Available online at www.sciencedirect.com
ScienceDirect
Gene 410 (2008) 241–248
GENE
www.elsevier.com/locate/gene

Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes

Maria Costantini, Giorgio Bernardi *

Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

Received 12 July 2007; received in revised form 13 November 2007; accepted 5 December 2007
Available online 27 December 2007
Received by M. Di Giulio

Abstract

Many years ago compositional correlations were found to hold between coding and contiguous non-coding sequences. These correlations were essentially studied in whole genomes of mammals, which are characterized by strong compositional heterogeneities. Here we investigated whether these correlations also hold within the much more homogeneous isochore families. This point was checked not only in the case of mammals, but also in that of phylogenetically distant vertebrates, which are characterized by very different compositional patterns. Indeed, these are remarkably different in cold- and warm-blooded vertebrates. Fish genomes, for instance, are much more homogeneous than those of mammals and birds. The compositional correlations between coding sequences and the corresponding introns, or their 5′ and 3′ flanking regions, were studied in the isochore families of the fully sequenced genomes from four fishes (*Brachydanio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus* and *Tetraodon nigroviridis*), human and chicken.

Keywords: Isochore families; Fishes; Chicken; Human

1. Introduction

Thirty years ago, hybridization of appropriate probes allowed us to localize coding sequences in compositional fractions from mouse, rabbit, chicken and human genomes (Cuny et al., 1978; Cortadas et al., 1979). This provided the first indication (Bernardi, 1979, 1984) that the composition of coding sequences and long surrounding sequences were correlated. When a number of coding sequences were localized and their GC levels were plotted against the GC levels of the large DNA fragments that harbored them, linear correlations were found (Bernardi et al., 1985; Bernardi and Bernardi, 1986). This was also the case when plotting GC values (or GC_3 values, the GC levels of third codon positions) of exons against GC values of introns, or of the genome sequences in which the genes were embedded (Clay et al., 1996).

These findings were of great relevance because they showed the existence of compositional correlations between coding sequences, which only represent 1–2% of the genomes investigated and contiguous non-coding sequences (introns and intergenic sequences, which represent 98–99% of the genomes). These correlations established the view of the genome as an integrated ensemble, which was under compositional constraints due, in our opinion, to natural selection (Bernardi and Bernardi, 1986). At the same time, we found that the distribution of genes in the genomes of warm-blooded vertebrates was strongly biased towards GC-rich isochores (Bernardi et al., 1985; Mouchiroud et al., 1991; Zoubak et al., 1996). These observations rejected the widely accepted view of genes being distributed at random in non-coding "junk DNA" (Ohno, 1972).

All the results mentioned so far essentially concerned whole genomes from warm-blooded vertebrates, which are characterized by a remarkable compositional heterogeneity (Filipski

Abbreviations: GC, molar fraction of guanine and cytosine in DNA; GC_1, GC_2, GC_3, GC levels of first, second and third codon positions; kb, kilobases.

* Corresponding author. Tel.: +39 081 583 3215; fax: +39 081 245 5807.
E-mail address: bernardi@szn.it (G. Bernardi).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.12.016

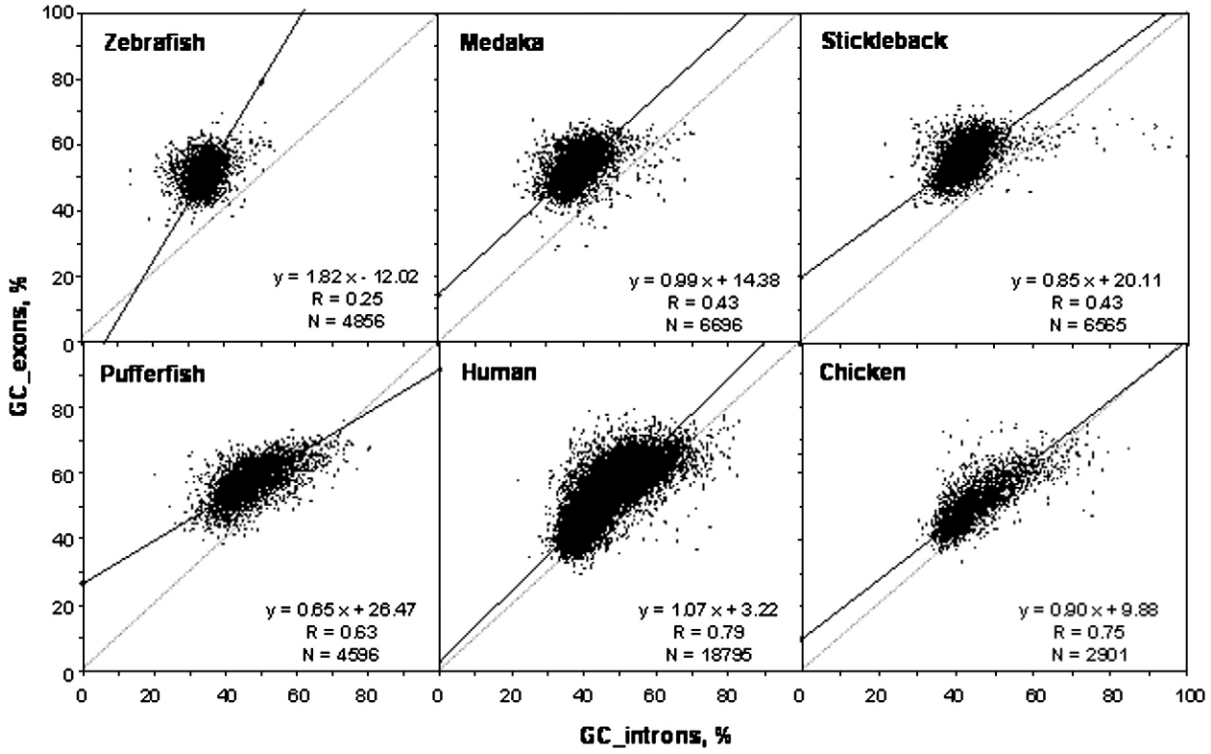

Fig. 1. Scatterplots of exon GC versus intron GC in the genomes analyzed. The orthogonal regression equations, the correlation coefficients (*R*) and the number of genes (*N*) are reported. The main diagonal is indicated by a broken line.

et al., 1973; Thiery et al., 1976; see also Bernardi 2004, 2007, for reviews), a factor which might influence the correlations.

In the present investigation we studied the compositional correlations between coding and contiguous non-coding sequences in individual human and chicken isochore families (Costantini et al., 2006, 2007a) and in the corresponding families from the much less heterogeneous genomes of fishes. Indeed, recent results (Costantini et al., 2007b) showed that fish genomes consist of one major isochore family (L1, L2 and H1 for zebrafish, medaka and stickleback, respectively) and a minor family (L2, H1 and H2, respectively); in the case of pufferfish, H1 and H2 isochore families are present in comparable amounts. Since the corresponding isochore families from the genomes of both fishes and warm-blooded vertebrates are compositionally very similar (this being the reason to give them the same names), it was of interest to see whether the compositional correlations were or were not the same, independently of the phylogenetic origin of isochore families.

2. Materials and methods

2.1. Analysis of database sequences

The zebrafish genes were retrieved from Hovergen (Release 47, July 2005), the genes from medaka (Release 41.1), stickleback (Release 41.1a) and pufferfish (Release 41.1 g), from Ensembl (http://www.ensembl.org/index.html), the human genes from GenBank. Partial, putative, synthetic construct, predicted, not experimental, hypothetical protein, r-RNA, t-RNA and mitochondrial genes were eliminated and then the cleanup program (Grillo et al., 1996) was applied in order to get rid of redundancies from the remaining nucleotide sequences. A script implemented by us was then used to identify the coding sequences that began with a start codon, ended with a stop codon and contained no internal stop codons, so as to calculate reliable GC, GC_1, GC_2 and GC_3 values.

In the case of the correlations between exons and introns, the sequences of introns were retrieved from the website http://www.ensembl.org/index.html. In order to analyze the correlations between coding sequences and 5′+3′ flanking sequences, the coordinates of the genes on chromosomes were retrieved from the website from which the chromosomes were downloaded. In particular, we considered flanking regions (5′+3′) of different lengths: 1, 2, 5, 10, 20 and 40 kb. In some cases we also masked the interspersed repeated sequences retrieved from the UCSC website (http://genome.ucsc.edu).

Isochore families from human, chicken and fish genomes were those investigated by Costantini et al. (2006, 2007a,b). It should be stressed that isochore families are defined by distributing isochores (as assessed according Costantini et al., 2006) in 1 or 0.5% GC bins.

2.2. Methods

Orthogonal regressions (Jolicoeur, 1990; Clay et al., 1996; Cruveiller et al., 2004) were used in all the plots. This approach minimizes the sum of square distances between points and regression line. The use of orthogonal regression is appropriate in the cases under consideration, since large differences exist in the spread of points along the two Cartesian axes. Moreover, our

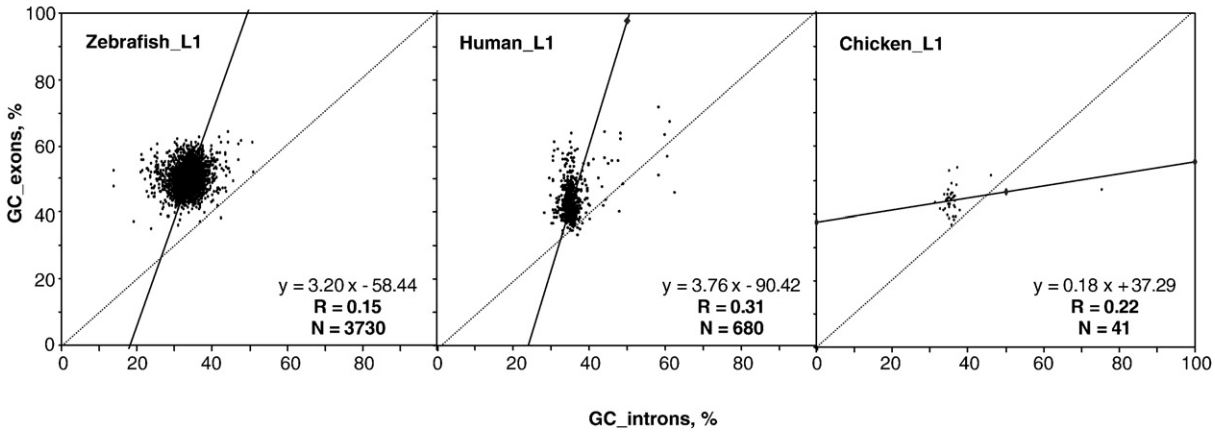

Fig. 2. Scatterplots of exon GC versus intron GC concerning the genes of Fig. 1 as partitioned according to isochore families from human, fish and chicken genomes (for other details see legend of Fig. 1).

aim here was to obtain a good representation of the scatterplot, rather than to find how one variable depended upon the other one (as with the linear regression).

3. Results

3.1. Compositional correlations of coding sequences with introns

Figs. 1–6 display the compositional correlations, the correlation coefficients and the orthogonal regression equations of GC levels of coding sequences versus the GC levels of the corresponding introns from whole genomes and for isochore families of fish, chicken and human. The results show that the base compositions of exons and introns are strongly correlated in the case of whole genomes, R values being in the 0.4–0.8 range, with a lower value (R=0.25) in the case of zebrafish. This value was at least in part due to the presence of interspersed repeated sequences, since when these were masked, the correlation coefficients increased to 0.33. The slopes of the regression line ranged from 0.65 to 1.07, with a higher

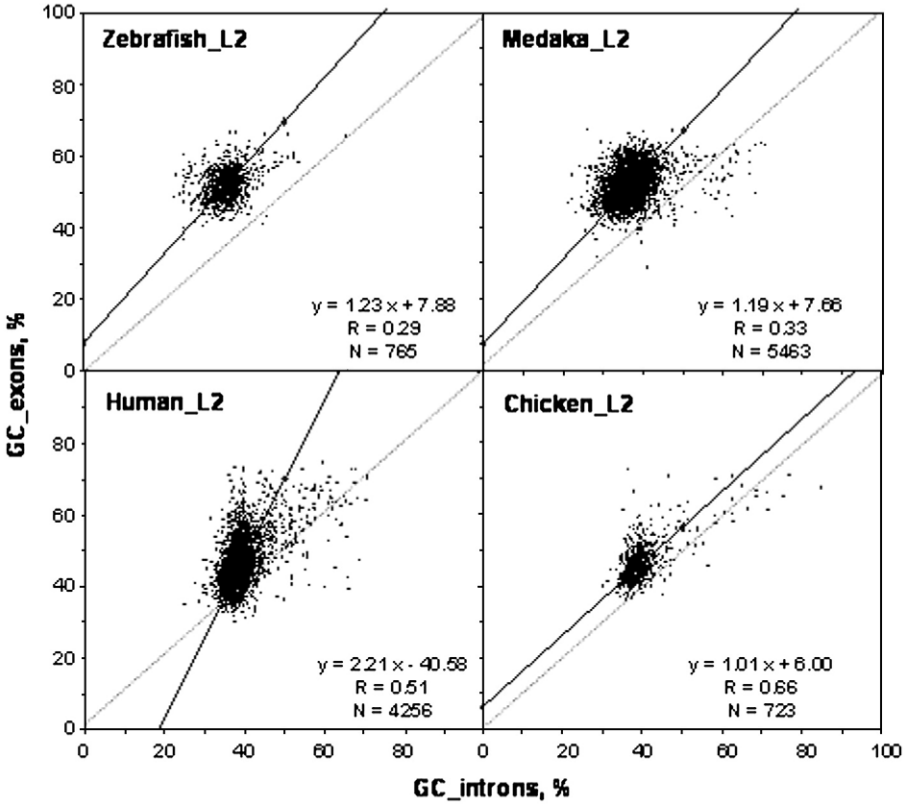

Fig. 3. Scatterplots of exon GC versus intron GC concerning the genes of Fig. 1 as partitioned according to isochore families from human, fish and chicken genomes (for other details see legend of Fig. 1).

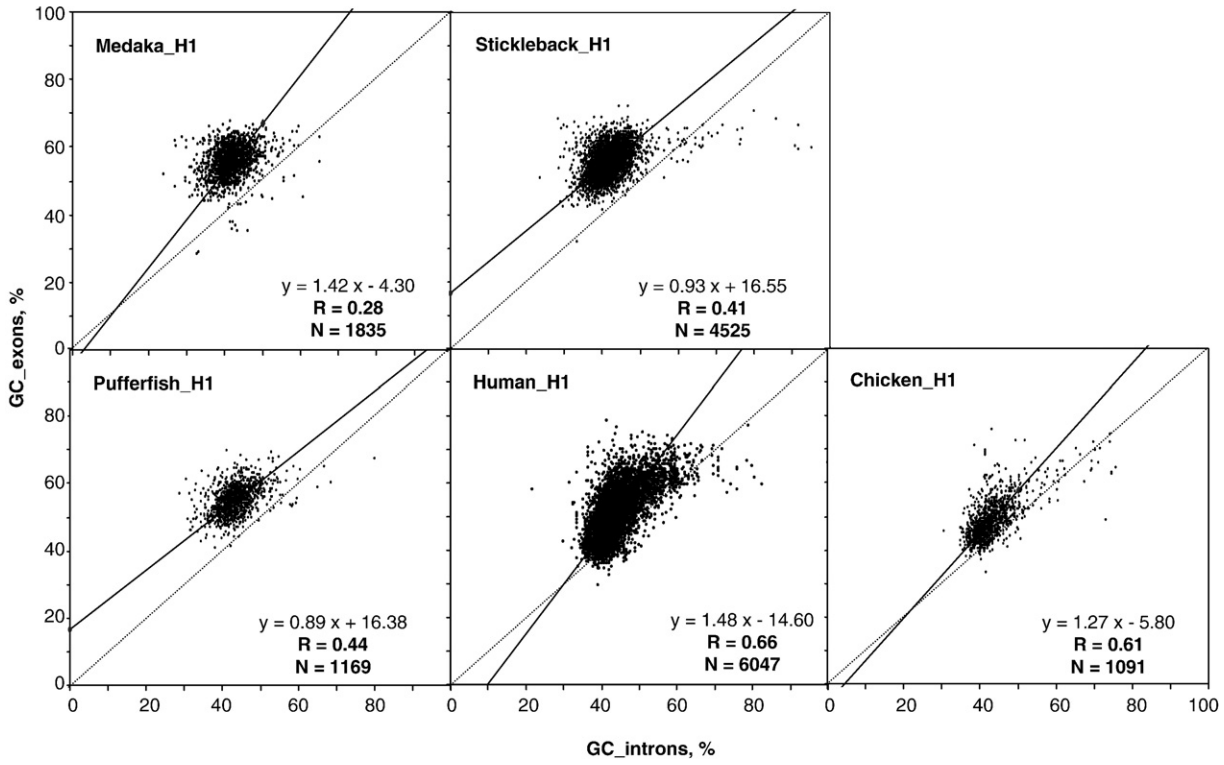

Fig. 4. Scatterplots of exon GC versus intron GC concerning the genes of Fig. 1 as partitioned according to isochore families from human, fish and chicken genomes (for other details see legend of Fig. 1).

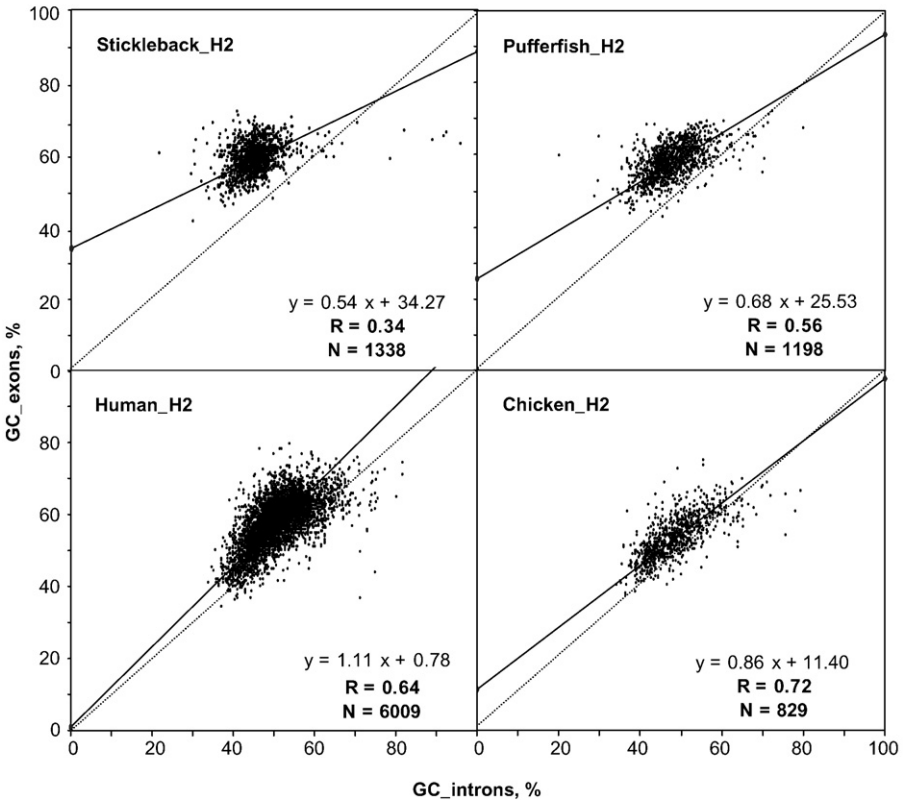

Fig. 5. Scatterplots of exon GC versus intron GC concerning the genes of Fig. 1 as partitioned according to isochore families from human, fish and chicken genomes (for other details see legend of Fig. 1).

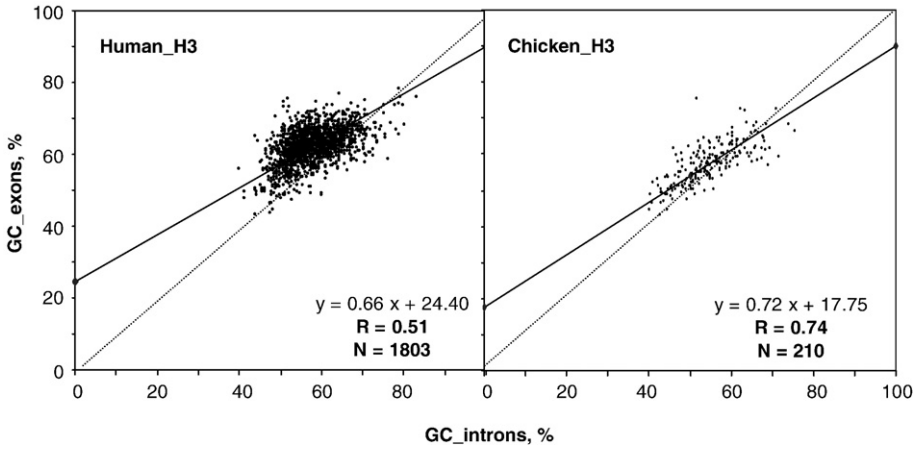

Fig. 6. Scatterplots of exon GC versus intron GC concerning the genes of Fig. 1 as partitioned according to isochore families from human, fish and chicken genomes (for other details see legend of Fig. 1).

value 1.82 in the case of zebrafish. As expected, strong correlations were also found between GC_3 and intron GC (Supplementary Figs. S1–S6), in which case slopes were regularly higher than in the case of the GC plots of Figs. 1–6.

In the case of individual isochore families, correlation coefficients were in the 0.3–0.7 range with the lower values in the GC poorer families. In the case of zebrafish, the very low value, 0.15, for the L1 family increased to 0.23 when the abundant repeated sequences were masked. The slopes of the regression lines were higher (S=3–4) in L1 families compared to all other families (S=0.5–1.5), with the single exception of the L1 family of chicken, where the number of genes was very small.

Compositional correlations between exons and introns from the corresponding isochore families of human and fish genomes (the total number of genes compared was 22,713 in the case of fishes, 18,795 in that of human) are compared in Fig. 7, which stresses their similarity.

The correlation coefficients linking the GC_1 and GC_2 levels of genes with those of introns for the whole genomes and for individual isochore families from zebrafish, medaka, stickleback, pufferfish, human and chicken were found to be lower than for GC and GC_3. In the case of GC_1, results show that the base compositions of exons and introns are strongly correlated in the whole genomes of pufferfish, human and chicken, R values being in the 0.4–0.6 range, whereas lower values (0.16, 0.18 and 0.18 respectively) were found for zebrafish, medaka and stickleback. In the case of individual isochore families, correlation coefficients were in the 0.11–0.47 range, the lower values being found in the GC poorer families.

In the case of GC_2, values of exons were correlated with GC of introns for whole genomes of human (R=0.44), stickleback, pufferfish and chicken genomes (R values being in the 0.12–0.32 range). Non-significant values were found in the case of zebrafish (R=0.03) and medaka (R=0.08). In the case of individual isochore families, correlation coefficients were comparable to, yet slightly lower, than those reported for GC_1 (Supplementary Table T3).

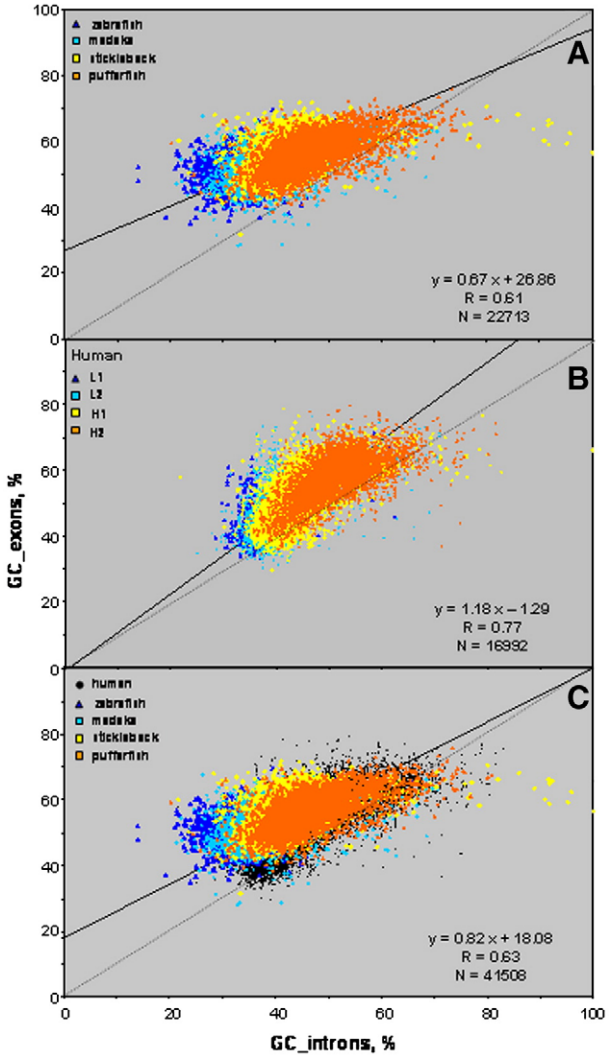

Fig. 7. Scatterplots of exon GC versus intron GC (A) in the genomes of four fishes, (B) in human isochore families, and (C) in human plus four fishes.

Table 1
Number of genes, correlation coefficients and orthogonal regression equations of GC of genes versus GC of flanking regions at different length and in the isochore families in zebrafish, pufferfish and human genomes

y	*x*	Species	Number of genes	Length_flanking (5'+3'), kb	Coefficient correlation	Regression equation	Coeff. correlation flanking masked	Regression equation flanking masked
$GC_{(gene)}$	$GC_{(flanking\ regions)}$	Zebrafish	5348	1	0.39	$y=0.712\times+24.10$		
				1 kb — repeats	0.39	$y=1.45\times-36.29$		
				2	0.34	$y=1.17\times+8.99$		
				5	0.27	$y=2.62\times-41.89$		
				5 kb — repeats	0.35	$y=2.20\times-27.47$		
				10	0.27	$y=3.94\times-89.78$		
				10 kb — repeats	0.33	$y=3.27\times-65.97$		
				20	0.28	$y=4.96\times-127.70$		
				40	0.26	$y=5.99\times-165.58$		
		Zebrafish_L1	4255	1	0.32	$y=0.57\times-29.31$	0.34	$y=0.60\times-27.76$
				2	0.24	$y=1.20\times-7.99$	0.30	$y=1.18\times-8.12$
				5	0.14	$y=5.54\times-143.13$	0.24	$y=3.57\times-75.37$
				10	0.08	$y=15.49\times-49.39$	0.18	$y=7.69\times-221.08$
				20	0.06	$y=27.94\times-938.68$	0.15	$y=12.17\times-381.72$
				40	0.06	$y-36.20\times-1237.55$	0.14	$y=16.86\times-550.80$
		Zebrafish_L2	944	1	0.49	$y=0.77\times+23.14$	0.51	$y=0.74\times+23.75$
				2	0.47	$y=1.09\times+11.72$	0.51	$y=1.05\times+13.47$
				5	0.35	$y=2.01\times-21.48$	0.43	$y=1.64\times-7.81$
				10	0.34	$y=2.76\times-50.33$	0.42	$y=2.53\times-42.12$
				20	0.33	$y=3.68\times-85.93$	0.41	$y=2.16\times-27.68$
				40	0.30	$y=5.02\times=138.38$	0.42	$y=3.04\times-62.15$
		Pufferfish	4351	1	0.43	$y=1.09\times-15.54$		
				2	0.43	$y=0.74\times+3.44$		
				5	0.44	$y=0.52\times+16.28$		
				10	0.45	$y=0.47\times+19.81$		
				20	0.44	$y=0.41\times+22.78$		
				40	0.40	$y=2.75\times-70.02$		
		Pufferfish_H1	1485	1	0.33	$y=0.80\times+19.44$		
				2	0.31	$y=1.63\times-16.52$		
				5	0.31	$y=3.21\times-85.46$		
				10	0.30	$y=4.49\times-142.90$		
				20	0.29	$y=6.15\times-216.99$		
				40	0.24	$y=8.29\times-311.80$		
		Pufferfish_H2	1509	1	0.40	$y=0.74\times+22.69$		
				2	0.38	$y=1.17\times+3.14$		
				5	0.39	$y=2.14\times-42.72$		
				10	0.40	$y=2.75\times-73.60$		
				20	0.37	$y=3.46\times-108.83$		
				40	0.33	$y=4.74\times-171.42$		
		Human	24346	1	0.69	$y=0.71\times+18.03$	0.70	$y=0.68\times+19.91$
				2	0.71	$y=0.87\times+11.06$	0.73	$y=0.78\times+15.55$
				5	0.73	$y=1.11\times+0.71$	0.75	$y=0.91\times+10.21$
				10	0.72	$y=1.28\times-6.58$	0.75	$y=0.99\times+7.03$
				20	0.71	$y=1.44\times-13.51$	0.74	$y=1.06\times-4.37$
				50	0.68	$y=1.64\times-22.42$	0.76	$y=1.19\times-1.26$
		Human_L1	908	1	0.41	$y=0.40\times+29.06$	0.38	$y=0.36\times+30.73$
				2	0.41	$y=1.15\times-12.75$	0.41	$y=0.59\times+22.55$
				5	0.38	$y=1.11\times+0.71$	0.41	$y=1.43\times-6.79$
				10	0.37	$y=1.28\times-6.58$	0.42	$y=2.84\times-55.79$
				20	0.33	$y=1.44\times-13.51$	0.39	$y=5.03\times-130.99$
				50	0.20	$y=19.15\times-652.48$	0.30	$y=10.16\times-305.15$
		Human_L2	5345	1	0.42	$y=0.53\times+24.39$	0.45	$y=0.52\times+25.24$
				2	0.41	$y=0.88\times-10.37$	0.45	$y=0.77\times+15.70$
				5	0.40	$y=2.54\times-55.09$	0.45	$y=1.71\times-19.72$
				10	0.36	$y=4.57\times-135.12$	0.43	$y=2.69\times-56.60$
				20	0.33	$y=6.99\times-231.06$	0.46	$y=3.73\times-95.35$
				50	0.24	$y=13.25\times-478.86$	0.39	$y=5.76\times-171.43$
		Human_H1	7876	1	0.55	$y=0.67\times+19.80$	0.58	$y=0.62\times+22.15$
				2	0.57	$y=0.97\times+6.26$	0.60	$y=0.81\times+14.28$
				5	0.56	$y=2.27\times-50.60$	0.62	$y=1.14\times-0.43$
				10	0.58	$y=1.63\times-22.29$	0.61	$y=1.45\times-12.40$

Table 1 (*continued*)

y	*x*	Species	Number of genes	Length_flanking (5'+3'), kb	Coefficient correlation	Regression equation	Coeff. correlation flanking masked	Regression equation flanking masked
				20	0.51	$y=3.12\times-87.70$	0.58	$y=1.79\times-27.06$
				50	0.43	$y=4.95\times-167.87$	0.52	$y=2.44\times-55.39$
		Human_H2	7788	1	0.55	$y=0.65\times+22.04$	0.58	$y=0.61\times-24.35$
				2	0.56	$y=0.90\times+9.16$	0.60	$y=0.78\times+15.92$
				5	0.57	$y=1.35\times-12.50$	0.60	$y=1.00\times-45.04$
				10	0.55	$y=1.80\times-34.69$	0.59	$y=1.19\times-4.44$
				20	0.51	$y=2.34\times-60.62$	0.57	$y=1.42\times-15.53$
				50	0.43	$y=3.48\times-116.42$	0.50	$y=1.91\times-40.28$
		Human_H3	2429	1	0.49	$y=0.51\times+30.96$	0.49	$y=0.53\times+30.06$
				2	0.49	$y=0.72\times+19.33$	0.50	$y=0.73\times-18.01$
				5	0.46	$y=1.03\times+2.11$	0.47	$y=0.98\times+3.98$
				10	0.44	$y=1.26\times-10.43$	0.47	$y=1.18\times-7.50$
				20	0.39	$y=1.62\times-30.08$	0.45	$y=1.47\times-24.11$
				50	0.34	$y=2.53\times-79.62$	0.39	$y=2.17\times-65.11$

3.2. Compositional correlations of coding gene sequences with flanking regions

Table 1 reports the lengths of the flanking regions explored, the correlation coefficients and the orthogonal regression equations linking the GC levels of coding sequences with those of flanking regions for the whole genomes from zebrafish, pufferfish and human and for the corresponding isochore families (the corresponding data for medaka, stickleback and chicken are reported in Supplementary Table T1). Supplementary Table T2 shows the data linking the GC_3 levels of coding sequences with those of flanking regions.

In the case of zebrafish, when the interspersed repeats that form 46.8% of the zebrafish genome (Costantini et al., 2007b) were removed from flanking regions, the correlation coefficients increased indicating that interspersed sequences may be largely responsible for the low correlation between coding sequences and flanking sequences. Incidentally, this was not the case for the 1-kb flanking sequences, in all likelihood because no interspersed repeats were present in the 500 bp flanking the gene on each side.

As far as isochore families are concerned, the trend towards a decrease of the correlation coefficient with increasing size of flanking sequences was also generally present. Masking interspersed repeated sequences led to higher correlation coefficients in the case of the two isochore families of zebrafish, whereas this effect was almost negligible in the case of human and chicken families.

As far as GC_1 and GC_2 in the case of whole genomes, correlation coefficients tended to be low ($R=0.10$–0.15) for whole genomes and individual isochore families for zebrafish, medaka and stickleback, but were higher for pufferfish, human and chicken (see Supplementary Tables T4 and T5).

4. Discussion

We should recall that compositional correlations between coding and contiguous non-coding sequences 1) were initially established twenty years ago (Bernardi et al., 1985; Bernardi and Bernardi, 1986) and were updated ten years ago (Clay et al., 1996); 2) concerned the small number of human genes which were available at those times; 3) covered the very wide spectrum of GC levels which is present in whole genomes from mammals. These points indicated that time was ripe for an updating, especially since the present situation in terms of available data is totally different from that prevailing in previous investigations. Indeed, a number of fully sequenced vertebrate genomes are now available and we could take advantage of this new situation. While there were good reasons for the present updating of the correlations, the analysis of individual isochore families was also important because the compositional heterogeneity of the genomes of warm-blooded vertebrates used could influence the correlations. Finally, this investigation covered human, chicken and four fishes, not only in order to explore those genomes, but also to compare the corresponding isochore families.

The main conclusions reached are (i) that significant ($p<10^{-4}$) compositional correlations between contiguous coding and non-coding sequences hold even when narrow compositional ranges such as those of isochore families are explored; and (ii) that the compositional correlations for corresponding isochore families are very similar, independently of the phylogenetic origin of the genome; the similarity of the compositional correlations among corresponding vertebrate families adds another feature to the compositional similarity of these families (Costantini et al., 2007a,b).

The differences in slopes and correlation coefficient found in corresponding isochore families may due to a number of reasons, such as isochore-specific abundance and/or the different extents of amplification of gene families and interspersed repeated sequences.

Incidentally, we found no indication for GC-rich isochores containing a mixture of GC-rich and GC-poor genes as claimed by Robins and Press (2005) and Press and Robins (2006). The discrepancy may be apparently due, however, to the much larger size of "isochores" as estimated by those authors compared to the "canonical" isochores of Costantini et al. (2006). Moreover, the data of Press and Robins (2006) were based on an approach which did not involve coding sequences that were localized on chromosomes, nor compared with their actual flanking sequences, as done here.

In conclusion, the existence of highly significant compositional correlations between coding and non-coding sequences is confirmed here on the basis of data which concern all individual isochore families from the genomes investigated. These correlations are a crucial property of the vertebrate genome for the reasons which are mentioned in the Introduction and which are discussed elsewhere (Bernardi, 2007).

Acknowledgements

We thank Fabio Auletta and Giuseppe Torelli for their help in computer work. We also thank Oliver Clay and Dr. Kamel Jabbari for helpful discussions and suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.12.016.

References

Bernardi, G., 1979. Organization and evolution of the eukaryotic genome. In: Morgan, J., Whelan, W.J. (Eds.), Recombinant DNA and genetic experimentation. Pergamon, New York, NY, USA, pp. 15–20.

Bernardi, G., 1984. Sequence organization of the vertebrate genome. In: Arber, W., Illmensee, K., Peacock, J., Starlinger, P. (Eds.), Genetic Manipulation: Impact on Man and Society. Cambridge University Press, Cambridge, UK, pp. 171–178.

Bernardi, G., 2004. Structural and Evolutionary Genomics. Natural Selection in Genome Evolution. Elsevier, Amsterdam, (The Netherlands). (reprinted in 2005).

Bernardi, G., 2007. The neoselectionist theory of genome evolution. Proc. Natl. Acad. Sci. U. S. A. 104 (20), 8385–8390.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Bernardi, G., et al., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and non-coding DNA: compositional correlations. Mol. Phylogenet. Evol. 5, 2–12.

Cortadas, J., Olofsson, B., Meunier-Rotival, M., Macaya, G., Bernardi, G., 1979. The DNA components of the chicken genome. Eur. J. Biochem. 99, 179–186.

Costantini, M., Clay, O., Auletta, F., Bernardi, G., 2006. An isochore map of human chromosomes. Genome Res. 16, 536–541.

Costantini, M., Di Filippo, M., Auletta, F., Bernardi, G., 2007a. Isochore pattern and gene distribution in the chicken genome. Gene 400 (1–2), 9–15.

Costantini, M., Auletta, F., Bernardi, G., 2007b. Isochore patterns and gene distributions in fish genomes. Genomics 90, 364–371.

Cruveiller, S., Jabbari, K., Clay, O., Bernardi, G., 2004. Incorrectly predicted genes in rice? Gene 333, 187–188.

Cuny, G., Macaya, G., Meunier-Rotival, M., Soriano, P., Bernadi, G., 1978. Some properties of the major components of the mouse genome. In: Boyer, W.H., Nicosia, S. (Eds.), Genetic Engineering. Elsevier, Amsterdam, The Netherlands, pp. 109–115.

Filipski, J., Thiery, J.P., Bernardi, G., 1973. An analysis of the bovine genome by Cs_2SO_4-Ag^+ density gradient centrifugation. J. Mol. Biol. 80, 177–197.

Grillo, G., Attimonelli, M., Liuni, S., Pesole, G., 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. Comput. Appl. Biosci. 12, 1–8.

Jolicoeur, P., 1990. Bivariate allometry: interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. J. Theor. Bio. 144, 273–285.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., Bernardi, G., 1991. The distribution of genes in the human genome. Gene 100, 181–187.

Ohno, S., 1972. So much "junk" DNA in our genome. Brookhaven Symp. Biol. 23, 366–370.

Press, W.H., Robins, H., 2006. Isochores exhibit evidence of genes interacting with the large-scale genomic environment. Genetics 174, 1029–1040.

Robins, H., Press, W., 2005. Human microRNAs target a functionally distinct population of genes with AT-rich 3′ UTRs. Proc. Natl. Acad. Sci. U. S. A. 102, 15557–15562.

Thiery, J.P., Macaya, G., Bernardi, G., 1976. An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. 108, 219–235.

Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.