Environmental genomics:

a tale of two fishes

Giuseppe Bucciarelli[1], Miriam Di Filippo[1], Domenico Costagliola[1],
Fernando Alvarez-Valin[2], Giacomo Bernardi[3], and Giorgio Bernardi[1*]

[1] *Stazione Zoologica Anton Dohrn, Villa Comunale, Napoli 80121, Italy*
[2] *Sección Biomatemática, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay*
[3] *Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, 95064, California, USA.*

*Corresponding author. Tel.: +39 081 5833402; Fax: +39 081 2455807; e-mail address: bernardi@szn.it
E-mail addresses: quao97@yahoo.com (G. Bucciarelli); miriam@szn.it (M. Di Filippo); domenico.costagliola@fastwebnet.it (D. Costagliola); falvarez@fcien.edu.uy (F. Alavarez-Valin); bernardi@biology.ucsc.edu (G. Bernardi).

The environment moulds the genome through natural selection

Keywords: Body temperature. Evolution. Genomes. Interspersed repeats. Isochores. Speciation.

ABSTRACT

The influence of the environment on two congeneric fishes, *Gillichthys mirabilis* and *Gillichthys seta*, that live in the Gulf of Mexico at temperatures of 10°- 25°, and up to 42°- 44°, respectively, was addressed by analyzing their genomes. Compared to *G. mirabilis*, *G. seta* showed some striking features: (i) extremely fast substitution rates in mitochondrial genes, indicating a divergence time of less than 0.66-0.75 million years ago; (ii) an expansion of a GC-rich minisatellite in the gene-rich regions of the nuclear genome; (iii) a decrease in DNA methylation; (iv) ratios of non-synonymous/synonymous changes (Ka/Ks) suggesting that some genes may be under positive selection; (v) high ratios of transversions over transitions and of AT to GC over GC to AT. These observations (i) indicate that the environment can rapidly mould the genome through natural selection and (ii) provide a model for the genome changes that accompany body temperature increases, as found after the emergence of homeothermy.

The sequences of the nuclear and mitochondrial genes were deposited in -------. Supplemental material includes supplemental Figs.S1-S3 and supplemental Tables S1-S4.

INTRODUCTION

Classically, sequence changes in the genome were visualized as resulting from point mutations and recombination. We found, however, that the vertebrate genomes underwent massive regional GC increases at the emergence of homeothermy (Thiery et al. 1976; Macaya et al. 1976), and proposed that these changes were due to the need of maintaining the thermodynamic stability of DNA, RNA and proteins (GC-rich codons preferentially encoding aminoacids that stabilize proteins) at the higher body temperature of warm-blooded vertebrates (Bernardi and Bernardi 1986; Bernardi 2004, 2007). This "thermodynamic stability hypothesis" was supported by finding that compositional changes affected only the gene-rich and not the gene-poor regions of the genome. Indeed, the gene-rich regions (the "genome core" see Bernardi 1993) are characterized by an open chromatin structure (Bernardi 2000; Saccone et al. 2002; Di Filippo and Bernardi 2008) and need an increased GC level to be stable at 37°- 40°, whereas the gene-poor regions (the "empty space", or "the genome desert" Bernardi 2004) are embedded in a closed chromatin structure (Saccone et al. 2008) which can by itself stabilize DNA.

A critical test to demonstrate that an environmental factor, such as temperature, can affect the structure of the genome is provided here by comparing the compositional patterns, the DNA methylation, and the nucleotide substitutions in the nuclear and the mitochondrial genomes of two congeneric goby fishes that live at very different temperatures (Huang et al. 2001; Fields et al. 2002). The sister relationship of these two species solves the problem we were confronted with in our initial work (Bernardi and Bernardi 1986) on the Death Valley pupfish, *Cyprinodon salinus*, and the Lake Magadi tilapia, *Oreochromis alcalicus grahami*, which showed regional GC increases in their genomes, but could only be compared with evolutionarily distant species.

The long-jawed mudsucker *G. mirabilis* inhabits salt water creeks in coastal California, Baja California, and the northern Gulf of California. The short-jawed mudsucker *G. seta*, a paedomorphic variant of *G. mirabilis* (Barlow 1961), is restricted to the uppermost tide pools, that are reached by sea water only rarely at the highest spring tides, in the northern Gulf of California. While *G. mirabilis* lives at 10°- 25°C, *G. seta* experiences temperatures that may reach 42°- 44°C, among the highest temperatures encountered by any fish (Nelson 2006) and hypoxia. *G. mirabilis* was previously studied in its hypoxia-induced gene expression (Gracey et al., 2001) and its response to heat stress (Buckley at al. 2006; Hochachka and Somero 2002; Cossins and Crawford 2005). In this study, we used two experimental approaches, working at the genome level and at the level of orthologous genes, respectively.

RESULTS

The compositional patterns of the nuclear genomes. In the first approach, the GC profiles of the two genomes were visualized by analytical cesium chloride centrifugation (Fig. 1a). As usual with most fish genomes (Bernardi and Bernardi 1990; Bucciarelli et al., 2002) , a fairly symmetrical DNA profile was observed for *G. mirabilis*, whereas *G. seta* exhibited a shoulder on the GC-rich side of the main peak. When the two DNAs were subjected to preparative ultracentrifugation (Fig. 1b), the CsCl profiles of the first two fractions, largely corresponding to the main peak, were identical in both species (Fig. 1 c, d). In contrast, the two GC-rich fractions showed a gradual increase in buoyant density in *G. mirabilis*, but a shoulder at about 40% GC and a major peak at about 46% GC in *G. seta* (Fig.1e, f).

When cDNAs from both fishes were hybridized on shallow gradient fractions from both DNAs (Fig. 2a)*,* the hybridization profiles in *G. mirabilis* largely followed the main DNA band, only a small shoulder appearing in the GC-rich range (44-47% GC), whereas the *G. seta* profiles showed a major peak in the GC-rich fractions (44-45% GC) and a minor peak corresponding to the main DNA band.

The expansion of a minisatellite in the *G. seta* genome. Since only very small compositional differences were found in the GC levels of orthologous coding sequences (see below), we explored the possibility that the shift of *G. seta* genes to higher densities was due to the expansion of interspersed GC-rich sequences. The GC-rich DNA fractions of *G. seta* were, therefore, cloned after AluI digestion. Out of 1000 clones, 10% consisted of arrays of a 38bp tandem repeat sequence (5'-

CTGGTTTGGGTTGGACCTGTTTCAGTCCCGTGTGAGTC-3') that exhibited a similarity of 80% among each other. The repeats showed a comparatively very high GC level (51% *vs* 39% for the main peak), a very strong strand asymmetry (A 10.8%, G 33.1%, C 18.7 %, T 37.4%), sizes in the 180-660 bp range and short internal repeats (see Supplemental Fig. S1). Specific primers designed on the basis of *G. seta* AluI repeats allowed us to sequence the corresponding family from *G. mirabilis*, which showed an interspecific similarity of 81%.

When AluI sequences of *G. seta* and *G. mirabilis* were hybridized on the shallow gradient fractions of the corresponding species, they produced a hybridization profile characterized by a major peak around 47% GC and a minor peak at 40-45% GC (Fig. 2b). The amount of the AluI sequences in the two genomes were then quantified by hybridization with total DNAs from both species and shown to be 2.5 times more abundant in *G. seta* compared to *G. mirabilis.* In contrast, the cloned DNA fragments obtained from PstI digestion (another family of repeats; see Supplemental Fig. S2 for the sequence) did not show any difference in amount in the two fishes and its hybridization followed the profile of the main peak (Fig. 2b). The presence of similar or identical AluI repeats in intergenic sequences of Zebrafish, Stickleback, Medaka and Tetraodon and the existence of "single-copy" sequences flanking the AluI repeats in *G. seta* (not shown) indicated that the AluI tandem arrays were interspersed in the genome.

Nucleotide substitutions in nuclear genes. In the second approach, 34 pairs of orthologous nuclear coding sequences (CDS) from the two species were investigated (see Supplemental Tables S1-S2). Their list is shown in Table 1 along with their sequenced sizes (49,331 bp out of a total size of 57,262 bp), the total number of non-synonymous

(NS) and synonymous (S) changes, the corrected proportion of non-synonymous (Ka) and synonymous (Ks) changes, the number of transitions (tr) and transversions (tv), and of the compositional changes (AT→GC; GC→AT). A detailed presentation of all nucleotide changes is given in Supplemental Table S3.

The data of Table 1 indicate that at least two Ka/Ks values, 0.43 (dio1) and 0.69 (E3), can be indicative of an acceleration in amino acid substitution rates, very likely due to positive selection (see ref. Tang and Wu 2006). Interestingly enough, in 18 genes (~53% of the sample) we observed a transition/transversion ratio well below the usual 2-10 range. Another remarkable feature exhibited by *G. seta* genes was the overall excess of AT→GC over GC→AT changes, this trend being especially pronounced in those genes with high Ka/Ks ratios. Note that although the direction of changes cannot be determined for every single site (due to the lack of an appropriate outgroup), this excess is unambiguously indicated by the overall figures which show a clear predominance of G and C in *G. seta* at evolutionaries variable sites. This contrasts with the normal AT bias found in vertebrate genomes (Bernardi 2004, 2007; Eyre-Walker 1999; Smith and Eyre-Walker 2001; Alvarez-Valin et al. 2002) Finally, an assessment of 5mC (5methylC) in the two genomes showed a significantly lower level (1.54%) in *G. seta* compared to *G. mirabilis* (2.20%).

Nucleotide substitutions in the mitochondrial genome. We also sequenced the mitochondrial genomes from the two fishes and observed a NS/S ratio equal to 0.15, which is not indicative of positive selection. Selection was, however, also investigated by comparing the genetic divergence at the control region, the most variable, presumed neutral mitochondrial locus, with the divergence in the protein coding genes. Out of 13 protein-coding gene pairs, 12 exhibited a higher or a similar divergence compared to the control region, whereas a lower divergence was expectedly found for all five pairs of sister fish species whose complete mitochondrial genomes were available (Fig. 3). Incidentally, preferential AT to GC changes were not observed, and most amino acid changes from *G. mirabilis* to *G. seta* were conservative hydrophobic changes or amphipathic to hydrophobic changes (see Supplemental Table S4; this assessment could not be done in the case of proteins encoded by nuclear genes because of the small number of changes). The usually conserved ribosomal genes (12S RNA and 16S RNA) also showed a higher relative rate compared to the other sister species.

DISCUSSION

The striking observations made on the nuclear and mitochondrial genomes of the two fishes can be summarized and commented upon as follows.

(i) The mitochondrial genomes of *G. seta* and *G. mirabilis* exhibit faster substitution rates in all protein-coding sequences (except for COX-1) compared to the D-loop, as opposed to the other pairs of congeneric fishes investigated (which exhibit the expected trend of faster D-loops). Based on mitochondrial cytochrome b sequences, the divergence between *G. mirabilis* and *G. seta* was originally dated at 4.6 to 11.6 Million years ago (Mya; Huang et al. 2001). Since the mutation rate in protein-coding genes is now known to be extremely fast, this divergence time certainly was a vast overestimate. A divergence assessment based on the control region, 10.4%, is likely to provide a more realistic estimate for the time of divergence between these species. Using a molecular clock based on the fish control regions (Domingues et al. 2005), the divergence time can be estimated at most at 0.66 - 0.75 Mya. Incidentally, the direction of changes is suggested by *G. seta* being a paedomorphic variant of *G. mirabilis* (Bois and Jeffreys 1999).

(ii) The amplification and/or expansion of the GC-rich AluI repeats in the *G. seta* genome is clearly responsible for the gene shift, as well as for the stabilization of the gene-rich regions. The tandem repetition of a 38 bp sequences, itself made up of shorter repeats, shows that the AluI sequence is a typical minisatellite. The instability and proneness to expansion of minisatellites (Bois and Jeffreys 1999; Richard and Paques 2000)' have apparently been favored by the temperature increase.

(iii) The excess of AT→GC over GC→AT changes (a feature particularly evident in genes with high Ka/Ks ratios) observed in *G. seta* is an indication that coding sequences in this species are also undergoing a GC increase. Yet, this increase takes places at a lower pace than the overall genomic GC increase. These different rates can be readily explained by the evolutionary intrinsic rates responsible for them, since changes in the copy number of a repetitive sequence (GC-rich AluI repeats) are more rapid than point mutations (Bois and Jeffreys 1999; Richard and Paques 2000).

(iv) The strong decrease of nuclear DNA methylation in *G. seta* compared to *G. mirabilis* fits with previous observations on the decrease of mC with increasing body temperature in fishes (Varriale and Bernardi 2006). This may open up genome regions that were locked by methylation.

(v) The genes that were sequenced can be assumed to correspond to a representative set of *Gillichtys* genes, since 12 *G. mirabilis* sequences were from GenBank and the other 56 sequences originated from our cDNAs libraries of both fishes. Here we found that the proportion of genes that are very likely under positive selection in *G. seta* appears to be high. Moreover, the observed value possibly is an underestimate due to methodological limitations.

(vi) Positive selection is classically considered to be very rare. In agreement with this view, only four genes of cichlid fishes (prime examples of adaptive radiation) out of 12,000 tested were found to be under positive selection (Salzburger et al. 2008). Examples of positive selection are however, increasing (see, for example, refs Endo et al. 1996; Bustamante et al. 2005; Voight 2006).

In conclusion, our results indicate that the action of environmental factors, in this case temperature, is accompanied by dramatic genome changes in minisatellites, in nuclear and mitochondrial coding sequences and in nuclear genome methylation. Such changes appear to be largely due to the need of responding to the thermodynamic requirements of the genome at the higher temperature and support the "thermodynamic stability hypothesis" about the genome changes accompanying the emergence of homeothermy in vertebrates (Bernardi and Bernardi 1986; Bernardi 2004, 2007). The very recent divergence of the *Gillichthys* sister species underscores the high rate at which such changes may occur. More generally, our results indicate that the environment can not only affect gene expression (Gracey et al. 2001; Buckley et al. 2006), but also mould the genome through natural selection. It has not escaped our notice that such moulding of the genome may also affect the tempo and mode of speciation of *Gillichtys*.

METHODS

Fish samples and nucleic acid preparation

G. seta was collected from Estero la Pinta, *G. mirabilis* from Estero Morhua, Sonora, Mexico (the two sites are approximately five kilometers apart). DNA and mRNA were prepared from either liver or muscle, using the method as described in Kay et al. (1952) and the Invitrogen Fast Track 2.0 Kit, respectively.

Ultracentrifugation

Analytical ultracentrifugation was performed using a Beckman Optima XL-A ultracentrifuge. GC values were calculated from the modal buoyant densities as described by Schildkraut et al. (1962). Shallow gradient ultracentrifugation was performed using a Beckman Optima XL-100K ultracentrifuge as described by DeSario et al. (1995); 200µg, 120µg and 50µg of total DNA were run according to whether fractions had to be used for subsequent analytical ultracentrifugation analysis, cloning, or hybridization analysis, respectively.

Cloning of shallow gradient fractions

Three *G. seta* DNA shallow gradient fractions averaging 44-45% GC were pooled, dialyzed against distilled water, digested for 3 hours at 37°C with restriction enzyme AluI (one of the only two restriction enzymes tested that are able to degrade repeated DNA from *G.seta*, the other one being PstI) and run in a 1.5% agarose gel. In order to obtain AluI repeats, gel slices corresponding to 400 - 2500 bp fragments were cut off and extracted using the QIAquick Gel extraction Kit by Qiagen. The purified DNA was cloned using the EcoRV site of a pBluescript plasmid. Transformation was performed on electrocompetent *E. coli* cells and positive clones were sequenced. Sequence similarity was calculated by the sequence identity matrix of the Bioedit programme.

cDNA preparation and PCR amplification

cDNAs from *G. mirabilis* and *G.seta* were synthesized on mRNAs templates using the Invitrogen Copy Kit; cDNA libraries were prepared by Invitrogen and cloned in pCMV-SPORT 6.1. PCR amplifications were performed in a 50µl final volume of 1x PCR buffer (Roche) containing also $MgCl_2$, 0.5µM primers, 0.2mMdNTP and 0.04U/µl Taq Polimerase.

Hybridization

An aliquot of each shallow gradient fraction was diluted 70 times with 0.4M NaOH to a final volume of 200µl. 100µl of each diluted fraction was dot-blotted on a positively charged nylon membrane (Hybond-N+, Amersham Pharmacia). DNA probes (either cDNAs or PCRs) were radiolabeled using the Random Oligo Labelling method and [α-^{32}P]CTP and [α-^{32}P]ATP as radioactive nucleotide precursors.

Hybridization was performed overnight at 65°C in a 1M EDTA, 0.5M Na_2HPO_4 (pH7.2), 7% SDS solution. Filters were then washed once at room temperature in a 2x SSC and 0.1% SDS solution, once at 65°C for 30 minutes in a 2x SSC and 0.1% SDS solution and for an additional 30 minutes at the same temperature in a 0.5x SSC and 0.1% SDS solution. Hybridization intensity analysis was performed using a Typhoon Trio (Amersham Biosciences) and the associate intensity quantification software.

Gene collection and analysis

This operation was very laborious and time consuming because we needed a collection of CDS as completely sequenced as possible from both species. The strategy used can be described as follows: (i) 12 complete CDS of *G. mirabilis* available in GenBank were used as templates for designing primers for PCR amplification of the orthologous CDS from *G. seta* cDNA; (ii) the other (partial) CDS were derived from our cDNA library of *G. mirabilis* (some of these sequences being also found as partial sequences in GenBank); (iii) partial CDS were sequenced, and used to find complete orthologous sequences in *Tetraodon nigroviridis*, *Takifugu rubripes*, *Danio rerio*, and *Oryzias latipes*; the heterologous fish sequences were then used to design degenerate primers for PCR amplification of the orthologous sequences present in *G. mirabilis* and *G. seta* cDNAs. In the case of eno1 gene, two paralogous genes were found. After alignment of the homologous genes of the two species, compositional changes in 1st, 2nd and 3rd codon position (→GC /→AT changes) along with NS/S were calculated. The MEGA version 4.0 program (Tamura et al. 2007) was used to calculate the rates of synonymous (Ks) and non-synonymous (Ka) substitutions of the coding region following Pamilo-Bianchi-Li method (Pamilo and Bianchi 1993; Li 1993;.

Mitochondrial genome analysis

The complete mitochondrial genomes of *G. mirabilis* and *G seta* were sequenced (see Supplemental Table S2). In order to determine if there was evidence of selection on mitochondrial genomes, we used a comparative method, similar in concept to the classical neutrality test (Hudson et al., 1987). We took a region that is presumably under no or weak selection, the non-coding control region (the D-loop), and determined the ratio of genetic divergences between a given gene and the control region between the two species. These ratios are expected to be smaller than 1, because the control region, being a non-coding region, is freer to vary than protein coding genes (see Fig.3 b).

ACKNOWLEDGEMENTS

We wish to thank our colleague Annalisa Varriale for a number of experiments including the determination of 5mC that were carried out in the early phase of this work. We also thank the Molecular Biology Service of the Stazione Zoologica Anton Dohrn for the library screening and the clone sequences of the present paper, in particular Dr. Elio Biffali, Dr. Marco Borra and Mrs. Elvira Mauriello.

FIGURE LEGENEDS

Figure 1. (**a**) Analytical ultracentrifugation profiles of DNAs from *G. mirabilis* and *G. seta*; buoyant density values were converted to GC values (see Method Summary). Optical density values at 260 nm (A260) are given on the ordinate. (**b**) Shallow gradient profiles of DNAs; in this case, abscissa values are fraction numbers. (**c-f**) Analytical ultracentrifugation profiles of shallow gradient fractions indicated by the arrows in **b**. The two vertical lines correspond to the modal buoyant density of the main peak and of the heavy fraction of *G. seta*, respectively. Blue colour corresponds to *G. mirabilis*, red colour to *G. seta*.

Figure 2. (**a**) Hybridization of *G. mirabilis* and *G. seta* cDNAs on shallow gradient fractions of both genomes. The solid curves show the A260 profiles of the shallow gradient, the dotted curves the hybridization signals. Blue colour corresponds to *G. mirabilis*, red colour to *G. seta*. (**b**) Hybridization of the *G. seta* AluI sequences on shallow gradient fractions. Again, the solid curves represent the shallow gradient A260 profile, the dotted curves the hybridization signals. The hybridization of the *G. seta* PstI sequences (another repeated sequence, whose primary structure is shown in Supplemental Fig.S2) on shallow gradient fractions of *G. seta* DNA is represented by the green dotted curve. Otherwise, as in **a** and in Fig. 1, blue colour corresponds to *G. mirabilis*; red colour to *G. seta*. Hybridization levels (radioactivity) are normalized to match the A260 profiles.

Figure 3. Histograms representing the ratios of sequence divergence between mitochondrial genes and the control region of congeneric fish species. The horizontal broken line indicates a ratio of 1. The cytochrome oxydase 2 (COX-2) gene of the tuna pair was an exception, possibly related to the unusual thermal control of these species Species pairs are *G. mirabilis/G. seta* (red), *Beryx decadactylus/B. splendens* (yellow), *Salvelinus alpinus/S. fontinalis* (pale blue), *Thunnus alalunga/T. thynnus thynnus* (peach), *Anguilla anguilla/A. rostrata* (royal blue), *Chaunax tosaensis/C. abei* (green).

Table 1. List of the orthologous coding sequences (CDS) investigated. Full and sequenced sizes of CDS (in base pairs), and the number of non-synonymous and synonymous nucleotide changes, transitions and transversions, and AT→GC, GC→AT changes from *G. mirabilis* to *G. seta* [a] are presented.

	gene	CDS size	sequenced %	NS[b]	S[b]	NS/S[b]	Ka[d]	Ks[d]	Ka/Ks	tr[c]	tv[c]	tr/tv[c]	AT→GC[e]	GC→AT[e]	AT→GC/GC→AT[e]
1	E3	1041	86	7	4	**1.8**	0.011	0.016	***0.69***	7	4	1.8	**6**	**4**	1.5
2	gygl	1018	85	8	7	**1.1**	0.011	0.036	***0.31***	8	7	1.1	**9**	**4**	2.3
3	dio1	765	82	4	3	**1.3**	0.009	0.021	***0.43***	4	3	1.3	**2**	**3**	0.7
4	Qtrtd1	1030	79	5	7	**0.7**	0.008	0.036	0.22	7	5	1.4	**5**	**4**	1.3
5	NDPK-B	450	94	1	2	**0.5**	0.003	0.014	0.21	2	1	2.0	**0**	**2**	GC→AT
6	RL27	411	94	2	4	**0.5**	0.008	0.032	0.25	5	1	5	**4**	**2**	2.0
7	RL11	537	95	2	4	**0.5**	0.006	0.022	0.27	6	0	tr	**4**	**2**	2.0
8	ampd	2199	78	2	4	**0.5**	0.002	0.009	0.22	4	2	2.0	**2**	**3**	0.7
9	S23	432	94	3	7	**0.4**	0.012	0.071	0.17	8	2	4.0	**8**	**1**	8.0
10	eno1	1299	93	3	11	**0.3**	0.003	0.039	0.08	8	6	1.3	**8**	**6**	1.3
11	eno1b	1299	93	2	11	**0.2**	0.002	0.039	0.05	8	5	1.6	**7**	**6**	1.2
12	sdhb	882	82	1	3	**0.3**	0.002	0.018	0.11	2	2	1.0	**2**	**1**	2.0
13	atp5b	1554	98	2	8	**0.3**	0.002	0.019	0.11	7	3	2.3	**1**	**8**	0.1
14	GDI	621	78	1	4	**0.3**	0.002	0.032	0.06	3	2	1.5	**2**	**2**	1.0
15	rplp1	342	92	1	6	**0.2**	0.004	0.050	0.08	5	1	5.0	**4**	**2**	2.0
16	aco2	2428	76	4	22	**0.2**	0.003	0.041	0.07	16	10	1.6	**15**	**9**	1.7
17	TPI	747	95	1	6	**0.2**	0.002	0.030	0.07	5	2	2.5	**5**	**2**	2.5
18	GADPH	1002	94	1	6	**0.2**	0.001	0.019	0.05	6	1	6.0	**4**	**3**	1.3
19	pgam	768	85	1	11	**0.1**	0.002	0.059	0.03	10	2	5.0	**11**	**1**	11.0
20	S19	444	93	0	2	S				1	1	1.0	**1**	**1**	1.0
21	GABA	369	93	0	1	S				1	0	tr	**0**	**1**	GC→AT
22	RL23	423	94	0	1	S				1	0	tr	**1**	**0**	AT→GC
23	S13	456	94	0	2	S				1	1	1.0	**1**	**1**	1
24	RL24	474	95	0	1	S				0	1	tv	**0**	**0**	-
25	NRGN	198	87	0	2	S				2	0	tr	**1**	**1**	1
26	LDH-A	999	100	0	3	S				2	1	2.0	**1**	**2**	0.5
27	pvalb1	330	93	0	1	S				0	1	tv	**1**	**0**	AT→GC
28	RXR	1114	60	0	3	S				2	1	2	**2**	**0**	AT→GC
29	S5	612	94	0	4	S				4	0	tr	**1**	**3**	0.3
30	S18	459	94	0	4	S				6	1	6.0	**2**	**5**	0.4
31	znf207	1443	63	0	4	S				2	2	1.0	**2**	**2**	1.0
32	btub	1338	92	0	14	S				11	3	3.7	**5**	**8**	0.6
33	S8	627	96	0	0	S									
34	HSPG	520	93	0	0										
total		28631	85[f]	51	172					154	71		117	89	
sequenced[g]		49331													1.3
average ratio						0.5						2.2			

a See Supplemental Tables S1-S4 for additional informations.

b NS and S indicate non-synonymous or synonymous changes, respectively. In the ratio column, S indicates the presence of only synonymous changes.

c Ka and Ks indicate the rates of non-synonymous or synonymous substitutions, respectively.

d Transitions and transversions are indicated by tr and tv. In the ratio column, tr and tv indicate the presence of only transitions and only transversions.

e AT→GC and GC→AT changes and their ratios are shown. Here AT→GC means that either T and A is observed in *G. mirabilis* and G or C in *G.seta*. Likewise, GC→AT means that G or C is observed in *G. mirabilis* and A or T in *G.seta*.

f Weight average.

g This size concerns the orthologous genes from both fishes.

REFERENCES

Alvarez-Valin, F., Lamolle, G. and Bernardi, G. Isochores, GC3 and mutation biases in the human genome. *Gene* **300:**161-8.

Barlow, G.W. 1961. Gobies of the genus Gillichthys with comments on the sensory canals as a taxonomic tool.*Copeia* **1961:** 423-437.

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

Bernardi, G. and Bernardi, G. 1990. Compositional patterns in the nuclear genome of cold-blooded vertebrates. *J. Mol. Evol.* **31:** 282–293.

Bernardi, G. 1993. The vertebrate genome. Isochores and evolution. *Mol. Biol. Evol.* **10:** 186-204.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241:** 3-17.

Bernardi, G. 2004 reprinted in 2005. *Structural and Evolutionary Genomics, Natural Selection in Genome Evolution*. Elsevier, Amsterdam.

Bernardi, G. 2007. The neoselectionist theory of genome evolution. *Proc. Natl. Aca.d Sci. U.S.A.* **104:** 8385-8390.

Bois, P. and Jeffreys, J. 1999. Minisatellites instability and germline mutation.*Cell. Mol. Life Sci*. **55:** 1636–1648.

Bucciarelli, G., Bernardi, G. and Bernardi, G. 2002. An ultracentrifugation analysis of 200 fish genomes. *Gene* **295:** 153-62.

Buckley, B.A., Gracey, A.Y. and Somero**,** G.N. 2006. The cellular response to heat stress in the goby Gillichthys mirabilis: a cDNA microarray and protein-level analysis. *J. Exp. Biol.* **209:** 2660-2677.

Bustamante, C.D. Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. 2005. Natural selection on protein coding genes in the human genome. *Nature* **437:** 1153-1157.

Cossins, A.R. and Crawford, D.L. Fish as models for environmental genomics. 2005. *Nat Rev Genet.* **6:**324-333.

De Sario, A., Geigl, E.M. and Bernardi, G. 1995. A rapid procedure for the compositional analysis of yeast artificial chromosomes.*Nucleic Acids Res*. **23:** 4013-4014.

Di Filippo, M. and Bernardi, G. 2008. Mapping Dnase-I hypersensitive sites on human isochores. *Gene* **419:** 62-65.

Domingues, V.S., Bucciarelli, G., Almada, V.C. and Bernardi, G. 2005. Historical colonization and demography of the Mediterraneandamselfish, Chromis chromis. *Mol Ecol.* **14:** 4051-63.

Endo, T., Ikeo, K. and Gojobori, T. 1996. Large-Scale Search for Genes on Which Positive Selection May Operate. *Mol. Biol. Evol.* **13:** 685-690.

Eyre-Walker, A. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152:** 675-683.

Fields, P.A., Kim, Y.S., Carpenter, J.F. and Somero, G.N. 2002. Temperature adaptation in *Gillichthys* (Teleost: Gobiidae) A_4-lactate dehydrogenase: identical primary structures produce subtly different conformations. *J. Exp. Biol.* **205:** 1293–1303.

Gracey, A.Y., Troll, J.V., and Somero, G.N. 2001. Hypoxia-induced gene expression profiling in the euryoxic fish Gillichthys mirabilis. *Proc. Natl. Aca.d Sci. U.S.A.* **98:** 1993-1998.

Hochachka, P.W. and Somero, G.N. 2002. *Biochemical adaptation: mechanism and process in physiological evolution***.** pp.290-312. Oxford University Press, Oxford; New York,

Huang, D. and Bernardi, G. 2001.Disjunct Sea of Cortez - Pacific Ocean *Gillichthys mirabilis* populations and the evolutionary origin of their paedomorphic relative, *Gillichthys seta*. *Mar. Biol.* **138:** 421-428.

Hudson, R. R., Kreitman, M. and Aguadé, M. 1987. A test of neutral molecular evolution based on nucleotide data.Genetics **116:** 153–159.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and non-synonymous substitution. *J. Mol. Evol.* **36:** 96-99 .

Kay, E.R.M., Simmons, N.S. and Dounce, N.L. 1952.An improved preparation of sodium desoxyribonucleate. *J. Am. Chem. Soc.* **74:** 1724–1726.

Macaya, G., Thiery, J.P. and Bernardi, G. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* **108**: 237–254.

Nelson, J. S. 2006. *Fishes of the world*. John Wiley and Sons, Inc. New York.

Pamilo, P. and Bianchi, N.O. 1993. Evolution of the Zfx and Zfy, genes: Rates and interdependence between the genes. *Mol. Biol. Evol.* **10:** 271-281.

Richard, G. and Paques, F. 2000. Mini- and microsatellite expansions: the recombination connection. *EMBO reports* **1:** 122-126.

Saccone, S., Federico, C. and Bernardi, G. 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* **300:** 169-178.

Salzburger, W., Renn, S., Steinke, D., Braasch, I., Hofmann, H.A. and Meyer A. 2008. Annotation of expressed sequence tags for the East African cichlid fish Astatotilapia burtoni and evolutionary analyses of cichlid ORFs. *BMC Genomics.* **9:** 96.

Schildkraut, C.L., Marmur, J. and Doty P. 1962. Determination of the base composition of DNA from its buoyant density in CsCl. *J. Mol. Biol.* **4:** 430-443.

Smith, N.G. and Eyre-Walker, A. 2001. Nucleotide substitution rate estimation in enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. *Mol Biol Evol.* **18:** 2124-2126.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2004. *MEGA4*: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol. Evol.* **24:** 1596-1599.

Tang, H. and Wu, C. 2006.A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. *Mol. Biol. Evol.* **23:** 372-379.

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

Varriale, A. & Bernardi, G. 2006. DNA methylation and body temperature in fishes. *Gene* **385:** 111-112.

Voight, F., Kudaravalli, S., Wen, X. and Pritchard, J.K. 2006. A Map of Recent Positive Selection in the Human Genome *Plos Biology* **4:** e72.

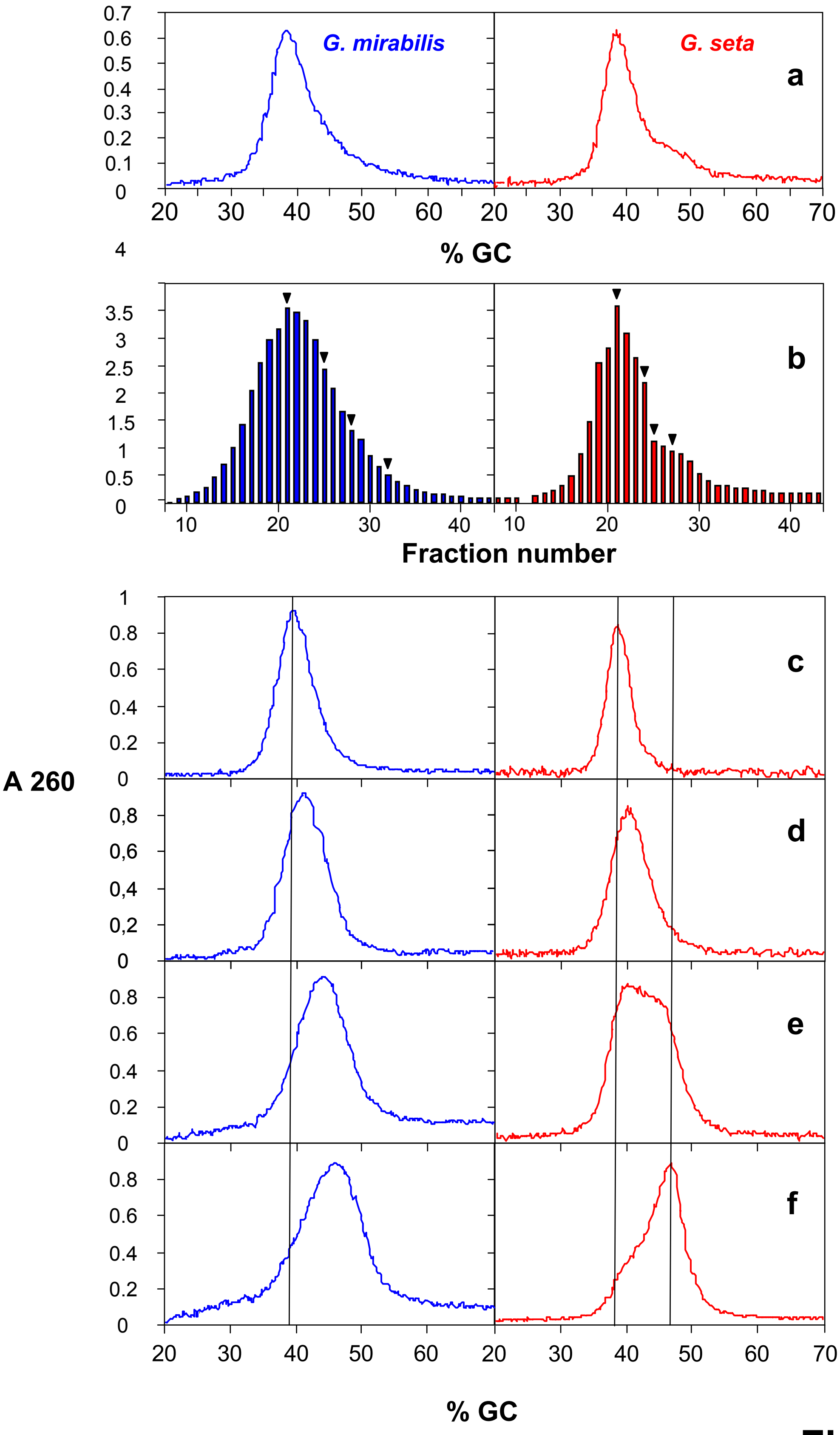
G. mirabilis
G. seta
a
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70
% GC
b
4
3.5
3
2.5
2
1.5
1
0.5
10
20
30
40
Fraction number
A 260
c
d
e
f
1
0.8
0.6
0.4
0.2
0,8
0,6
0,4
0,2
% GC

Fig.1

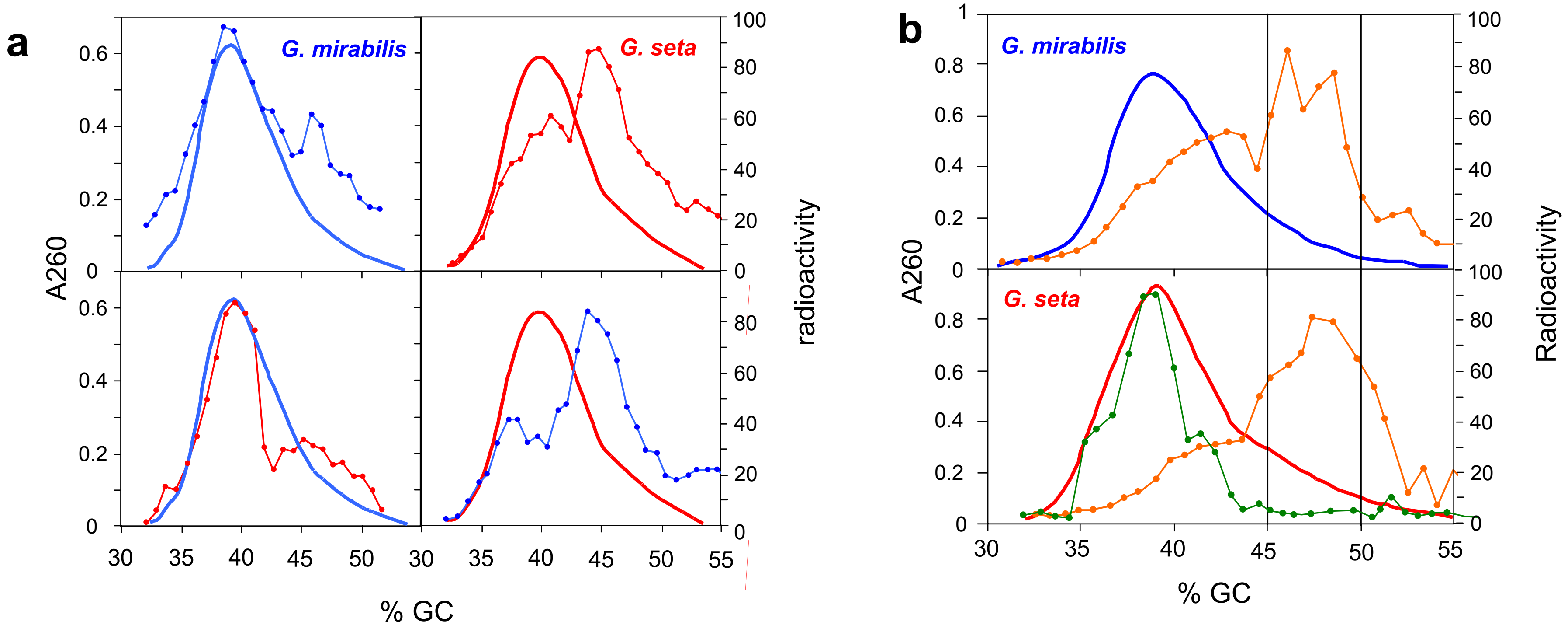
a
G. mirabilis
G. seta
A260
radioactivity
% GC
0
0.2
0.4
0.6
0
20
40
60
80
100
30
35
40
45
50
55
b
G. mirabilis
G. seta
A260
Radioactivity
% GC
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100

Fig.2

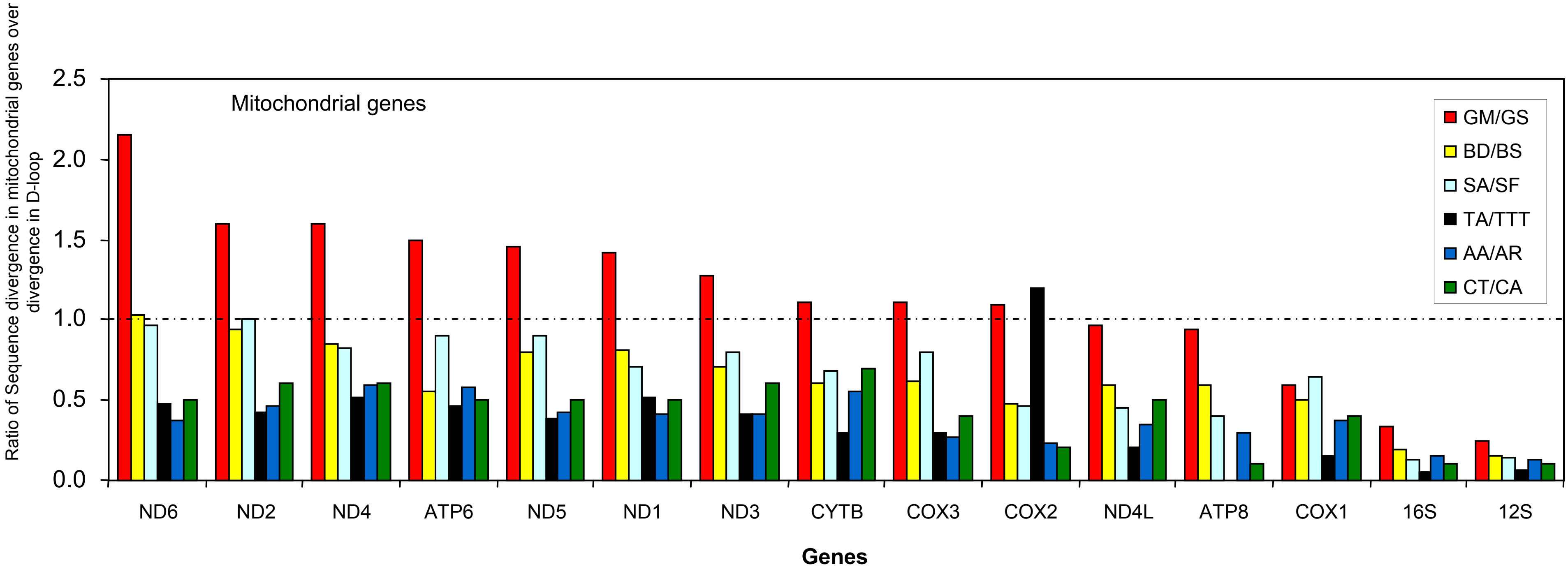
Mitochondrial genes
Ratio of Sequence divergence in mitochondrial genes over divergence in D-loop
2.5
2.0
1.5
1.0
0.5
0.0
GM/GS
BD/BS
SA/SF
TA/TTT
AA/AR
CT/CA
ND6
ND2
ND4
ATP6
ND5
ND1
ND3
CYTB
COX3
COX2
ND4L
ATP8
COX1
16S
12S
Genes

Fig. 3