. These results suggest constraints on the splicing machinery to splice out very long or very short introns. It is remarkable to see that though chromosome 1 is the largest chromosome neither the gene with the maximum number of exons nor the gene with the longest intron or the longest gene reside on chromosome 1. An average human gene contains about 6–9 introns. The average number of introns per gene is 7.8. This number is considerably variable with ranges from 0 in about 3,362 genes (Single exonic genes) to 147 introns in NEB (Nebulin) on chromosome 2.

Correlations between chromosome size and total length in exons, introns

The total length in exons is 39,841,315 bp and that in introns is 1,123,657,235 bp. A moderate correlation of $r = 0.77$ is observed for total length in exons (bp) and chromosome size (Figure 1(a)). This is very similar to the correlation ($r = 0.73$) for genes and chromosome size. Since the average number of exons is more or less same for all chromosomes, this suggests higher number of genes on larger chromosomes. This hints that there are other factors that determine chromosome size and architecture. This probed us to explore the possibility of correlations between non-coding DNA (introns and intergenic DNA) and chromosome size. A very strong positive correlation is observed ($r = 0.95$) between total length in introns (bp) and chromosome size (bp) (Fig. 1(b)). A similar positive correlation ($r = 0.97$) is also observed between intergenic DNA and chromosome size (intergenic DNA = determined chromosome size – (length in exons + length in introns)) (Fig. 1(c)). This suggests that for larger chromosomes more regions are covered in introns and intergenic DNA. These observations indicate on the important role of introns and intergenic DNA in chromatin structure and chromosome architecture (since introns and integenic DNA account for major component of the determined chromosome size [Venter *et al*., 2001; Lander *et al*., 2001]). Lengyel and Penman showed that the size of hnRNA (heterogeneous nuclear RNA), but not mature mRNA, increases with genome size in dipterans. This observation, dated before the discovery of the intervening sequences or introns in 1977, was the first indication of a positive relationship between genome size and total intron length [Lengyel and Penman, 1975]. A significant, although weak, positive relationship between intron and genome size has now been established for many eukaryotes [Hughes and Hughes, 1995; Moriyama *et al*., 1998; Deutsch and Long, 1999; Vinogradov, 1999]. In all cases, however, the differences in intron size alone cannot fully account for the differences in euchromatic genome size, indicating that a single class of non-coding

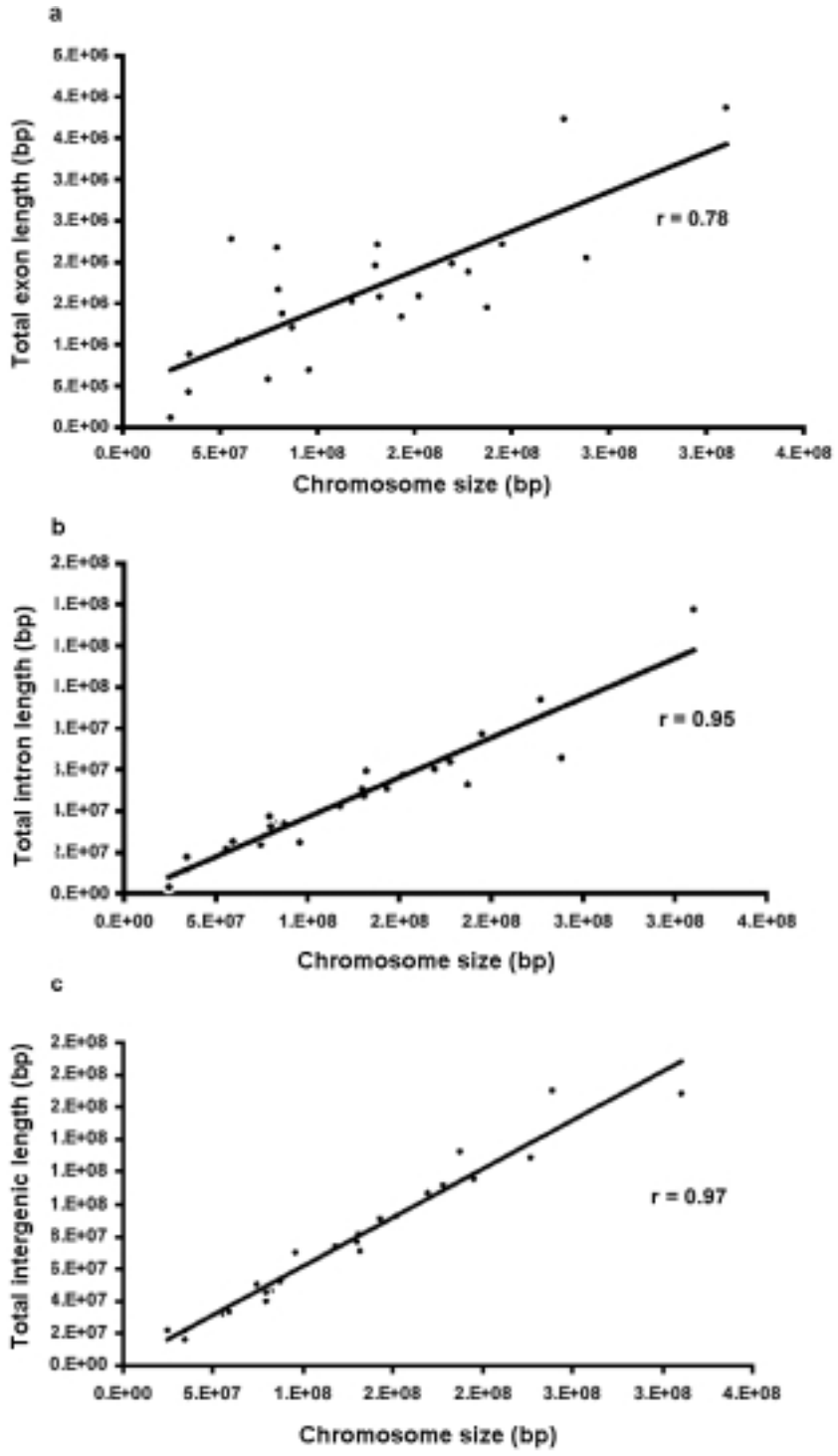

Fig. 1. Correlation between (a) total exon length, (b) total intron length, and (c) total intergenic length in bp and determined chromosome size.

DNA does not easily explain the differences in genome size. Our results suggest that variation in genome size among organisms is usually associated to congruent changes across different classes of non-coding DNA (e.g. introns and intergenic regions) uniformly across the genome. Recently, Morey and colleagues argued for the role of non-coding RNAs in epigenetic regulation [Morey and Avener, 2004]. Therefore, understanding the functions of these so called "non-coding sequences" in addition to the proteins themselves will be vital to understanding the genetics, biology and evolution of humans.

However, the numbers and the analysis need to be taken with caution because they are based on the genome annotations that sometimes are not very precise [Zhang, 2002].

CAVEATS

It must be noted that the traditional gene finding algorithms treat the translation start site as the 5' boundary of the gene and there are currently no computational tools to predict the non coding first exons or non coding portions of the first exon except where the true full-length mRNA sequences are available [Galas, 2001; Stormo, 2000; Davuluri *et al*., 2001]. As this analysis is strictly based on CDS feature in genome data, it does not take into account the first exon and is biased towards internal coding exons of the gene. Nonetheless, this analysis hints at the possible role of non-coding DNA in genome architecture and design and provides a platform for understanding the human genome and issues in gene evolution.

ACKNOWLEDGMENTS

This work is supported by A*STAR-BMRC, Singapore, Grant # 03/1/22/19/242.

REFERENCES

- 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.
- 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.
- Deutsch, M. and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. **27**, 3219-3228.
- Dorit, R. L., Schoenbach, L. and Gilbert, W. (1990). How big is the universe of exons? Science **250**, 1377-1382.
- Fedorov, A., Suboch, G., Bujakov, M. and Fedorova, L. (1992). Analysis of nonuniformity in intron phase distribution. Nucleic Acids Res. **20**, 2553-2557.
- Galas, D. J. (2001) Sequence interpretation. Making sense of the sequence. Science **291**, 1257-1260.
- Grover, D., Majumder, P. P., Rao, C. B., Brahmachari, S. K. and Mukerji, M. (2003). Nonrandom distribution of Alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. Mol. Biol. Evol. **20**, 1420-1424.
- Hawkins, J. D. (1988). A survey on intron and exon lengths. Nucleic Acids Res. **16**, 9893-9906.
- Hughes, A. L. and Hughes, M. K. (1995). Small genomes for better flyers. Nature **377**, 391.
- Koenig, M., Monaco, A. P. and Kunkel, L. M. (1988). The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. Cell **53**, 219-228.
- Lander, E. S., *et al*.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature **409**, 860-921.
- Lengyel, J. and Penman, S. (1975). hnRNA size and processing as related to different DNA content in two dipterans: Drosophila and Aedes. Cell **5**, 281-290.
- Moriyama, E. N., Petrov, D. A. and Hartl, D. L. (1998). Genome size and intron size in Drosophila. Mol. Biol. Evol. **15**, 770-773.

- Morey, C. and Avner, P. (2004). Employment opportunities for non-coding RNAs. FEBS Lett. **567**, 27-34.
- 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.
- Nishio, H., Takeshima, Y., Narita, N., Yanagawa, H., Suzuki, Y., Ishikawa, Y., Ishikawa, Y., Minami, R., Nakamura, H. and Matsuo, M. (1994) Identification of a novel first exon in the human dystrophin gene and of a new promoter located more than 500 kb upstream of the nearest known promoter. J. Clin. Invest. **94**, 1037-1042.
- Palmer, J. D. and Logsdon, J. M., jr. (1991). The recent origins of introns. Curr. Opin. Genet. Dev. **1**, 470-477.
- Sakharkar, M., Passetti, F., de Souza, J. E., Long, M. and de Souza, S. J. (2002). ExInt: an Exon Intron Database. Nucleic Acids Res. **30**, 191-194.
- Stormo, G. D. (2000). Gene-finding approaches for eukaryotes. Genome Res. **10**, 394-397.
- Venter, J. C., *et al*. (2001). The sequence of the human genome. Science **291**, 1304-1351.
- Vinogradov, A. E. (1999). Intron-Genome size relationship on a large evolutionary scale. J. Mol. Evol. **49**, 376-384.
- Zhang, M. Q. (2002). Computational prediction of eukaryotic protein coding genes. Nat. Rev. Genet. **3**, 698-709.