Isochore patterns and gene distributions in fish genomes

Maria Costantini, Fabio Auletta, Giorgio Bernardi *

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

Received 3 April 2007; accepted 11 May 2007
Available online 21 June 2007

Abstract

The compositional approach developed in our laboratory many years ago revealed a large-scale compositional heterogeneity in vertebrate genomes, in which GC-rich and GC-poor regions, the isochores, were found to be characterized by high and low gene densities, respectively. Here we mapped isochores on fish chromosomes and assessed gene densities in isochore families. Because of the availability of sequence data, we have concentrated our investigations on four species, zebrafish (Brachydanio rerio), medaka (Oryzias latipes), stickleback (Gasterosteus aculeatus), and pufferfish (Tetraodon nigroviridis), which belong to four distant orders and cover almost the entire GC range of fish genomes. These investigations produced isochore maps that were drastically different not only from those of mammals (in that only two major isochore families were essentially present in each genome vs five in the human genome) but also from each other (in that different isochore families were represented in different genomes). Gene density distributions for these fish genomes were also obtained and shown to follow the expected increase with increasing isochore GC. Finally, we discovered a remarkable conservation of the average size of the isochores (which match replicon clusters in the case of human chromosomes) and of the average GC levels of isochore families in both fish and human genomes. Moreover, in each genome the GC-poorest isochore families comprised a group of “long isochores” (2–20 Mb in size), which were the lowest in GC and varied in size distribution and relative amount from one genome to the other.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Compositional heterogeneity; Gene distribution; Genomes; Isochores

Investigations published 30 years ago led to the discovery of phylogenetic differences at the macromolecular level in eukaryotic genomes [1]. In particular, a major compositional difference was found between the genomes of warm- and cold-blooded vertebrates. While the former were very heterogeneous, as first observed in the “main band” (satellite DNAs are neglected here) of calf DNA [2], the latter were characterized by a much lower compositional heterogeneity.

The genomes of mammals are in fact mosaics of isochores [3], compositionally fairly homogeneous regions that can be assigned to a small number of families. This discontinuous compartmentalization covering a very wide range (34 to 59% GC in the human genome) was recently confirmed [4] at the sequence level. In the case of a typical mammalian genome, the human genome, the ~3200 isochores were mapped on chromosomes, where they form the ultimate bands and allow a precise definition of band borders at the classical 400- and 850-band resolution [5]. The average size of human isochores is 0.9 Mb; the standard deviations of GC levels are around 1% GC within isochores covering 85% of the genome and around 2% GC in the remaining, mostly GC-rich, isochores. If isochores are pooled in bins of 1% GC, their distribution confirms that they belong in the five families that we have previously described (L1, L2, H1, H2, and H3, in order of increasing GC; [6]).

The assessment of gene density in compositional DNA fractions led us to the discovery that genes are not uniformly distributed in mammalian genomes [6–8]. Indeed, in the human genome, almost two-thirds of protein-coding genes are concentrated in the GC-richest isochrome families H2 and H3, the “genome core” [9], which represent only 15% of the genome, the rest being thinly spread over the other isochrome families, the “genome desert” [10]. These two “gene spaces” [10] are different, not only in gene density, but also in a number of other basic structural and functional properties, such as

* Corresponding author. Fax: +39 081 2455807.
E-mail address: bernardi@szn.it (G. Bernardi).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.05.006

As far as investigations on the genomes of cold-blooded vertebrates are concerned, our initial observations [1], concerning only two reptiles, two amphibians, and two fishes, showed remarkable differences compared with the genomes of warm-blooded vertebrates, in that the former were much less heterogeneous in composition than the latter. While the work on reptile and amphibian genomes was resumed only later, an important step was made with fish genomes already by Hudson et al. [12], whose study of 33 fish species representing 12 orders of Teleostei and one of Chondrichthyes put on a wider basis the large differences found at the DNA level between the genomes of cold- and warm-blooded vertebrates.

The major conclusion was that the main CsCl bands of fish genomes were in most cases characterized by very low compositional heterogeneities. One of the fish, Arothron diadematus, having the typical very small genome of Tetraodontids (c = 0.4–0.5 pg; [13]), was studied in more detail and was shown to comprise “single-copy” sequences that represented as much as 87% of its DNA [14], as opposed to about 50% in human DNA.

A detailed study of the genomes of cold-blooded vertebrates [15,16], including 122 fish species from 21 orders of Osteichthyces and 3 orders of Chondrichthyces (a sample later expanded to 201 species by [17]), showed a broad range of average base composition and genome sizes (neglecting polyploidy), but low compositional heterogeneities of CsCl profiles compared to warm-blooded vertebrates. Preparative fractionation in Cs2SO4/BAMD density gradients (BAMD is bis(acetatomercurimethyldioxane)) of several fish DNAs (i) stressed the differences between the GC-poorest genome of Cyprinus carpio, a cyprinid, and the GC-richest genome of a Tetraodontid, A. diadematus; (ii) confirmed the positive correlation between DNA heterogeneity and staining contrast of chromosomal bands [18]; and (iii) confirmed the differences [19] that exist between the genomes of fishes living at high temperatures (40°C) and those of their congeners living at lower temperatures (20°C).

As far as gene distribution is concerned, the general conclusion was that the well-established bimodality shown by mammalian (and avian) genomes, with high gene densities in the GC-rich isochores and low gene densities in GC-poor isochores, was also present in the genomes of cold-blooded vertebrates (see [11] for a review). Since the genomes of zebrafish (Brachidionrio rerio), medaka (Oryzas latipes), stickleback (Gasterosteus aculeatus), and pufferfish (Tetraodon nigroviridis) are now very largely or fully sequenced, we investigated these genomes in both isochore patterns and gene distribution. Some comparative data from fugu (Takifugu rubripes), whose genome sequence is still incomplete, are also presented.

Results

A histogram derived from Bucciarelli et al. [17] presenting the number of fish species against the modal buoyant densities and GC levels of the corresponding genomes is displayed in Fig. 1 to show that the four genomes investigated here not only derive from fishes that belong to distant orders and are very different in genome size (see Ref. [17]), but also cover almost the full compositional range of isochore families and provide, therefore, a general picture of fish genomes.

Fig. 2 shows the GC profiles (using a fixed window of 100 kb; see [4], Methods, and Supplementary Fig. S1) of the first seven chromosomes from the four fishes. A full display of them is presented in Supplementary Figs. S2A–S2D. As shown in Fig. 2, the chromosomes of zebrafish consist almost solely of GC-poor isochores L1 and L2, whereas those of pufferfish are predominantly characterized by GC-rich isochores, the chromosomes from the other two fishes showing compositionally intermediate GC profiles. Interestingly, in all cases telomeres tend to be more GC rich than the other regions of chromosomes. Another evident feature is the various genome sizes of the four genomes under investigation (see also below).

Fig. 3 displays isochore maps of some ~40-Mb-size chromosome regions from the four fishes. They were constructed using the methodology of Costantini et al. [4], which is based on the standard deviation of GC levels in adjacent 100-kb sequences as scanned from the end of each chromosome. A 1% GC standard deviation was accepted, larger GC jumps being taken as borders between subsequent isochores. Supplementary Tables T1–T4 provide the coordinates, the GC level, and the standard deviation for each isochore from the four fish genomes.

If isochores are pooled in bins of 0.5% GC (Fig. 4), zebrafish was practically made up of only L1 and L2 isochores, with a predominance of the former family (75.7% vs 23.3%), whereas the isochores of pufferfish consist of H1 (55.7%) and H2
families with a minor presence of L2 (3.5%) and H3 isochores (2.5%). The compositionally “intermediate” genomes of medaka and stickleback consist essentially of L2/H1 and H1/H2 isochores, respectively, the first family being predominant in each case. These remarkable distributions, with almost no overlap between zebrafish and pufferfish, are compared in Fig. 4 with the human isochore pattern, which covers a very wide spectrum of isochores, in fact a spectrum as wide as that covered by all fish genomes.

Fig. 4 also shows the dramatically narrower distribution of fish genomes compared with the human genome, which is a good representative of mammalian genomes. Indeed, in each of the four fish genomes studied, practically only two major isochores families are represented, L1 and L2 in zebrafish, L2 and H1 in medaka, and H1 and H2 in stickleback and pufferfish, the first family being predominant except in the last case. Because of this situation, only two different arrangements of flanking isochores are predominantly found in chromosomes, namely GC-poor isochores flanked by GC-rich isochores and GC-rich isochores flanked by GC-poor isochores (see Supplementary Fig. S3). Indeed, the other families are so underrepresented that they can be neglected. A detailed assessment of flanking isochores, including the minor families, is, however, given in Supplementary Fig. S4. It should be noted that all four genomes exhibit “transition isochores,” in that one flanking isochore is higher, the other lower (as in the case of the human genome; see [5]). Such transition isochores represent only 5.6% of the genome for zebrafish and 5.9% for medaka, but 17.5% for stickleback and 18.6% for pufferfish. The higher percentage of transition isochores in the two GC-rich genomes can be understood because they correspond to steps in the formation of the blocks of isochores that are assembled into the high-resolution chromosomal bands. This situation is reminiscent of that reported for the GC-rich vs the GC-poor isochores of the human genome [5].
Interestingly, Fig. 4 and Table 1 show that the isochore families are very close in average GC levels from fish to human, the maximum variation from the overall average being within 1% GC. This point will be commented upon under Discussion.

As far as the average isochore size is concerned, the GC-poorest isochore families (L1 in zebrafish and human, L2 in medaka, and H1 in stickleback) are bimodal in that they comprise two size groups (see Supplementary Fig. S5). Table 1 also presents the percentages of the isochores that are below 2 and 3 Mb in size, respectively, the corresponding percentages of the longer isochores being given by the difference from 100%. Very interestingly, the first group (labeled A in Supplementary Fig. S5) is remarkably constant in size, as are the isochores belonging to the other families (see Table 2). In contrast, the second group (labeled B in Supplementary Fig. S5) is different in relative amounts (see Table 1) and size distribution (see Supplementary Fig. S5) in different genomes. Table 1 also shows that while there is no compositional difference in the first group of the GC-poorest isochore family, whether the chosen upper limit is 2 or 3 Mb, the long isochores of the second group are regularly lower in GC. It should be noted that the size distribution of the second group may be affected by some errors because of the large
number of gaps (or the surprising lack of them in the case of stickleback). Finally, the second group is practically nonexistent in the next GC-rich family of isochores (L2 relative to L1, H1 relative to L2, etc.; see Supplementary Fig. S5).

The distribution of coding sequences in different isochore families (Fig. 5) showed that gene concentration was higher in GC-rich compared to GC-poor isochores, with the only apparent exception of a higher concentration in L1 compared to L2 isochores in the case of zebrafish. This case was investigated further (see Discussion). Again, a comparison with human DNA is also shown for the purpose of emphasizing the much steeper gene concentration gradient of mammalian genomes, which parallels the steeper compositional gradient.

Coding sequences of the four fishes showed compositional differences that were smaller compared to isochores (as seen in Supplementary Figs. S6A–S6D for GC, GC1, GC2, and GC3 levels). While isochore GC showed an 8.7% difference between pufferfish and zebrafish, the difference between the overall GC
levels of their coding sequences was only 5.7%. Because it was due mainly to GC3 ($\Delta$GC3 was 8.8%), this difference was smaller in GC1 (2.5%) and GC2 (5.6%). Interestingly, a difference was also found between the two Tetraodontids, tetraodon always reaching higher values compared to fugu. The difference between GC levels of isochores and GC levels of coding sequences prompted an analysis of interspersed repeats, which revealed that they formed 46.8% of the zebrafish genome, but only 1.9% of pufferfish genome (confirming our previous results; see [14]). A histogram of interspersed repeats in the isochores of zebrafish (Supplementary Fig. S7) showed that the proportion of DNA that they contribute in each GC interval remains roughly constant, but for each of the four major classes of repeats the sequences present in L1 were GC poorer than those present in L2 (see Supplementary Table T5), following a compositional matching common for interspersed repeated sequences (see [11]). Finally, Supplementary Fig. S8 confirms the previously reported inverse correlation [15] between genome average GC and genome size for the four fishes investigated here.

### Discussion

The results obtained in the present investigations are of interest for three different reasons. The first is that they confirmed and extended previous findings from our laboratory [15,16], the second is that they led to the discovery of some novel features of isochores, the third that they have important implications as far as genome evolution is concerned.

### Table 1

(A) Average GC levels of isochore families in the four fish genomes and in the human genome

<table>
<thead>
<tr>
<th></th>
<th>L1 (%)</th>
<th>L2 (%)</th>
<th>H1 (%)</th>
<th>H2 (%)</th>
<th>H3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zebrafish</td>
<td>36.0</td>
<td>38.2</td>
<td>42.3</td>
<td>47.3</td>
<td>54.7</td>
</tr>
<tr>
<td>Medaka</td>
<td>39.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stickleback</td>
<td>44.2</td>
<td>44.4</td>
<td>43.5</td>
<td>48.1</td>
<td>54.6</td>
</tr>
<tr>
<td>Human</td>
<td>36.0</td>
<td>38.9</td>
<td>43.1</td>
<td>48.7</td>
<td>54.5</td>
</tr>
<tr>
<td>Overall</td>
<td>36.0</td>
<td>39.0</td>
<td>43.5</td>
<td>48.1</td>
<td>54.6</td>
</tr>
</tbody>
</table>

(B) Average GC levels of the GC-poorest isochore families, as determined in three different size ranges

<table>
<thead>
<tr>
<th></th>
<th>&lt;2 Mb</th>
<th>&lt;3 Mb</th>
<th>&gt;3 Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zebrafish_L1</td>
<td>37.6</td>
<td>37.5</td>
<td>36.1</td>
</tr>
<tr>
<td>Medaka_L2</td>
<td>40.1</td>
<td>40.1</td>
<td>39.2</td>
</tr>
<tr>
<td>Stickleback_H1</td>
<td>44.4</td>
<td>44.4</td>
<td>43.7</td>
</tr>
<tr>
<td>Pufferfish_H1</td>
<td>44.1</td>
<td>44.1</td>
<td>43.9</td>
</tr>
<tr>
<td>Human_L1</td>
<td>36.1</td>
<td>36.1</td>
<td>35.3</td>
</tr>
</tbody>
</table>

(C) Relative amounts of the GC-poorest isochore families, as determined in two different size ranges

<table>
<thead>
<tr>
<th></th>
<th>&lt;2 Mb</th>
<th>&lt;3 Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zebrafish</td>
<td>73</td>
<td>83</td>
</tr>
<tr>
<td>Medaka</td>
<td>68</td>
<td>77</td>
</tr>
<tr>
<td>Stickleback</td>
<td>67</td>
<td>77</td>
</tr>
<tr>
<td>Pufferfish</td>
<td>94</td>
<td>98</td>
</tr>
<tr>
<td>Human</td>
<td>76</td>
<td>86</td>
</tr>
</tbody>
</table>

### Table 2

Average sizes of isochores belonging to different families

<table>
<thead>
<tr>
<th></th>
<th>L1 (%)</th>
<th>L2 (%)</th>
<th>H1 (%)</th>
<th>H2 (%)</th>
<th>H3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zebrafish</td>
<td>0.8</td>
<td>0.5</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Medaka</td>
<td>0.7</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Stickleback</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Pufferfish</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Human</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

In the GC-poorest families of each fish genome, values concern isochores less than (A) 2 or (B) 3 Mb in size.
Along the first line, one should mention the results concerning the compositional patterns and gene distribution. The two major compositional features of fish genomes, the wide intergenomic spread of base composition (Fig. 1) and the narrow intragenomic distribution (Fig. 4), were confirmed on the basis of sequences. This provided, as expected, a more precise picture compared to the results obtained by the ultracentrifugation approach previously used. In each fish genome essentially only two isochore families were present, one of them being predominant (although less so in the case of pufferfish), whereas in the human genome five families are present. The compositional spread of the fish genomes analyzed here, on the other hand, was so wide that there was essentially no overlap between the isochore families of zebrafish and pufferfish.

Both the narrow intragenomic distribution and the wide compositional spread of fish genomes can be understood in terms of adaptation to environmental factors. While the narrow intragenomic distribution may be visualized as an adaptation to a particular ecological niche, the wide compositional spread implies the existence of compositional transitions involving whole genomes and responding to changes in environmental conditions and new adaptations. As already pointed out, such transitions may occur among genomes of fishes belonging to different orders, or even to different families independent of geological time (see [16]). What is remarkable is that those “whole genome shifts” lead to different patterns of isochore families, which are centered on the same values from fish to human, as if only “quantum jumps” into allowed states were permitted.

As far as the mechanisms leading to the “whole genome shift” discussed in this paper, they appear to be due mainly to mutations in the sequences coding for protein subunits of the replication machinery that lead to directional changes (AT → GC, GC → AT) as in the case of “mutator mutations” of prokaryotes [21,22]. Variations in genome size may also contribute, since increases and decreases in GC-poor intergenic and intronic sequences lead to compositional changes.

In some cases, selective advantages linked to compositional transitions could be identified. For instance, the differences found between two Tetraodontids, fugu and tetraodon (see Supplementary Fig. S6), may be correlated with the two different body temperatures of these closely related fishes. The former “cold” marine fish showed a distribution of DNA [20] and coding sequences (see Supplementary Fig. S6) that do not reach the same level as in the case of the latter, a tropical, freshwater fish. Another example, also associated with body temperature differences, is the higher GC level attained by Gillichthys seta, a gobid living in up to 40°C water, vs Gillichthys mirabilis, a congeneric species living at about 20°C (Bucciarelli et al., paper in preparation).

Concerning gene distribution, gene density increased with increasing GC level of isochores as indicated by previous results in both warm- and cold-blooded vertebrates. As in the case of mammals, the gene density gradient in fish genomes is probably correlated with the open chromatin structure linked to the higher GC level of isochores. An apparent exception to this general rule is the zebrafish, in which case gene density was found to be higher in L1 compared to L2 isochores. This may, however, be due to artifactual reasons, such as the relatively small gene sample used, the small amount of DNA in L2 isochores, or the expansion of the GC-poorest interspersed repeats.

Some major novel findings of this work suggest correlations with underlying genomic features common to all vertebrates (and possibly extending to other eukaryotes). The first is that the size of isochores from different families shows a remarkable conservation from fishes up to human, with the exception of the group of “long isochores,” which represent the GC-poorest tail of the GC-poorest family of each genome and are different in relative amounts and size distributions in different genomes.

The conservation of isochore size may be related to the fact that the human isochore map of Costantini et al. [4] perfectly coincides with the map of replicon clusters as defined by Watanabe et al. [23] for chromosomes 21q and 11q. More specifically, GC-rich isochores correspond to early replicating units, GC-poor isochores to late replicating units.

The second finding is that the average GC level of isochore families of fishes and human agree within 1% (see Table 1). The reasons for such discrete, conserved distribution are not yet clear. In any case, the newly discovered features of isochores that are evolutionarily conserved in vertebrates reinforce the idea that isochores represent “a fundamental level of genome organization” [24].

Finally, it should be mentioned that the results obtained in the investigations on the fish genomes and the comparisons with the human genome have important implications for the wider issue of genome evolution in vertebrates, a problem that has been recently discussed elsewhere [25].

Methods

Isochore mapping

The methodology used for isochore mapping was described by Costantini et al. [4]. The entire chromosomal sequences of the finished genome assembly for B. rerio (UCSC Release danRer4, http://genome.ucsc.edu), O. latipes (Ensembl Release 41.1, http://www.ensembl.org/index.html), G. aculeatus (Ensembl Release 41.1a), and T. nigroviridis (UCSC Release tetNgl) were partitioned into nonoverlapping 100-kb windows, and their GC levels were calculated using the program draw_chromosome.gc.pl [26,27] (http://genomat.img.cas.cz).

As far as the name of each isochore band is concerned we decided to use a convention in which the first number in the name represents the chromosome (or the group) number (the Roman number in the case of stickleback), the following two letters are the initials of the scientific name of the fish under consideration, and the last number identifies the band (see Supplementary Tables T1–T4).

Analysis of genes

The zebrafish genes were retrieved from Hovergen (Release 47, July 2005). The genes from medaka (Release 41.1), stickleback (Release 41.1a), tetraodon (Release 41.1 g), and fugu (Release 43.4e) were retrieved from Ensembl (http://www.ensembl.org/index.html). Partial, putative, synthetic construct, predicted, not experimental, hypothetical protein, r-RNA, t-RNA, ribosomal, and mitochondrial genes were eliminated and then the cleanup program [28], a fast computer program for cleaning nucleotide sequence databases of redundancies, was applied. For the remaining genes a script implemented by us was used to identify the coding sequences beginning with a start codon, ending with a stop codon, and containing no internal stop codons so as to
calculate reliable GC, GC1, GC2, and GC3 values. Using this protocol, we obtained from the four fishes complete coding sequences: 5348 for zebrafish, 7682 for medaka, 8818 for stickleback, 6244 for pufferfish, and 4496 for fugu.

The coordinates of the genes on the chromosomes were retrieved from the Web site from which the chromosomes were downloaded and used to calculate the gene density in the isochore families for the four fishes under analysis.

**Interspersed repeats**

Repeated sequences were retrieved from the UCSC Web site (http://genome.ucsc.edu). We retrieved in the annotation database the files rmsk.txt.gz, which contain information on the classification of repeats. To calculate the percentage of repeated sequences in chromosomes we retrieved the sequences of masked chromosomes, in which repeats (identified by RepeatMasker and Tandem Repeat Finder) are in lowercase and nonrepeating sequences are in uppercase.

**Acknowledgments**

We thank Kamel Jabbari for helpful discussions and Oliver Clay and Giacomo Bernardi for comments. We also thank Giuseppe Torelli for his help in the computer work.

**Appendix A. Supplementary data**

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.05.006.

**References**


**Web site references**

UCSC Genome Bioinformatics: http://genome.ucsc.edu.