Testing the Metabolic Rate Hypothesis by Analyzing Vertebrate Genomes

A thesis submitted to
University of Naples "Federico II", Naples, Italy
for the degree of

Doctor of Philosophy

In
"**Computational Biology and Bioinformatics**"
XXIV cycle

by

Ankita Chaurasia
November, 2011

University of Naples “Federico II”, Naples, Italy

Research Doctorate in
Computational Biology and
Bioinformatics
XXIV cycle

Ankita Chaurasia

The research activities described in this thesis were performed at Laboratory of Animal Evolution and Physiology, Group Genome Organization and Evolution, Stazione Zoologica Anton Dohrn (SZN), Naples, Italy.

Tutor
Dr. Giuseppe D’Onofrio

Co-tutors
Dr. Diego di Bernardo
Prof. Gennaro Miele

Doctoral Coordinator
Prof. Sergio Cocozza

November 30, 2011

To Ma and Papa

Abstract

The nature of the forces driving the genomic GC content of both prokaryotes and eukaryotes is a matter of debate among evolutionists. At present several hypotheses have been proposed: the mutational bias, the biased gene conversion (BGC), the thermal stability and the metabolic rate. The thesis mainly focused on testing the role played by the metabolic rate in shaping the base composition of genomes of vertebrates, namely: teleosts and mammals. Interesting results were also obtained analyzing non-vertebrate genomes (tunicates).

Focusing on teleostean fish, the mass specific routine metabolic rate temperature-corrected using the Boltzmann's factor (MR) and base composition of genomes (GC%) were re-examined and related with their major habitat: polar, temperate, sub-tropical, tropical and deep-water. Fish of the polar habitat showed the highest MR and that of temperate fish was significantly higher than that of tropical one, showing the lowest average value. The GC% of polar and temperate fish both showed significantly higher values than that of tropical and sub-tropical fish. Plotting MR *vs.* GC%, a significant correlation was found. The deamination process, transforming the 5-methylcytosine (5mC) to thymine, and thus 5mCpG doublets into the derivative ones, i.e. TpG and CpA, is well known to affect the genomic GC content and to be temperature dependent. The 5mC level was reported to decrease from polar to tropical fish gnomes. The frequencies of CpG, TpG and CpA in five teleostean genomes living in different habitat excluded the temperature effect of 5mC on the genomic GC content in fish, further supporting a link between environmental metabolic adaptation and genome base composition.

Several observations reported that GC-rich genes preferably arbor short non-coding sequences. A comparative analysis of orthologous introns (assigned throughout gene orthology) among five sequenced teleostean genomes, was carried out. The preliminary results highlighted a link between GC content and intronic length, hence supporting the energetic cost on transcriptional activity hypothesis.

Human genes were assigned to three large functional categories according to the KOG database: information storage and processing, cellular processes and signaling, and metabolism. The GC3 level was significantly increasing from the former to the latter. This specific compositional pattern was found, as footprint, in all mammalian genomes analyzed, but not in frog and lizard ones. In the same comparative analysis among vertebrate genomes, it was found that human genes involved in the metabolic processes underwent to the highest GC3 increment.

Compositional analysis of tunicate genomes showed that *C. savignyi* is GC-richer than *C. intestinalis*. Interestingly, preliminary data showed a same trend for the oxygen consumption.

In conclusion, the data produced in the present thesis, analyzing available vertebrate and non-vertebrate genomes, all converged towards evidences supporting the metabolic rate as one of the key forces driving the base composition variability observed among living organisms. Natural (negative) selection may essentially explain the GC variability among organisms.

Acknowledgements

This dissertation is the result of three years of research at Stazione Zoologica 'Anton Dohrn', Naples, Italy. I would like to express my sincere gratitude to many people whose continual support, both professional and personal, has made this dissertation possible.

First and foremost, I would like to thank my advisor Dr. Giuseppe D'Onofrio for his mentoring and continuous support throughout these years. His commitment towards science and innovative ideas has been a source of inspiration. He motivated me to work hard and follow ethical practices towards research. I regard his influence as the single biggest factor contributing to my growth from a student to a researcher.

I am equally thankful to my co-tutors Dr. Diego di Bernardo and Prof. Gennaro Miele for their insightful comments during the course of my research. I would like to thank Prof. Claudio Agnisola for his close supervision especially regarding metabolic rate calculation. He afforded me the opportunity to work together in his lab (Department of Biological sciences, UNINA) as well as during my stint at the University of California Santa Barbara. My sincere gratitude to Dr. Claudia Angelini, IAC-CNR for helping me with statistical analysis, which proved to be an important tool for conducting my research.

My special acknowledgement to external collaborators Dr. Erin Newman and Prof. William Smith, University of California Santa Barbara, California for offering me the privilege of working in their lab.

I am grateful to Prof. Giovanni Miano and Prof. Sergio Cocozza, the former and present Coordinator of the Research Doctorate in Computational Biology and Bioinformatics (UNINA) respectively, for their support and guidance, since my arrival in Naples. I would also like to thank other members of the PhD committee for accepting my candidature as a prospective doctoral student in 2009. I wish to extend my sincere gratitude to all researchers and staff members of SZN.

I am thankful to Mr. Guido Celentano, Secretary Doctorate School, for his administrative assistance.

Many thanks to Dr. Luisa Berná for all the stimulating discussions and answering the silliest questions. Special thanks to Dr. Erminia Uliano for sharing her expertise on fish and metabolic rate experiments. More so for the memorable time I spent with her during sample collection and visit to UCSB, CA. The collaboration-through-caffeine research methodology of Mr. Salvatore Bocchetti and Andrea proved surprisingly helpful.

I am highly indebted to my former research supervisors Prof. Ashok Kumar Gupta and Prof. Dwijendra K. Gupta, University of Allahabad, India without whom I would never have developed an interest in Bioinformatics.

I have spent an exciting and pleasant three years of research at SZN. I cannot sufficiently thank my colleagues: Chiara, Daniela, Vasco, Leo, Claudia, Elena, Lenina, Susy and Ida for their love and support. The fun-filled days and lunch/coffee break discussions will forever remain etched in memory. Thank you, Nicole for bringing in the high-glucose treats that kept me going. Thanks a lot for everything and especially for a beautiful friendship.

My colleagues and friends were my surrogate family away from home. Their love and care never let homesickness dampen my spirits. Thanks Atul for a wonderful friendship and for all emotional support during stressful times. There are more than a few names to mention: Dr. Pravin, Naresh bhaiya, Bhavna, Chingoi, Ankush, Deepak, Swaraj, Gauri and my flat mates Rosa and Alessia with whom I shared a special personal bond.

Very special thanks to Saurabh.

I would like to share this great moment of my life with my family. I find myself at a loss of words while expressing my appreciation for my parents for their unconditional love, support and prayers. Their persistent confidence in me has always been my greatest strength. Special thanks to dear Mausa ji and Mausi ji for their guidance and care. My brother is the person who mentored me ever since I was a child. Thanks a lot bhaiya, bhabhi and Vihaan. Thanks to Shilpi didi, Amitava jeejaji, Preeti, Rishika, Shivam and Shobhit for constant love and care.

Finally, I thank Gaurav for giving me the life I ever dreamed.

Contents

List of figures

List of tables

List of Supplementary materials

List of Abbreviations

AT	Molar ratio of **a**denine + **t**hymine in DNA
BGC	**B**iased **G**ene **C**onversion
bpi	**b**ase **p**air in introns
BM	**B**ody **m**ass
CDS	**C**o**D**ing **S**equence: region of nucleotides that corresponds to the sequence of amino acids in the predicted protein
CpG	**c**ytosine occurring next to a **g**uanine, in the linear sequence of bases along DNA, separated by one phosphate that links the two nucleosides together
CpA	**c**ytosine occurring next to a **a**denine, in the linear sequence of bases along DNA, separated by one phosphate that links the two nucleosides together
COG	**C**luster of **O**rthologous **G**roups of proteins
Fish	In order to avoid confusion between Fish and Fishes, Fish was used
3′GC	Molar ratio of **g**uanine and **c**ytosine at **3′** of the CDS
5′GC	Molar ratio of **g**uanine and **c**ytosine at **5′** of the CDS
GC	Molar ratio of **g**uanine + **c**ytosine in DNA
GC1	Molar ratio of **g**uanine and **c**ytosine at the **first** codon position
GC2	Molar ratio of **g**uanine and **c**ytosine at the **second** codon position
GC1+2	Molar ratio of **g**uanine plus **c**ytosine at **first** and **second** positions
GC3	Molar ratio of **g**uanine and **c**ytosine at the **third** codon position
GCcds	Molar ratio of **g**uanine + **c**ytosine of **co**d**i**ng **s**equences
GCg	Molar ratio of **g**uanine + **c**ytosine of the whole **g**enome
GCi	Molar ratio of **g**uanine and **c**ytosine in the **i**ntronic sequences
HACNS	**h**uman-**a**ccelerated **c**onserved **n**on-coding **s**equences
HAR	**h**uman-**a**ccelerated **r**egion
kb	**k**ilo **b**ase pairs = 1,000 of DNA
KOG	Eu**k**aryotic clusters of **O**rthologous **G**roups of proteins
LDH	**L**actate **de**h**y**drgenase
MR	mass specific routine **m**etabolic **r**ate, temperature-corrected using the Boltzmann's factor
MTE	**M**etabolic **T**heory of **E**cology
RBH	**R**eciprocal **B**est **H**its
SNP	**S**ingle-**n**ucleotide **p**olymorphism
TpG	**t**hymine occurring next to a **g**uanine, in the linear sequence of bases along DNA, separated by one phosphate that links the two nucleosides together
5mC	**5-M**ethyl**c**ytosine: a methylated form of the DNA base cytosine
UTD	**U**niversal **T**emperature **D**ependence

1 Introduction

One of the basic questions of genome evolution is centered around the processes affecting the base compositional variability among genomes. The first evidence that the amount of the "four bricks" of a DNA molecule, split by the Chargaff's rule in the AT and GC pairs, were not in a constant ratio among genomes came out from the pioneering studies of Sueoka (1959, 1962). Using the cesium chloride (CsCl) centrifugation technique described by Meselson, Stahl and Vinograd (1957), Sueoka first observed, indeed, that the DNA base composition (generally defined as GC%, i.e. the molar ratio of guanine plus cytosine) greatly vary among bacteria. This result was the starting point of the neutralist-selectionist debate on the nature of forces driving the base composition of a genome. Until now the question remains the "heart of the problem" in the field of molecular evolution, because understanding the mechanisms affecting nucleotide substitution in DNA are fundamental to comprehend evolutionary biology, population genetic and mutation research. Mutation, indeed, is the ultimate source of genetic variation, or, to put it another way, one of the fundamental forces of evolution enabling evolutionary changes (Barton, 2010; Loewe and Hill, 2010).

.

1.1 Causes of compositional change

Regarding the compositional variation of genomes, there is still an open debate between internal and external forces. The former are mainly based on stochastic events arising during intracellular processes, such as DNA duplication, repair, and recombination. The latter takes into account the role of adaptive processes resulting from the interaction of the organism with the surrounding environment.

Based on internal forces, two main hypotheses had been proposed, namely: Mutational Bias Hypothesis (Sueoka, 1962,1988) and Biased Gene Conversion (Eyre-Walker, 1993; Eyre-Walker and Hurst, 2001; Galtier, Piganeau et al., 2001; Galtier and Duret, 2007). Whereas, in the frame of the selectionist point of view, the proposed adaptive hypotheses focus on two important environmental factors: temperature (Bernardi, Olofsson et al., 1985) and metabolic rate (Vinogradov, 2001, 2005).

1.1.1 Mutational Bias hypothesis

In the frame of the neutral theory, quantitative theory of mutational bias hypothesis was first proposed to explain the great variation of the genomic GC content among different bacteria, and later on extended to high vertebrates (Sueoka, 1962,1988; Lobry and Sueoka, 2002). According to this, the major reason that brings the change in GC content of an organism is the mutation between an α pair and a **γ** pair (details in Box.1.1). Change in the equilibrium has directionality, i.e. directional mutational pressure, towards increment or decrement of GC content.

Box 1.1: Theory of Directional Mutation Pressure.

Existence of non-random overall mutation pressure namely, directional mutation pressure toward the ***α*** pair (A-T or T-A) or toward the **γ** pair (G-C or C-G), was first suspected from the variation of DNA base composition [expressed as G + C content, **γ/ (α + γ)]** among DNAs of different bacterial species.

The major cause for a change in DNA G + C content of an organism is the mutation between an ***α*** pair and a ***γ*** pair. When there are no selective constraints

$$\underset{(p)}{\gamma} \underset{v}{\overset{u}{\rightleftharpoons}} \underset{(1-p)}{\alpha},$$

where, ***p*** is the fractional G + C content of DNA and ***u*** and ***v*** are mutation rates per generation per base. At equilibrium

$$\hat{p} = \frac{v}{u+v}, \text{ or } \frac{v}{u} = \frac{\hat{p}}{1-\hat{p}}.$$

Here, ***p*** is the G + C content at equilibrium. The large variation of G + C content among DNAs of different bacteria was explained mainly by differences in ***v***/***u*** rather than by selection.

Directional Mutation Pressure: A definition of directional mutation pressure is necessary for the quantitative analysis of the effect of directional mutation on the G + C content of DNA. The directional mutation pressure (***μ_D***) is now defined as

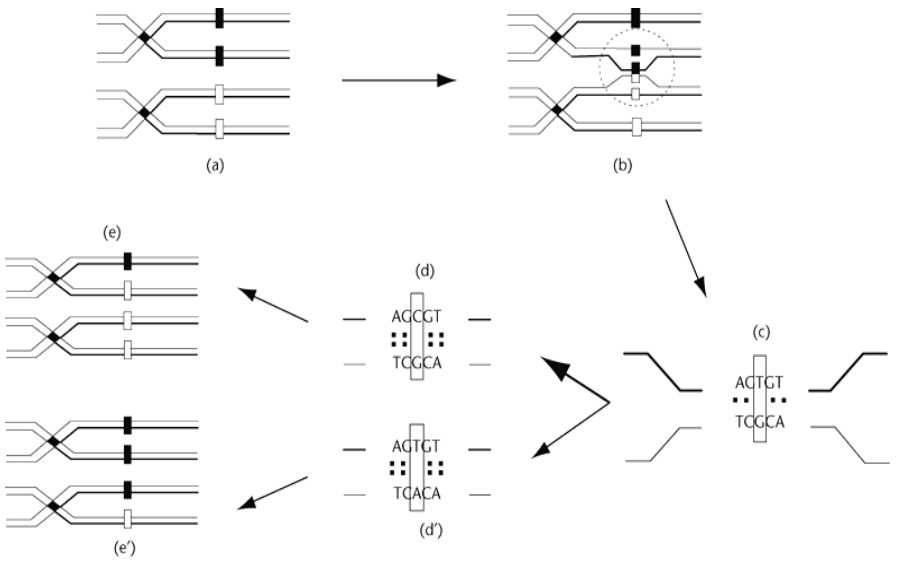
$$\mu_D = \frac{v}{u+v}.$$

Here, ***μ_D*** > 0.5 indicates that the mutation pressure favors y over at and ***μ_D*** < 0.5 indicates a preference for ***α*** over ***γ***. At equilibrium, the G + C content of a neutral nucleotide position equals mutation pressure ***D***; ***p*** is therefore termed as the G + C content at "mutational equilibrium."

1.1.2 Biased Gene Conversion

In the same frame, the biased gene conversion hypothesis (BGC), essentially based on the synergy between recombination events and biased DNA repair (Eyre-Walker, 1993; Galtier and Duret, 2007; Duret and Galtier, 2009), was first proposed to explain the formation of GC-rich isochores in mammals, and afterward extended to prokaryotes, since the finding of a GC biased repair system in them (Birdsell, 2002). Gene conversion is a molecular mechanism associated with recombination in which a genomic fragment is "copied/ pasted" onto another homologous fragment. Both DNA fragments therefore share identical sequences after a conversion event. Allelic conversion occurs during the process of meiotic recombination. A DNA heteroduplex is formed during a recombination event, involving the plus strand of one chromosome and the minus strand of the sister chromosome (Box 1.2, b). If this region of heteroduplex includes a heterozygous site, then a mismatch will occur, e.g. T:G mismatch in the heteroduplex (Box 1.2, c). This mismatch may be recognized and repaired either to C:G (Box 1.2, d) or A:T (Box 1.2, d'), resulting in non-Mendelian segregation of alleles (Box 1.2, e, e'), with the two possible ways of repairing a mismatch having equal probabilities.

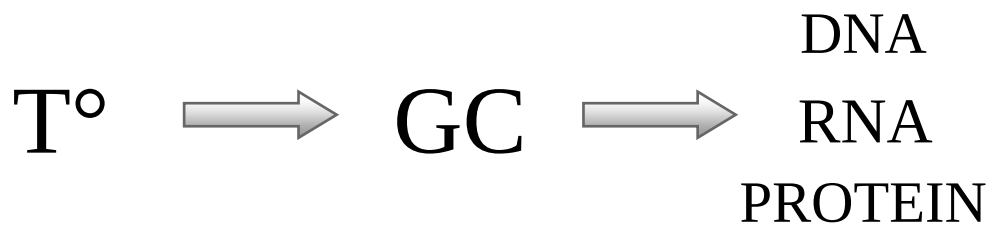
Box 1.2: Gene conversion during a recombination event. Solid rectangles, A:T pair; Open rectangles, G:C pair.



1.1.3 Thermodynamic stability hypothesis

According to the thermodynamic stability hypothesis, an increment of the environmental or body temperature favors a GC increment (Bernardi, Olofsson et al., 1985). The hypothesis was based on two main points: i) the fact that DNA and RNA structures are both stabilized by the increment of the GC pairs linked by a triple hydrogen bond; and ii) on the observation that proteins are stabilized by the increment of the hydrophobicity, through the increment of hydrophobic amino acids mainly encoded by GC-rich codons (D'Onofrio, Jabbari et al., 1999).

Box 1.3: Representation summarizing thermodynamic stability hypothesis.



1.1.4 Metabolic Rate Hypothesis

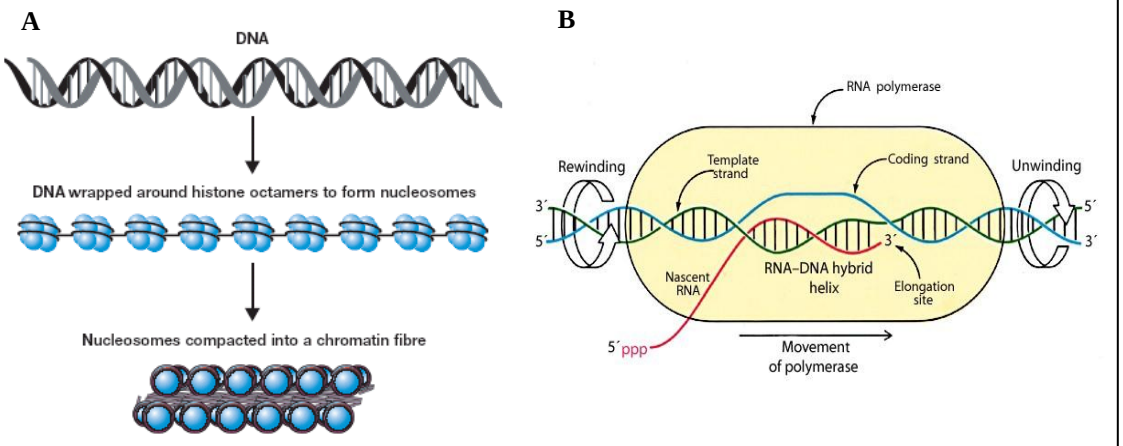
The metabolic rate hypothesis was based on the observation that two DNA physical properties, Bendability (Vinogradov, 2001) and Nucleosome Formation Potential (Vinogradov, 2005), were both significantly correlated with the GC%. More precisely, a decrement of the nucleosome formation potential (triggering the unfolding of the chromatin structure) paralleled by an increment of the DNA bendability (throughout the transcriptional process) should favor GC-rich genes. The hypothesis was supported by experimental data showing: i) the correlation between chromatin structure and GC content (Saccone and Bernardi, 2001); and ii) the higher transcriptional activity of the GC-rich genes (Versteeg, van Schaik et al., 2003; Arhondakis, Auletta et al., 2004).

Nucleosome Formation Potential (NFP)

NFP is the probability of DNA to wrap around histone protein (Box 1.4, A). A negative correlation holds between the NFP and the GC content of exons, intergenic spacers and

introns of human genes, (Vinogradov, 2005). Hence, the increase of the GC content associated with the reduction of NFP, should unfold the chromatin structure, and consequently facilitate the binding of transcription factors. The negative correlation between NFP and GC% was in very good agreement with the data of Saccone and Bernardi (2001) and Versteeg and van Schaik (2003). Indeed, *in-situ* hybridization of GC-rich and GC-poor chromosome probes showed that, in the nuclei of human blood lymphocyte, the former hybridized with a chromatin region localized in the center of the nucleus and characterized by a spread out confirmation, while the latter hybridized with a chromatin region squeezed at the periphery of the nucleus and characterized by a condensed structure (Saccone and Bernardi, 2001). Further, Versteeg and colleagues (2003) showed the chromosomal gene expression profiles of 62 SAGE libraries of at least 50,000 transcript established by the human transcriptome map revealing a clustering of highly expressed genes which was also marked by a high GC content (Versteeg, van Schaik et al., 2003).

Box 1.4: Cartoon representation of the two DNA physical properties: (A) NFP; (B) Bendability.



DNA Bendability

As very clear from its name Bendability (bend-ability) is the property of being easily bent without breaking. The DNA helix should be often bent and unbent in its transition from

packaged to extended state to comply during the transcriptional process (Box 1.4, B). DNA bendability of human exons and introns was found to be increasing, seemingly faster than that of the random sequences, and it correlates positively with the GC-content in both exons and introns. This correlation was stronger in case of the introns as compared with exons. The author concluded that GC-enrichment of both coding and non-coding regions can be favored by selection for increased bendability of DNA (Vinogradov, 2001).

1.2 How universal are the forces in shaping the base composition of genomes?

In the frame of the neutral theory, mutational bias was first reported behind the great variability of DNA base composition among bacterial genomes, analyzed using density gradient centrifugation (Sueoka, 1959,1962). After the availability of sequence data, the hypothesis was extended to all living organisms (Sueoka, 1988). However, further analysis of the mutation pattern and of human polymorphism data sets (Lander, Linton et al., 2001) showed an excess of GC ⇒ AT mutations. Bearing in mind the transitional mode of evolution (Figure 2.1), according to which during the evolution from cold- to warm-blooded vertebrates the GC content increased, then the excess of GC ⇒ AT mutations hardly explains the formation of the GC-rich isochores in the warm-blooded vertebrates (Bernardi, 2004, for review). Regarding bacteria, very recent analysis showed the existence of an AT mutational bias (Hildebrand, Meyer et al., 2010), casting doubts about the Sueoka's hypothesis, because not explaining the evolutionary steps leading to the formation of GC rich bacterial genomes.

In the same frame, the biased gene conversion hypothesis (BGC) was grounded on a significant correlation between GC content and re-combination process. The shortcoming of the BGC hypothesis is that, in spite of the fact that the genome structure appears to be well conserved among mammalian genomes (i.e. isochore organization) (Costantini, Cammarano et al., 2009), hot-spot recombination sites were reported to be species like chimpanzee and human (Ptak, Hinds et al., 2005). Moreover, the size of the

not phylogenetically, even in closely related species like chimpanzee and human (Ptak, Hinds et al., 2005). Moreover, the size of the elements known to evolve according to the BGC, namely HARs and HACNSs, is within the range of 200-1000 bp (Duret and Galtier, 2009), an order of magnitude far from the mega-bases of the isochores (Bernardi, Olofson et al., 1985). In case of yeast the AT/GC substitution pattern was "not correlated with recombination, indicating that GC content is not driven by recombination" (Marsolier-Kergoat and Yeramian, 2009).

Biased gene conversion was also proposed to explain the base compositional variability among bacterial genomes. However, studies on single nucleotide polymorphism (SNP) at synonymous positions, showed an excess of GC ⇒ AT polymorphism, without evidences of recombination (Hildebrand, Meyer et al., 2010).

Hence, if one side mutational bias and BGC failed to be supported by the analysis of bacterial genomes, the thermal stability hypothesis (Bernardi, Olofson et al., 1985) and the metabolic rate hypothesis (Vinogradov, 2001, 2005) first proposed for high vertebrates have been called into question also to explain the GC% variability among bacterial genomes (Naya, Romero et al., 2002; Rocha and Danchin, 2002; Musto, Naya et al., 2006). The thesis focuses on these two adaptive hypotheses.

1.3 Aim and strategies of thesis

The thesis has been designed to gain insight on the external forces playing the most relevant evolutionary role in driving the base compositional evolution as proposed in the frame of the thermal stability (Bernardi, Olofsson et al., 1985) and the metabolic rate hypothesis (Vinogradov, 2001, 2005). Hence, comparative genome analyses using different approaches and strategies, planned by available data, were designed to analyze the compositional pattern of vertebrate genomes. We decided to split aquatic from terrestrial organisms. The rationale of the choice grounded on two considerations. First, aquatic and terrestrial organisms are characterized by two different mode of genome evolution, i.e. the shifting and the transitional mode (**Chapter 2: "Compositional study of vertebrate genomes"**). Second, aquatic organisms, differently from terrestrial ones, live in an environment where the available oxygen, dictated by the Henry's law, is a limiting factor.

Regarding aquatic organisms, data on temperature (T°) metabolic rate (MR) and genomic base composition (GC%) were collected for ~200 bony fish (teleosts) that were grouped in the five major habitats: polar, temperate, subtropical, tropical and deep water. To compare the metabolic rate of fish living in different habitats, avoiding the effect of temperature, the Boltzmann correction was applied (**Chapter 2: Section 2.2**). To assess the role played by the CpG dinucleotide through the methylation/deamination process on the genomic GC% of fish living at different temperature five completely sequenced genomes were analyzed, namely: *D. rerio*, *O. latipes*, *T. rubripes*, *G. aculeatus* and *T. nigroviridis*. (**Chapter 2: Section 2.3**). The above cited genomes were further investigated in order to shed light on current hypotheses linking the length of the intronic sequences with the energetic cost of the transcriptional activity (**Chapter 2: Section 2.4**).

Regarding terrestrial organisms, comparative analyses were carried out: i) among thirteen mammalian genomes, in order to highlight common conserved features; and ii) between mammals and amphibians, as well as reptiles, in order to highlight genomic differences between poikilotherms and homeotherms. The approach focused on available data about gene functional classes. Indeed, according to the KOG database gene functional classes are grouped in three main categories: i) information storage and

processing; ii) cellular processes and signaling; and iii) metabolism. The compositional analysis of the coding sequences was directed to understand if genes involved in metabolic processes hold peculiar compositional features. (**Chapter 3: "Functional organization of mammalian genomes"**).

Finally, preliminary analyses were carried out on two completely sequenced invertebrate genomes, *C. intestinalis* and *C. savignyi*. Compositional and metabolic rate analyses were carried out. (**Chapter 4: "Compositional study of tunicate genomes"**).

Interestingly, all different approaches converged towards the conclusion that metabolic rate play a crucial role in shaping the genome base composition.

1.4 References

Arhondakis, S., F. Auletta, G. Torelli and G. D'Onofrio (2004). "Base composition and expression level of human genes." Gene **325**: 165-169.

Barton, N. H. (2010). "Mutation and the evolution of recombination." Philos Trans R Soc Lond B Biol Sci **365**(1544): 1281-94.

Bernardi, G. (2004). Sructural and Evolutionary Genomics. Natural Selection in Genome Evolution. Amsterdam, Elsevier.

Bernardi, G., B. Olofson, J. Filipski, M. Zerial, et al. (1985). "The mosaic genome of warm-blooded vertebrates." Science **228**: 953-958.

Birdsell, J. A. (2002). "Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution." Mol Biol Evol **19**(7): 1181-97.

Costantini, M., R. Cammarano and G. Bernardi (2009). "The evolution of isochore patterns in vertebrate genomes." BMC Genomics **10**: 146.

D'Onofrio, G., K. Jabbari, H. Musto and G. Bernardi (1999). "The correlation of protein hydropathy with the base composition of coding sequences " Gene **238**(1): 3-14.

Duret, L. and N. Galtier (2009). "Biased gene conversion and the evolution of mammalian genomic landscapes." Annu Rev Genomics Hum Genet **10**: 285-311.

Duret, L. and N. Galtier (2009). "Comment on "Human-specific gain of function in a developmental enhancer"." Science **323**(5915): 714; author reply 714.

Eyre-Walker, A. (1993). "Recombination and mammalian genome evolution." Proc. R. Soc. Lond. **B 252**: 237-243.

Eyre-Walker, A. and L. D. Hurst (2001). "The evolution of isochores." Nat Rev Genet **2**(7): 549-555.

Galtier, N. and L. Duret (2007). "Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution." Trends Genet **23**(6): 273-277.

Galtier, N., G. Piganeau, D. Mouchiroud and L. Duret (2001). "GC-content evolution in mammalian genomes: the biased gene conversion hypothesis." Genetics **159**(2): 907-911.

Hildebrand, F., A. Meyer and A. Eyre-Walker (2010). "Evidence of selection upon genomic GC-content in bacteria." PLoS Genet **6**(9).

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lobry, J. R. and N. Sueoka (2002). "Asymmetric directional mutation pressures in bacteria." Genome Biol **3**(10): RESEARCH0058.

Loewe, L. and W. G. Hill (2010). "The population genetics of mutations: good, bad and indifferent." Philos Trans R Soc Lond B Biol Sci **365**(1544): 1153-67.

Marsolier-Kergoat, M. C. and E. Yeramian (2009). "GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome." Genetics **183**(1): 31-8.

Meselson, M., F. W. Stahl and J. Vinograd (1957). "Equilibrium Sedimentation of Macromolecules in Density Gradients." Proc Natl Acad Sci U S A **43**(7): 581-8.

Musto, H., H. Naya, A. Zavala, H. Romero, et al. (2006). "Genomic GC level, optimal growth temperature, and genome size in prokaryotes." Biochem Biophys Res Commun **347**(1): 1-3.

Naya, H., H. Romero, A. Zavala, B. Alvarez, et al. (2002). "Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes." J Mol Evol **55**(3): 260-4.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, et al. (2005). "Fine-scale recombination patterns differ between chimpanzees and humans." Nat Genet **37**(4): 429-34.

Rocha, E. P. and A. Danchin (2002). "Base composition bias might result from competition for metabolic resources." Trends Genet **18**(6): 291-4.

Saccone, S. and G. Bernardi (2001). "Human chromosomal banding by in situ hybridization of isochores." Methods Cell Sci **23**(1-3): 7-15.

Sueoka, N. (1959). "A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation." Proc. Natl. Acad. Sci. USA **45**: 1480-1490.

Sueoka, N. (1962). "On the genetic basis of variatrion and heterogeneity of DNA base composition." Proc. Natl. Acad. Sci. USA **48**: 582-592.

Sueoka, N. (1988). "Directional mutation pressure and neutral molecular evolution." Proc Natl Acad Sci U S A **85**(8): 2653-7.

Versteeg, R., B. D. van Schaik, M. F. van Batenburg, M. Roos, et al. (2003). "The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes." Genome Res **13**(9): 1998-2004.

Vinogradov, A. E. (2001). "Bendable genes of warm-blooded vertebrates." Mol Biol Evol **18**(12): 2195-200.

Vinogradov, A. E. (2005). "Noncoding DNA, isochores and gene expression: nucleosome formation potential." Nucleic Acids Res **33**(2): 559-63.

2 Compositional study of vertebrate genomes

2.1 The state of the art about vertebrate genome evolution

Two modes of evolution have been found analyzing vertebrate genomes: the transitional and the shifting (Bernardi, 1990; D'Onofrio, Jabbari et al., 1999; Cruveiller, D'Onofrio et al., 2000). These two different modes were first observed by the cesium chloride (CsCl) ultracentrifugation technique (Thiery, Macaya et al., 1976), and later on confirmed by the comparisons of completely sequenced genomes.

2.1.1 Transitional Mode It was mainly observed while comparing the "cold- with warm-blooded vertebrates" (terminology first used by Bernardi and Bernardi, 1986) to stress the fact that, considering temperature, both environmental and body temperature should be taken into account). Precisely, during the transition from cold- to warm-blooded vertebrate, an increment in heterogeneity was observed (Figure 2.1).

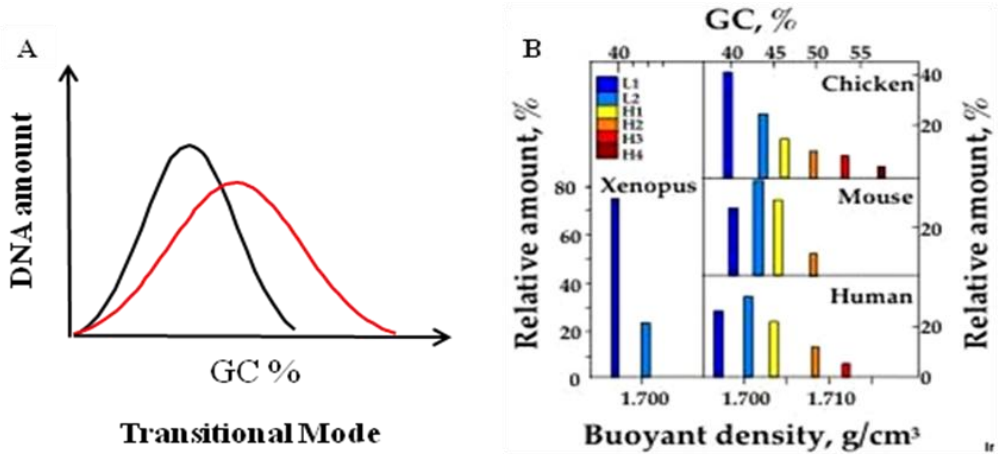


Figure 2.1: **(A)** Schematic representation of transitional mode in vertebrates and **(B)** isochore families obtained from density gradient centrifugation (From Bernardi, 2004).

Indeed, comparing CsCl profiles of amphibian/reptile genomes against those of mammals, it was observe that the increment of temperature was paralleled by an increment of the genome heterogeneity due to the increment of the GC content, with the appearance of the GC-rich isochores in mammals [i.e. the DNA regions characterized by fairly homogeneous base composition (Bernardi, Olofsson et al., 1985)] (Bernardi, 2004, for review).

2.1.2 Shifting Mode The second mode of evolution was mainly observed comparing fish genomes. In this case compositional changes are not paralleled by an increment of the genome heterogeneity, but by a complete shift of the compositional profile, thus affecting the whole genome (Figure 2.2, panel A). Comparing the isochore maps of fish genomes (Figure 2.2, panel B), no overlap between compositionally far genomes was observed, as for example in the case of zebrafish and fugu showing an average GC content of 36.9% and 44.0%, respectively (Costantini, Auletta et al., 2007).

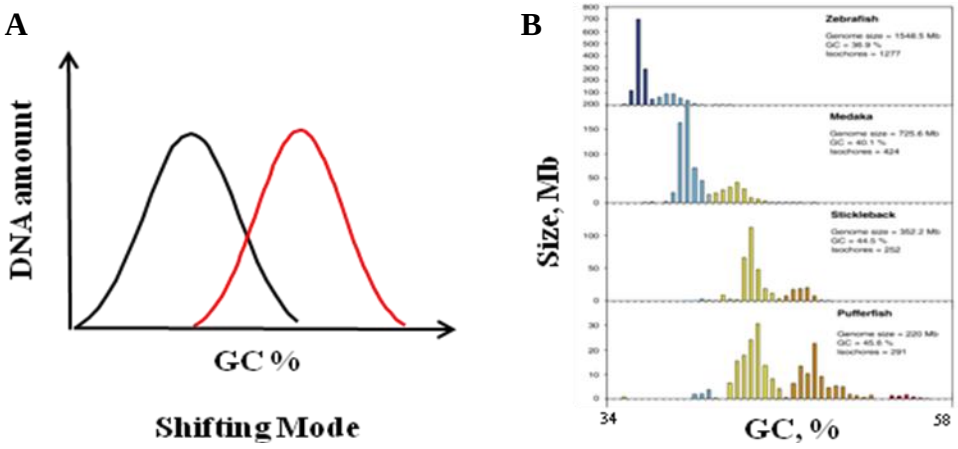


Figure 2.2: **(A)** Schematic representation of shifting mode in vertebrates and **(B)** distribution of isochores according to GC levels of fish genomes (From Costantini and Bernardi, 2007).

2.2 Teleostean Genomes

2.2.1 Introduction

2.2.1.1Teleosts

Teleostei (Actinopterygii), also known bony fish, are the most diverse and the largest vertebrate group, roughly making up half of the extant vertebrate species, accounting for more than 99% of ray-finned fish, with about 27,000 species in about 40 orders (Nelson, 2006). Palaeontological evidence suggests that the radiation of teleosts occurred between 150 and 250 Mya, in the Triassic period. The ray-finned fish vary widely in genome size, morphology behavior, and adaptations. This huge variability makes them extremely attractive for the study of many biological questions, particularly of those related to evolution. The diversity has been attributed to a whole-genome duplication (WGD) event in the ray-finned fish lineage (Ravi and Venkatesh, 2008). The sequencing of genes and gene families from teleost fish had unexpectedly revealed the presence of duplicate teleost genes for several human genes. Recently, the sequencing and comparative analysis of whole-genome sequences of teleost fish such as fugu, tetraodon, and medaka (Aparicio, Chapman et al., 2002; Jaillon, Aury et al., 2004; Kasahara, Naruse et al., 2007) have provided compelling evidence for the WGD event in the fish lineage. Timing the gene duplication events using molecular clock and phylogenetic analyses indicated that the duplication have occurred around 350 Mya, before the diversification of teleosts (Christoffels, Koh et al., 2004; Hoegg, Brinkmann et al., 2004; Vandepoele, De Vos et al., 2004).

Genomes Fish genomes seem to be ‘plastic’ in comparison with other vertebrate genomes because genetic changes, such as polyploidization, gene duplications, gain of spliceosomal introns and speciation, are more frequent in fish (Venkatesh, 2003). Traditionally fish had been the subject of comparative studies but recently there has been an increased interest in these vertebrates as model organisms in genomics and molecular genetics. Indeed, the second vertebrate genome to be sequenced completely after human

genome was of *Takifugu rubripes* (Aparicio, Chapman et al., 2002). The analyses of the fish genome sequences have provided useful information for understanding the structure, function and evolution of vertebrate genes and genomes. Teleost genomes have recently received much attention and a large amount of genomic sequence information has become available (see Appendix 2, Table 7.1).

The rationale of focusing our attention on the Teleostei was grounded on the two facts:

(i) Aquatic organism, different from terrestrial one, lives in an environment where the available oxygen, dictated by the Henry's law, is a limiting factor.

(ii) Teleosts are a highly diverse group of animals occupying all kind of aquatic environments. Hence the apparently high variability of metabolic rate around the classical allometric relationship (Clarke and Johnston, 1999), makes them an ideal choice for analyzing metabolic adaptation independent from mass and temperature.

The compositional studies were performed to test the two main adaptive hypotheses proposed to explain the GC variability among organisms: the thermal stability (Bernardi, Olofsson et al., 1985) and metabolic rate (Vinogradov, 2001, 2005; Vinogradov and Anatskaya, 2006). Our goal was to answer the following question: "Which of the two or both hypotheses drive the 'shifting mode' of evolution observed among fish genomes?"

2.2.1.2 Temperature and GC content

Doubts about the link between T° and GC% were raised by comparing data on genomic GC% with the corresponding habitat temperature of several fish (Table 2.1). Indeed, the GC content of *G. aculeatus*, living at a temperature range of 4 - 20°C, was around 44.5%. On the contrary *O. latipes,* living at a temperature range of 18 - 24°C, showed a GC% value of 40.1% (Table 2.1). Hence the GC% doesn't seem to fit perfectly with the corresponding living temperature range, therefore casting doubts on the thermodynamic stability hypothesis according to which increment of environment (or body) temperature produces a GC increment to stabilize DNA, RNA and proteins.

Table 2.1: GC % and habitat temperature of four teleosts.

	D. rerio	***O. latipes***	***G. aculeatus***	***T. nigroviridis***
GC%	36.9	40.1	**44.5**	45.6
Habitat Temperature (°C)	18-24	18-24	**4-20**	24-28

The above consideration drove our attention towards the role played by the metabolic rate in shaping the genomic GC content in fish genomes. Before tackling this point, let's consider shortly the metabolic rate and the factors affecting this parameter.

2.2.1.3 Metabolism and metabolic rate

Metabolism is the bio-chemical process by which energy and material are transformed within an organism and exchanged between the organism and the environment (Brown, Gillooly et al., 2004). Organisms convert the acquired resources from the environment to biologically usable forms and allocate them to the vital processes of survival, growth, and reproduction, and eliminate the altered forms back into the environment. On one hand metabolism determines the demands that organisms place on their environment while on the other, environmental factors constraints the allocated of resources to sustain life by influencing the organismic performance, and hence their energetic demands.

The complex network of biochemical reactions (organized into metabolic pathways) are catalyzed by enzymes regulating the rates of reactions. The overall rate of the processes (rate at which energy and materials are taken up, transform, and used up) is the metabolic rate is related to molecular processes like DNA repair and mutation, because metabolism produces oxygen radicals (highly reactive molecules with free electrons) which mediates the oxidative damage to DNA. Naturally this damage is continuously repaired and mutation may occur by incorrect repair. Hence species with higher metabolic rates should have higher DNA substitution rates, due to the fact that DNA damage and mutation rate are positively correlated (Martin and Palumbi, 1993; Gillooly, Brown et al., 2001; Gillooly, Allen et al., 2005).

2.2.1.4 Factors affecting metabolic rate

Body mass and temperature are the two major factors affecting metabolism (Gillooly, Brown et al., 2001), according to,

$$B = B_0 M^{-1/4} e^{-E / kT}$$

Where, *B* is the mass specific metabolic rate, B_0 is the coefficient independent of body size *M*, $e^{-E/kT}$ is the Boltzmann factor with *k* as Boltzmann constant, activation energy *E* and temperature *T*.

2.2.1.4.1Body mass

Mass specific metabolic rate varies with body size elevated to a factor that approximates a quarter-power (Kleiber, 1932; West, Brown et al., 1997,1999).At present it is still questioned if this represents a universal scaling law. The debate is mainly dealing with the meaning of the species-related variability in the normalizing coefficient *a* and the scaling exponent *b* of the allometric equation, $R_b = aM^b$, where R is the resting metabolic rate and M is the body mass. Recently a curvilinear rather than linear relationship between LogR and LogM has been proposed (Kolokotrones, Van et al., 2010). A metabolic level boundaries (MLB) hypothesis stressing on the boundary constrains that limit the scaling of metabolic rate has also been proposed (Glazier, 2010) and applied to teleost fish (Killen, Atkinson et al., 2010) suggesting that lifestyle, swimming mode and temperature affect the intraspecific scaling of resting metabolic rate.

2.2.1.4.2 Temperature

Within the range of biologically relevant values (approximately 0 - 40 °C), temperature affects metabolism mainly via its effects on the rates of biochemical reactions, whose kinetics varies according to the Boltzmann's factor ($e^{-E/kT}$), where *E* is the activation energy, *k* is Boltzmann's constant, and *T* is absolute temperature, known as the universal temperature dependence (UTD) (Gillooly, Brown et al., 2001).

Thus, metabolic rate, the rate at which organisms transform energy and materials is governed largely by two interacting processes: first the quarter-power allometric relation, which describes how biological rate processes scale with body size and second the Boltzmann factor, which describes the temperature dependence of biochemical processes (UTD).The assumption of a universal significance of both scaling and UTD forms the basis of the Metabolic Theory of Ecology (MTE), proposed by (Brown, Gillooly et al., 2004), that stresses the ecological relevance of the mass and temperature dependence of metabolic rate. Despite the fact that this hypothesis has been questioned (Glazier, 2005; Enquist, Allen et al., 2007), UTD can be considered in any case a useful statistical tool to describe the relationship between temperature and basal metabolic rate (Clarke, 2006). In particular, the methodological approach of MTE could be used to separate the effects of mass and temperature from those of other sources of basal metabolic rate variability, including those related with life history and specific environmental adaptations of a species or group of species. In this view, mass and temperature correction of metabolism within a group of phylogenetically related organisms may reveal a broad tendency to adapt metabolism to different environments.

2.2.2 Results

Teleost fish were grouped in the five major habitat: polar, temperate, subtropical, tropical and deep-water. Data about genomic GC levels were obtained from available literature (Bucciarelli, Bernardi et al., 2002; Varriale and Bernardi, 2006). Data about taxonomic classification, geographical distribution and metabolism were retrieved from public database (www.fishbase.org). For each species, mass specific metabolic rate values, expressed as milligrams of oxygen consumed per kilogram of wet weight per hour (mg kg^{-1} h^{-1}), were temperature-corrected by applying the Boltzmann's factor (see Appendix 6.4.1).

2.2.2.1 Body mass and metabolic rate according habitat Within each habitat both body mass (BM) and temperature-corrected metabolic rate (MR) showed skewed distributions. Therefore, log normalized values of BM and MR were plotted and visualized by box plots (Figure 2.3). Potential outliers (Figure 2.3; red dots) were detected according to (Barnett and Lewis, 1984) and removed from the dataset (see also Appendix 6.5.1). The range of temperature for each habitat was also reported, except that of deep-water because of its unavailability. The average values of BM and MR showed a decreasing trend, from polar to deep-water, when plotted against the habitat (Figure 2.3, panels A and B, respectively). As BM is expected to inversely correlate with MR, the similar trend of BM and MR among habitat confirmed that the effect of body mass on metabolic rate was removed after applying the Boltzmann's factor. Regarding MR (Figure 2.3, panel B; Table 2.2), polar teleostean fish showed the highest average value. Temperate fish displayed a significantly higher average MR than that of tropical ones, which had the lowest average value.

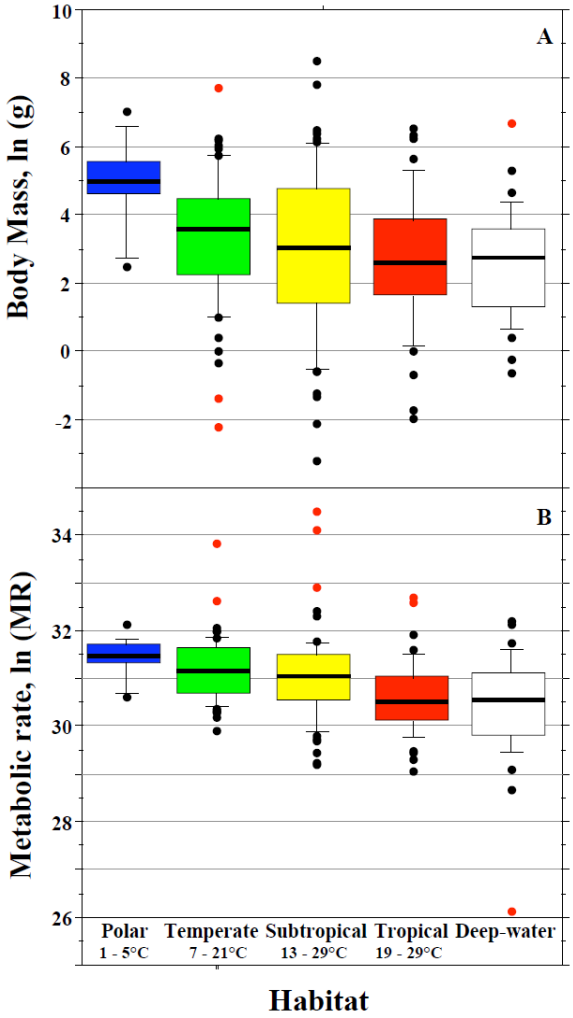


Figure 2.3: Box plot log-normalized distributions of body mass (panel A) and specific metabolic rate (panel B) in teleostean within each habitat group. Boxes are sorted according to the increasing temperature of habitat. Outliers are shown with red dots.

Table 2.2: *p*-values of Mann-Whitney test for MR of teleostean among different habitat.

		Polar	Temperate	Subtropical	Tropical
Teleostean	**Polar**	--			
	Temperate	$< 2.75 \times 10^{-2}$	--		
	Subtropical	$< 7.1 \times 10^{-3}$	ns	--	
	Tropical	$< 1.0 \times 10^{-4}$	$< 1.0 \times 10^{-4}$	$< 1.1 \times 10^{-3}$	--
	Deep-water	$< 5.0 \times 10^{-4}$	$< 2.2 \times 10^{-3}$	$< 5.0 \times 10^{-2}$	ns

2.2.2.2 Temperature and metabolic rate Temperature and metabolic rate are well known to be strongly correlated. In order to test independently the role played by the two variables on the GC content, Boltzmann's factor was used to disentangling temperature (T°C) from the metabolic rate. After correction, a negative correlation between average environmental T°C and MR was observed, $R^2 = 0.121$ and $p < 10^{-4}$ (Figure 2.4), for fish living in all habitat (except for deep-water fish, because of the unavailability of temperature data).

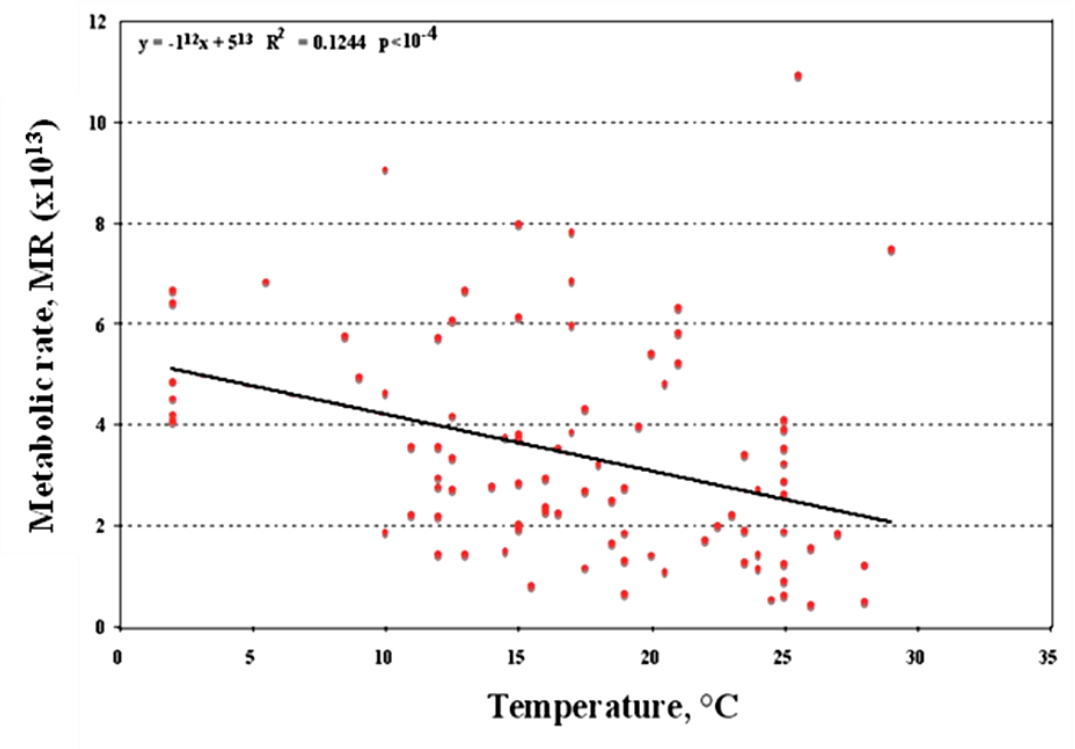


Figure 2.4: Plot of the mass specific metabolic rate, corrected according to the Boltzmann' factor (MR), against the average environmental temperature (T°C). The equation of the linear regression, the correlation coefficient (R^2) and the *p*-value are reported**.**

2.2.2.3 Effects of phylogenetic relationship Clarke and Johnston (1999) reported that MR was not affected by the phylogenetic relationship among species, an observation also confirmed by the present analysis. Restricting the analysis to the order of Perciformes, showing the largest MR range (**S1**), the same decreasing trend among habitat was observed (Figure 2.5). Indeed, the average MR value of polar fish was significantly the highest, and that of both temperate and subtropical fish was significantly higher than that of tropical one (Table 2.3). The limited number of deep-water Perciformes (only four) probably accounts for the non-significant pair-wise comparisons with temperate and subtropical fish (see also Table 2.1). Hence the result excluded the effect of phylogenetic relationship on MR.

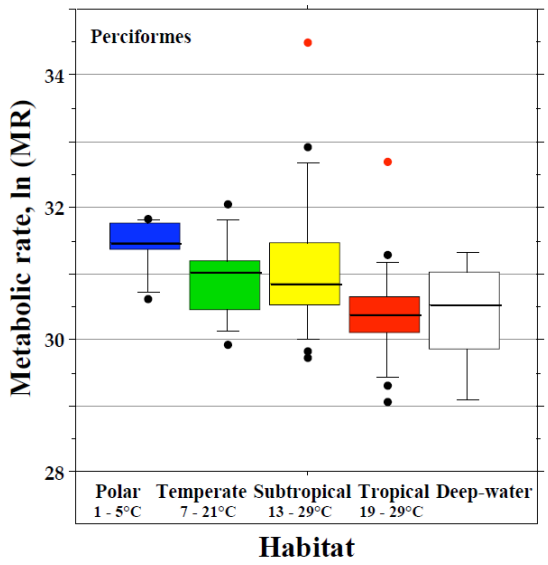


Figure 2.5: Box plot of log-normalized distribution of specific MR in Perciformes within each habitat group. Outliers are shown red dots.

Table 2.3: *p*-values of Mann-Whitney test for MR of Perciformes among different habitat.

		Polar	Temperate	Subtropical	Tropical
Perciformes	**Polar**	--			
	Temperate	$< 5.7 \times 10^{-2}$	--		
	Subtropical	$< 3.9 \times 10^{-2}$	ns	--	
	Tropical	$< 1.0 \times 10^{-4}$	$< 2.5 \times 10^{-2}$	$< 8.2 \times 10^{-3}$	--
	Deep-water	$< 1.0 \times 10^{-2}$	ns	ns	ns

2.2.2.4 GC content among habitat In the light of the current hypotheses of vertebrate genome evolution (Vinogradov, 2001; Bernardi, 2004, for review; Vinogradov, 2005), the distribution of the genomic GC levels among habitat was investigated. Unfortunately, no data were available about the genomic GC levels of fish living in the deep-water habitat. The GC%, like MR, showed a decreasing trend from polar to tropical habitat (Figure 2.6; outliers showed as red dots). However, pair-wise comparisons between polar and temperate as well as between subtropical and tropical showed that the differences were not statistically significant in both cases (Table 2.4). On the contrary, the GC% of both polar and temperate was significantly higher than that of subtropical and tropical habitat (Table 2.4).

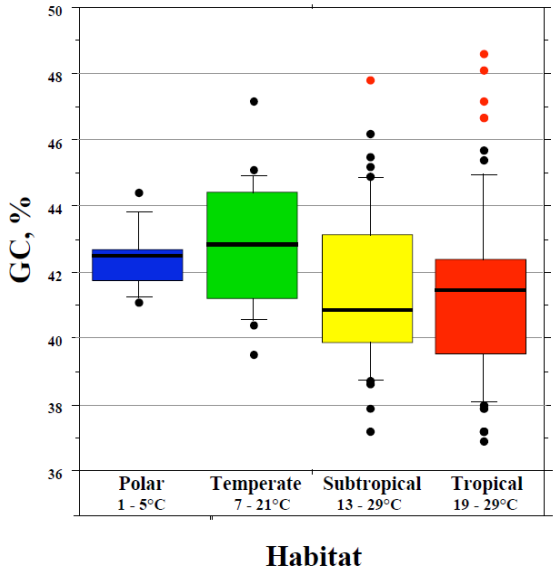


Figure 2.6: Box plot of GC% genomic levels distributions within each habitat group. Outliers are shown in red dots.

Table 2.4: *p*-values of Mann-Whitney test for GC levels among different habitat.

		Polar	Temperate	Subtropical
GC, %	**Polar**	--		
	Temperate	ns	--	
	Subtropical	$< 5.5 \times 10^{-2}$	$<5.0 \times 10^{-3}$	--
	Tropical	$< 1.9 \times 10^{-2}$	$< 1.0 \times 10^{-3}$	ns

2.2.2.5 Metabolic rate and GC In order to uncover the correlation between MR and GC%, the two databases were crossed and common data for 34 fish (nine polar, nine temperate, twelve subtropical and four tropical fish) were obtained (Table 2.5).

Table 2.5: Average metabolic rate and genome base composition.

Habitat	Species	MR ($x10^{12}$)[a]	GC%
Polar	*Boreogadussaida*	52.28	48.40
	Gymnodracoacuticeps	45.62	42.60
	Pagotheniaborchgrevinki	47.56	41.80
	Trematomusbernacchii	48.53	43.59
	Trematomuscentronotus	45.10	42.50
	Trematomushansoni	64.07	41.10
	Nototeniacoriiceps	42.01	44.40
	Nototeniarossii	66.60	44.52
	Gobionotothengibberifrons	19.88	42.62
Temperate	*Anguilla anguilla*	14.87	44.00
	Centropristismelana	83.43	44.90
	Fundulusheteroclitus	38.56	40.40
	Gasterosteusaculeatus	57.15	44.00
	Onchorhynchuskisutch	60.60	44.50
	Onchorhynchusmykiss	68.32	43.50
	Onchorhynchusnerka	26.65	44.40
	Salmofario	49.33	44.80
	Salmosalar	68.27	44.40
Subtropical	*Anguilla rostrata*	37.51	42.60
	Brevoortiatyrannus	41.69	43.40
	Carassiusauratus	20.65	37.90
	Chromischromis	33.19	40.10
	Clinocottusanalis	16.55	41.20
	Cyprinodonvariegatus	48.40	40.60
	Cyprinuscarpio	11.57	37.20
	Embiotocalateralis	22.65	40.10
	Gillichthys mirabilis	16.11	38.40
	Ophiodonelongatus	13.23	41.50
	Opsanus tau	54.27	40.90
	Orizyalatipes	58.02	40.10
Tropical	*Oreochromisaureus*	6.41	41.60
	Oreochromismossambicus	27.90	41.80
	Oreochromisniloticus	18.93	41.70
	Daniorerio	27.50	36.90

[a] Metabolic rate according to Boltzmann's correction (see Appendix 1, section 6.4.1).

The box plot of the average MR values in the four habitat was slightly different from that observed for the all dataset of teleostean fish (Figure 2.7, panel B). The average MR of polar fish, indeed, was not significantly different from that of temperate ones, but in turn both were significantly higher than those of fish living in subtropical and tropical habitat (Table 2.6). Similar results were obtained analyzing the average genomic GC levels (Figure 2.7, panel A; Table 2.6).

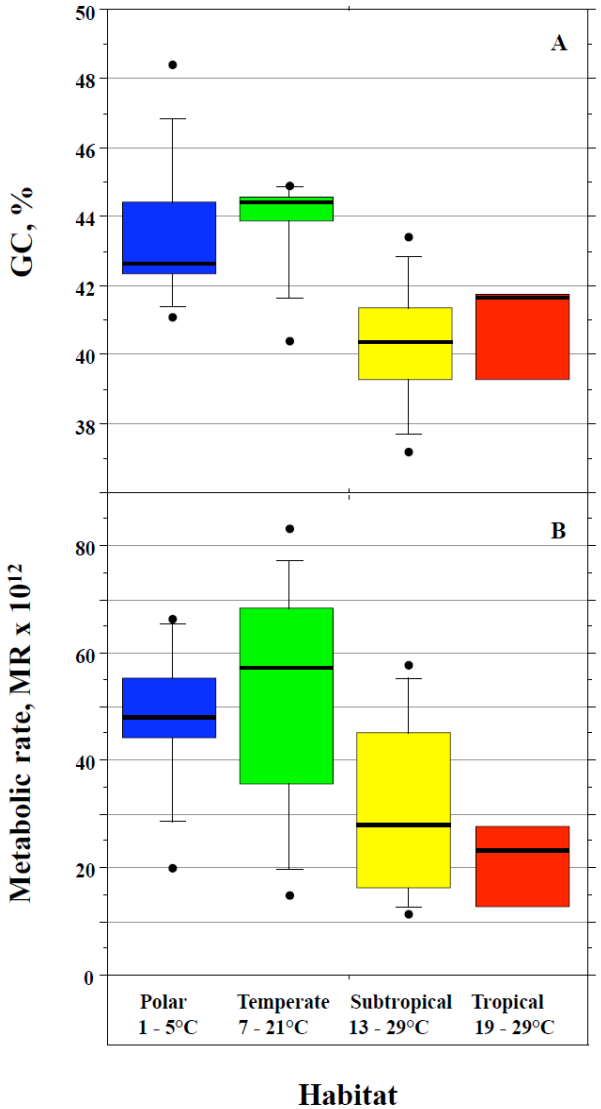


Figure 2.7: Box plot of genomic GC levels (panel A) and specific MR (panel B, corrected for the Boltzmann's factor) distributions within each habitat group.

Table 2.6: *p*-values of Mann-Whitney test for GC% and MR of fish among different habitat.

		Polar	Temperate	Subtropical
GC, %	**Polar**	--		
	Temperate	ns	--	
	Subtropical	$< 2.0 \times 10^{-3}$	$< 6.0 \times 10^{-4}$	--
	Tropical	$< 2.5 \times 10^{-2}$	$< 2.1 \times 10^{-2}$	ns
MR	**Polar**	--		
	Temperate	ns	--	
	Subtropical	$< 3.3 \times 10^{-2}$	$< 2.8 \times 10^{-2}$	--
	Tropical	$< 1.3 \times 10^{-2}$	$< 4.5 \times 10^{-2}$	ns

In view of these results, plotting the MR of the each 34 species against their corresponding genomic GC, positive and significant correlation was found, $R^2 = 0.252$ and $p < 2.5 \times 10^{-3}$ (Figure 2.8).

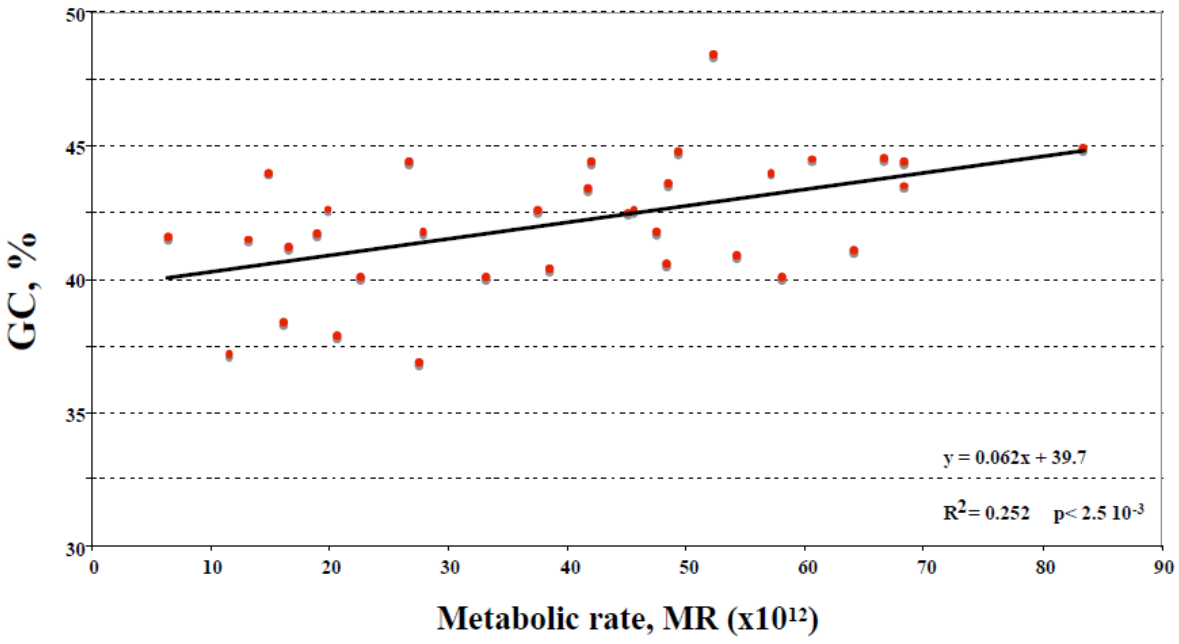


Figure 2.8: Plot of the specific MR, corrected for the Boltzmann's factor, against the average genome base composition, reported as GC%. The equation of the linear regression and the correlation coefficient (R^2) and the *p*-value are reported.

2.3 Role of the CpG doublet

2.3.1 Introduction

The high mutation rate of CpG doublets [due to the methylation of cytosine residues in the CpG dinucleotide with subsequent deamination of 5-methylcytosine (5mC) to thymine (Bird, 1986)] makes CpG the most under-represented doublet as compared to all possible DNA doublets (Josse, Kaiser et al., 1961). Therefore, deaminated 5mC have been reported in the literature as highly responsible for the GC ⇒ AT compositional change, thus affecting the genomic GC content (Fryxell and Zuckerkandl, 2000), since CpG frequency was reported to be linearly correlated with the GC content of the genome (Bernardi, 1985; Aissani and Bernardi, 1991). Figure 2.9 summarizes the "methylation/ deamination" processes, affecting the CpG doublets present in both DNA leading and lagging strands. The deamination process transforms 5mC to thymine, leading to the final derivatives doublets, not only TpG but also CpA (Duret and Galtier, 2000).

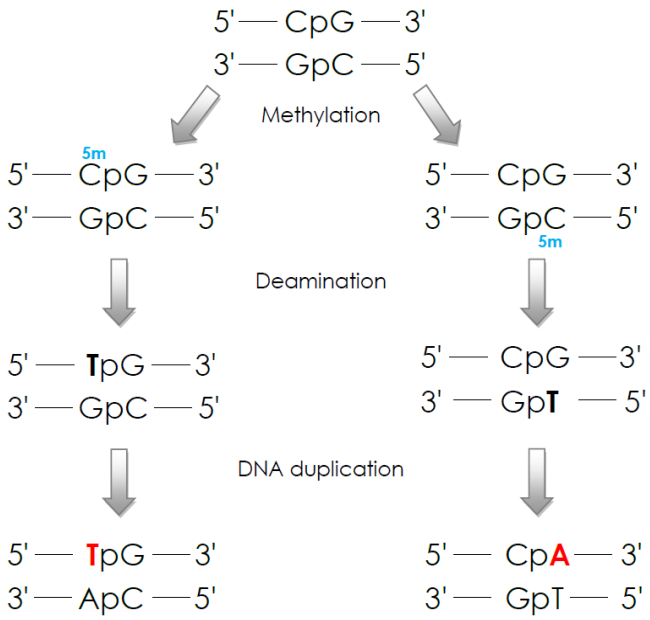


Figure 2.9: Scheme of methylation/deamination process of CpG resulting into TpG and CpA doublets.

In fish genomes the 5mC levels were reported to be different in fish living in different habitat. Precisely, 65 fish representing 12 teleostean orders were pooled into the three categories: polar, temperate and tropical. The average 5mC showed a decrease from polar to tropical fish (Figure 2.10) (Varriale and Bernardi, 2006).

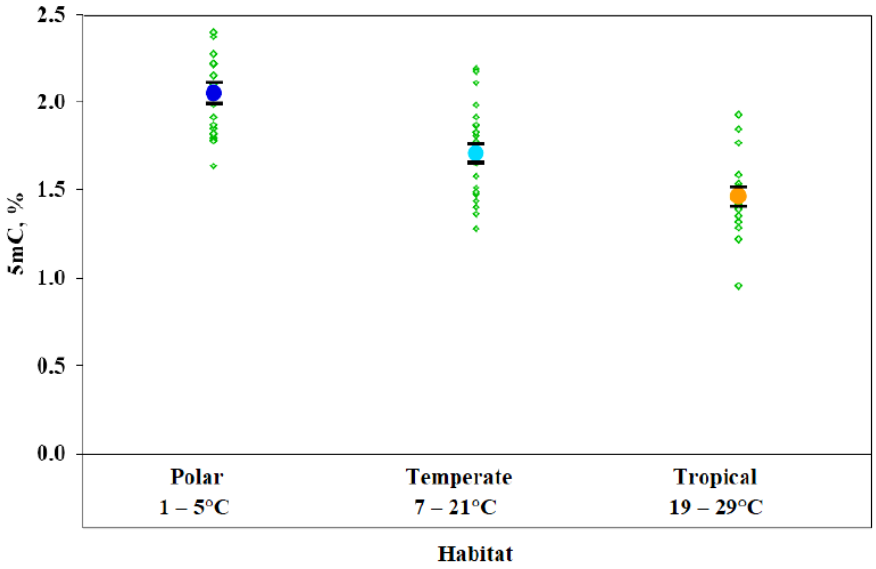


Figure 2.10: The cytosine methylation levels (5mC) in the genomes of fish living in different habitat. Standard error bars are shown. Modified from (Varriale and Bernardi, 2006).

Considering the temperature dependence of the 5mC deamination process (Shen, Rideout et al., 1994), it was hypothesized that low genomic GC% in fish living in warm habitat (Figure 2.6) could be due to the loss of CpG doublets (Varriale and Bernardi, 2006).

2.3.2 Results

The frequencies of CpG and those of the final derivative doublets of the methylation/deamination process, i.e. TpG and CpA, were checked in the available intronic sequences from five completely sequenced teleostean genomes, namely *D. rerio* (zebrafish, 36.9% GC), *O. latipes* (medaka, 40.1% GC), *G. aculeatus* (stickleback, 44.5% GC), *T. rubripes* (fugu, 44.0% GC) and *T. nigroviridis* (pufferfish, 45.69% GC). Intron analysis allowed not only to check the frequencies of the doublets of interest in a large amount of non-coding region, that could be considered well representative of the whole genome, but also, through gene orthology to compare corresponding DNA sequences in different genomes.

2.3.2.1 Dinucleotide % among habitat The box-plots of CpG and TpG + CpA for each species were arranged according to the increasing living temperature of organisms (Figure 2.11, top and bottom panel, respectively). Regarding the frequencies of the CpG doublet, the result were in rather good agreement with those reported by (Varriale and Bernardi, 2006), see Figure 2.10. Indeed, zebrafish and medaka, living in a tropical and subtropical habitat, respectively, showed lower values than those of stickleback and fugu, living in a temperate habitat. However, fugu showed very high CpG frequency, although living in a tropical habitat. In all five genomes, the frequency of the CpG doublet was highly correlated, as expected, with the genomic GC% (Table 2.7).

Regarding the frequencies of the TpG + CpA doublets (Figure 2.11, bottom panel), small and not significant variation was found among genomes living at different temperatures.

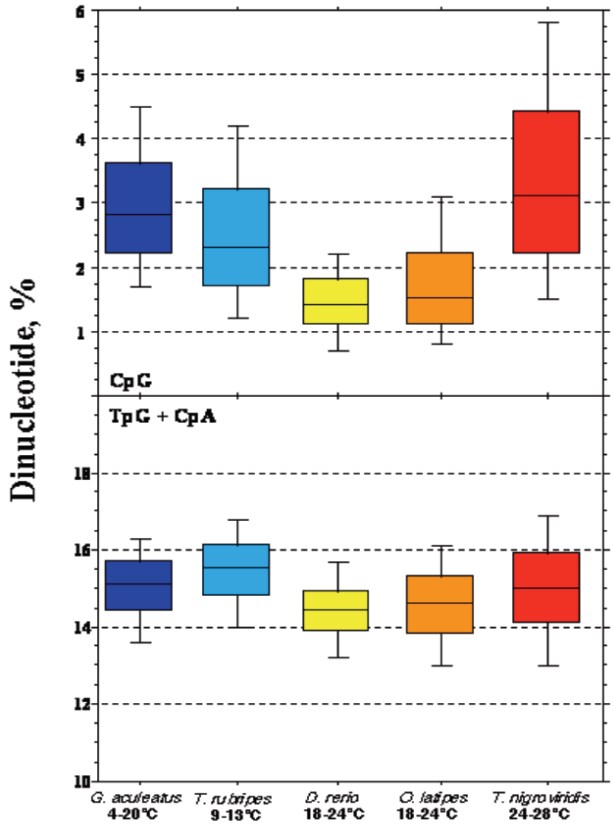


Figure 2.11: Box plot of the CpG (top panel) and TpG + CpA (bottom panel) frequencies in the intronic sequences of fish living at different temperatures.

The Spearman correlation test of CpG *vs.* TpG and CpG *vs.* CpA among genomes showed no significant correlation, p-values were <0.16 and <0.96, respectively.

2.3.2.2 CpG *vs.* TpG + CpA The above result questioned about the correlation between dinucleotides within each genome. Plotting the CpG frequency against that of TpG + CpA, low correlation coefficients were found (Figure 2.12).

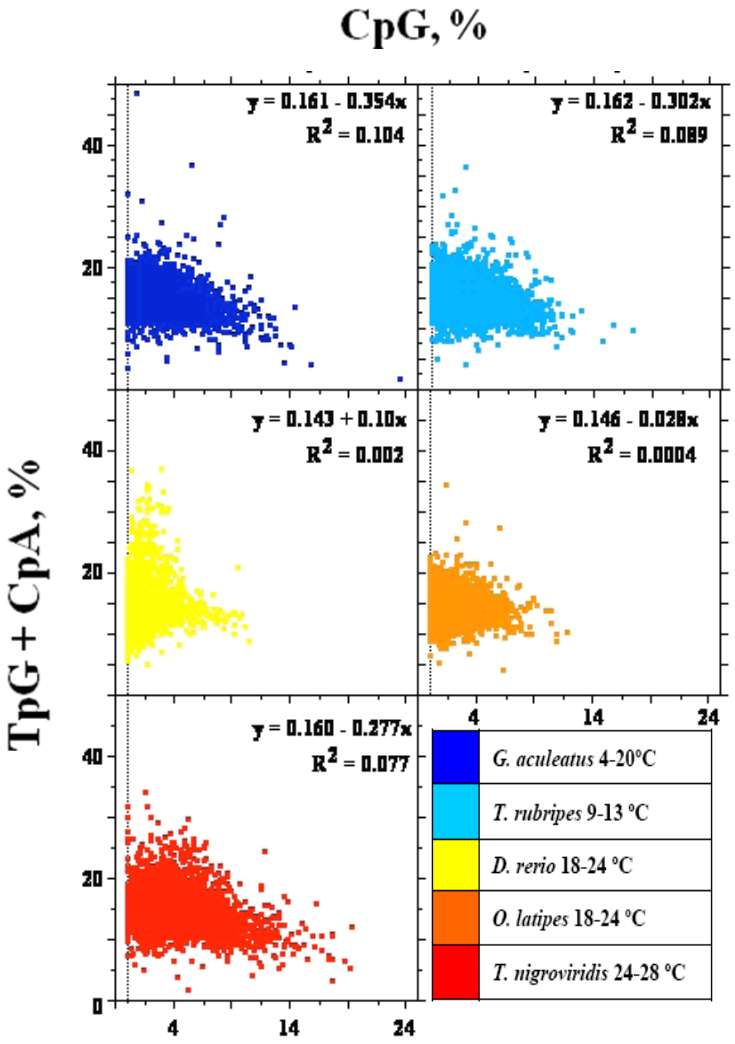


Figure 2.12: The plots show the correlations between CpG and TpG + CpA in the intronic sequences of fish living at different temperatures. The equations of the regression lines, as well as the correlation coefficients (R) are reported.

The result was not due to the cumulative effect of TpG + CpA. Indeed, Plotting CpG against TpG and CpA separately, low R values were also found (Table 2.7).

In order to avoid the affect caused by transposable and repetitive elements, the analysis was performed after removing interspersed repeats and low complexity DNA sequences using RepeatMasker. The overall picture remained unaffected, although the majority of correlations showed increased R values (Table 2.7, bottom).

Table 2.7: Intra-genomic correlation coefficient (R) before and after Repeat masker.

Before Repeat Masker	Species	GC *vs.* CpG	CpG *vs.* TpG	CpG *vs.* CpA
	G. aculeatus	0.773	0.188	0.223
	T. rubripes	0.803	0.166	0.213
	D. rerio	0.686	0.006	0.076
	O. latipes	0.741	0.010	0.037
	T. nigroviridis	0.837	0.118	0.241
After Repeat Masker	***G. aculeatus***	0.781	0.238	0.282
	T. rubripes	0.820	0.227	0.261
	D. rerio	0.693	0.077	0.044
	O. latipes	0.746	0.007	0.056
	T. nigroviridis	0.843	0.197	0.324

In order to avoid any bias caused by different sets of genes [the available number of intronic sequences greatly vary among fish genomes (ranging from the $10^4$of medaka to the $3.8x10^4$of fugu), the analysis was restricted to sets of orthologous pairs. It is worth to stress that orthology was based on the analysis of coding sequences and extended to the corresponding intronic sequences. The results showed the same trend already shown above, i.e. not clear cut trend with different habitat temperature. For example, in the set of orthologous intron pairs the intra-genome correlation coefficients (R) of the regression CpG *vs.* TpG + CpA: i) comparing *D. rerio vs. G. aculeatus* the values were in each genome 0.133 and 0.314, respectively; ii) comparing *O. latipes vs. G. aculeatus* they were 0.004 and 0.289, respectively; iii) comparing *D. rerio vs. T. rubripes* they were 0.133 and 0.316, respectively; and iv) comparing *O. latipes vs. T. rubripes* they were 0.013 and 0.312, respectively. Once more the intragenomic correlations between doublets turned out to be barely affected after removing repetitive elements.

2.4 Intron length

2.4.1 Introduction

Evolution of the intron architecture (like length, GC content) and its significance has been studied across the broad phylogenetic breadth. Using the gene expression data in *C. elegans* and *H. sapiens*, Castillo-Davis (2002), demonstrated that in highly expressed genes the introns length was substantially shorter than those present in low expressed genes. Also in *M. musculus*, moss *P. patens* and the pollen of *A. thaliana*, highly expressed genes have been found to have short introns (Duret and Mouchiroud, 1999; Seoighe, Gehring et al., 2005; Stenoien, 2007).

Several hypotheses have been proposed to explain the relationship between intron length and gene expression pattern. At present two main hypotheses are considered the most consistent: 'Genome design' and 'Transcription efficiency'.

Genome design The hypothesis suggests that the functional load is mainly responsible of the intron length. More precisely highly expressed genes are short because most of them are house-keeping genes whose epigenetic regulation is less complex than that of weakly expressed tissue-specific genes (Vinogradov, 2004). It has also been reported that not only the gene expression level, but also the expression breadth was strongly correlated (Eisenberg and Levanon, 2003; Vinogradov, 2004). But recently house-keeping genes are reported to be no more compact than the narrowly expressed genes, implying that the breadth plays role in compactness of genes more than expression level (Li, Feng et al., 2007).

Transcription efficiency The hypothesis is based on the fact that transcription is a slow and expensive process, hence suggesting that natural selection for transcriptional efficiency favors the compactness of highly expressed genes (Hurst, McVean et al., 1996; Castillo-Davis, Mekhedov et al., 2002; Eisenberg and Levanon, 2003).The transcription efficiency hypothesis further includes two sub-hypotheses: an energetic cost hypothesis and a time cost hypothesis. The energetic cost hypothesis argues that selection for short introns may be driven by minimizing the energetic cost of transcription (Castillo-Davis, Mekhedov et al., 2002; Zhu, He et al., 2008; Carmel and Koonin, 2009), while the time

cost hypothesis states that selection of short intron is likely due to the requirement to transcribe large amounts of mRNA molecules within limited periods (Chen, Sun et al., 2005; Huang and Niu, 2008). Although the time cost hypothesis got support from analysis on human antisense genes showed that the genes having very short response times have been found to have short introns, but this hypothesis has been argued (or overlooked) in many literatures. Even Seoighe (2005) pointed out that the transcription of multiple copies of mRNA does not necessarily require a much longer period of time than required to transcribe the first copy, because multiple polymerases may be simultaneously works on one template. On the other hand recent analysis on *G. gallus* genome presented evidences in support for the energetic cost hypothesis (Rao, Wang et al., 2010).

In the light of the energetic cost on transcriptional activity hypothesis, considering the negative correlation found in the intronic sequences between length and GC content, as well as and the higher average expression level of the GC-rich genes compared to the GC-poor ones (Arhondakis, Auletta et al., 2004), investigation on the pattern of intron length and GC content in teleost's genome would provide further evidences to understand the intronic architecture of genomes.

2.4.2 Results

The pattern of length (bpi) and GC content (GCi) in the intronic sequences of five completely sequenced teleostean genomes, namely: *D. rerio*, *O. latipes*, *T. rubripes*, *G. aculeatus* and *T. nigroviridis* were investigated. In order to avoid any bias due to a different occurrence of repetitive elements among genomes, hence affecting the intron length, RepeatMasker was used to remove both interspersed repeats and low complexity sequences from intronic sequences. For each genome the average base composition values of both whole genome (GCg%) and introns (GCi%), as well as the average intron length (bpi) are reported in Table 2.8.

Among the five teleosts, GCg and GCi showed the same trend, i.e. *D. rerio*<*O. latipes*<*T. rubripes*<*G. aculeatus*<*T. nigroviridis*, whereas no trend was observed regarding bpi.

Table 2.8: The average base composition (GCi%) and length (bpi) of intronic sequences in fish genomes.

	D. rerio	*O. latipes*	*T. rubripes*	*G. aculeatus*	*T. nigroviridis*
GCg%	36.9	40.1	44.0	44.5	45.6
GCi%	0.364	0.395	0.441	0.434	0.472
bpi	17992.57	3109.9	5366.9	5056.68	3011.24

The distribution of GCi% in five genomes is reported in Figure 2.13**,** therefore confirming the shifting mode of evolution.

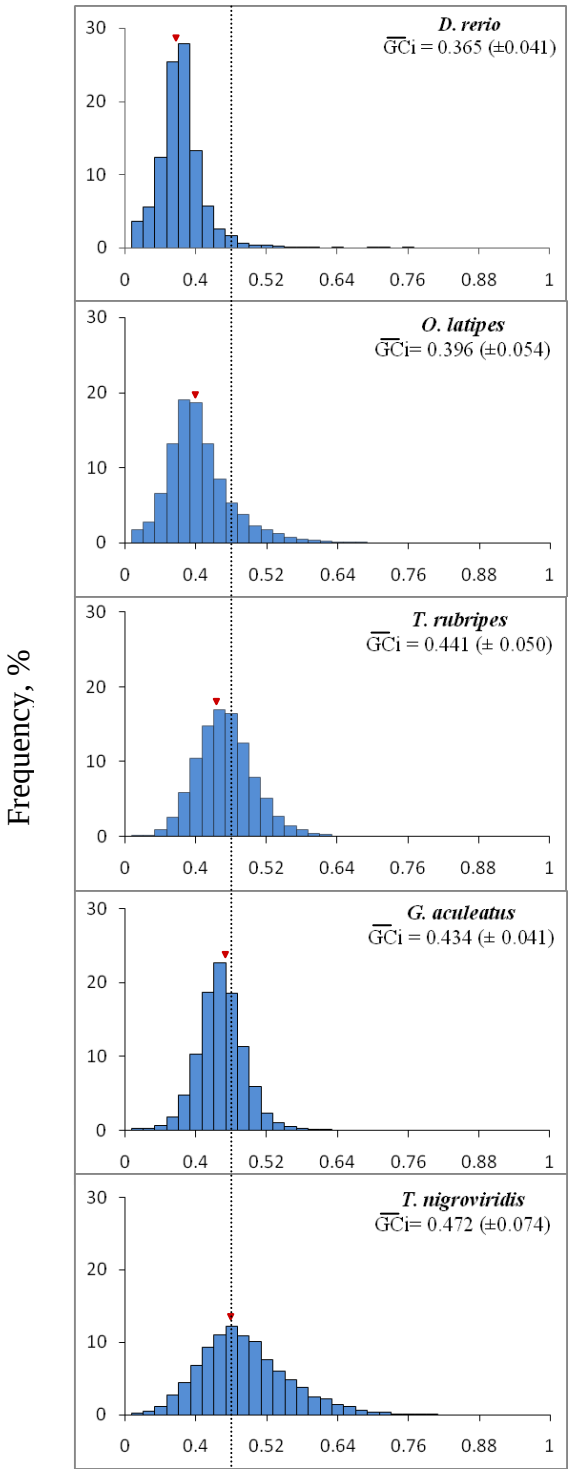


Figure 2.13: Distribution of GCi% in five teleostean genomes; average value of GCi and standard deviations are reported; average genomic GCg is marked with red arrow head.

It is worth to recall that the available number of intronic sequences greatly varies among fish genomes, hence, also in this case, in order to avoid any bias because of different sets of genes, the analyses of GCi and bpi were restricted to sets of orthologous genes (details in Appendix 1, section 6.2). The average GCi% and bpi of each genome, computed taking into account all intronic sequences (see Table 2.8), remained practically unchanged in the subset of orthologous introns obtained in each pairwise analysis (**Table S2, S3**). Hence, the subset of intronic sequences retrieved for each genome in pairwise comparisons could be considered representative of the whole genome.

The delta of GCi% (Figure 2.14, panels A to J**)** and the delta of bpi% (Figure 2.15, panels A to J) were computed in all genome pairwise comparisons. For example in Figure 2.14 panel A reports the ΔGCi% between zebrafish (z) and medaka (m), the value of each histogram bar was calculated as follows:

$$[(\mathrm{GCi_m} - \mathrm{GCi_z}) > 0 \div \sum_{i=1}^{n} \mathrm{xi}\ ] \text{ and } [(\mathrm{GCi_m} - \mathrm{GCi_z}) < 0 \div \sum_{i=1}^{n} \mathrm{xi}],$$

where, *n*= number of orthologous genes between species *m* and *z*. In Figure 2.15, panel A, the Δbpi% between *zebrafish* and *medaka* was reported. Values were computed as above, but disregarding Δbpi>|100|. The histogram bars reported in the two corresponding panels (panels A of Figure 2.14 and 2.15**)** showed an opposite trend.

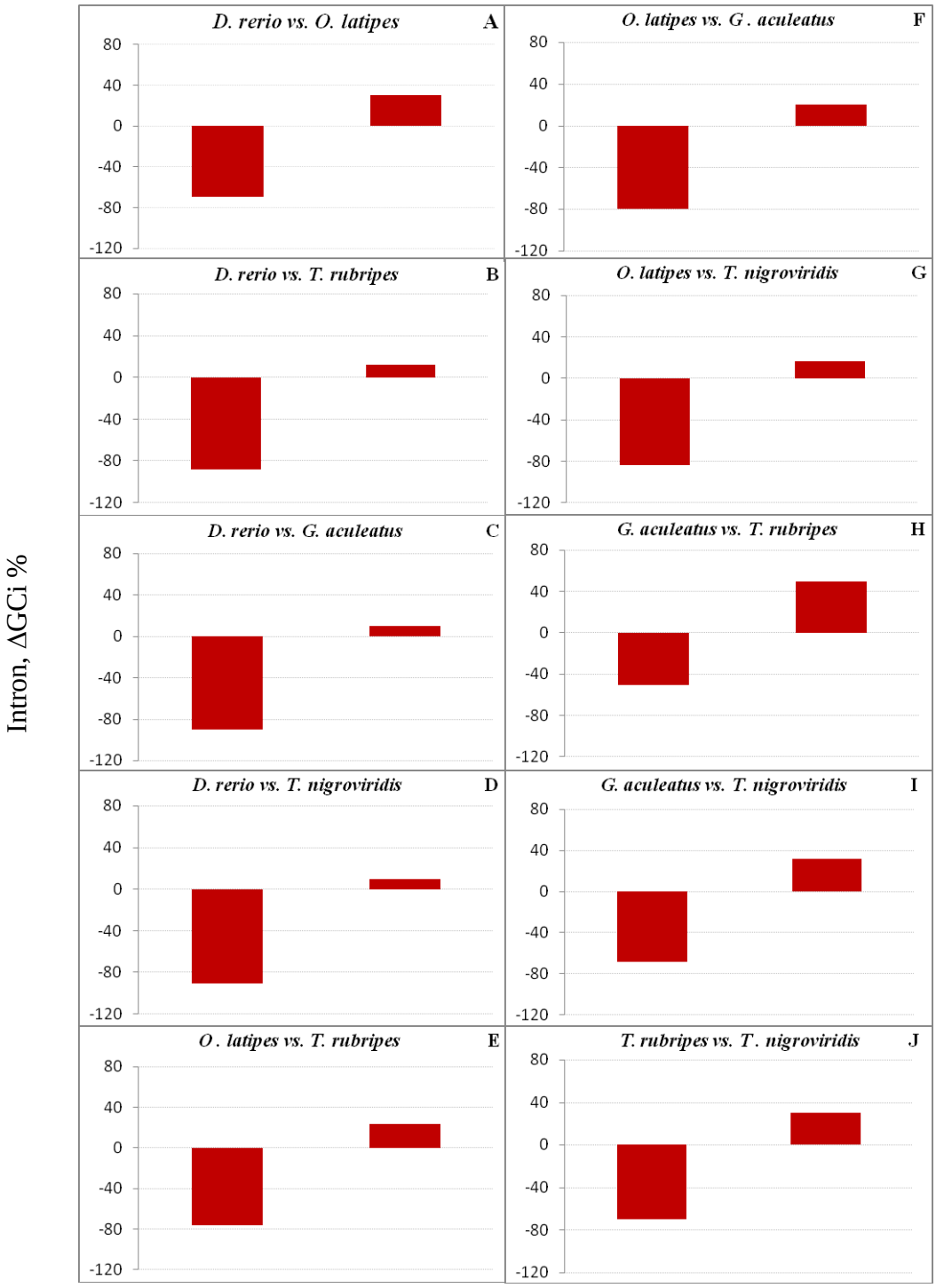


Figure 2.14: Histograms showing the percent of positive and negative values of ΔGCi computed in each subset of orthologous intron pairs (panels A to J).

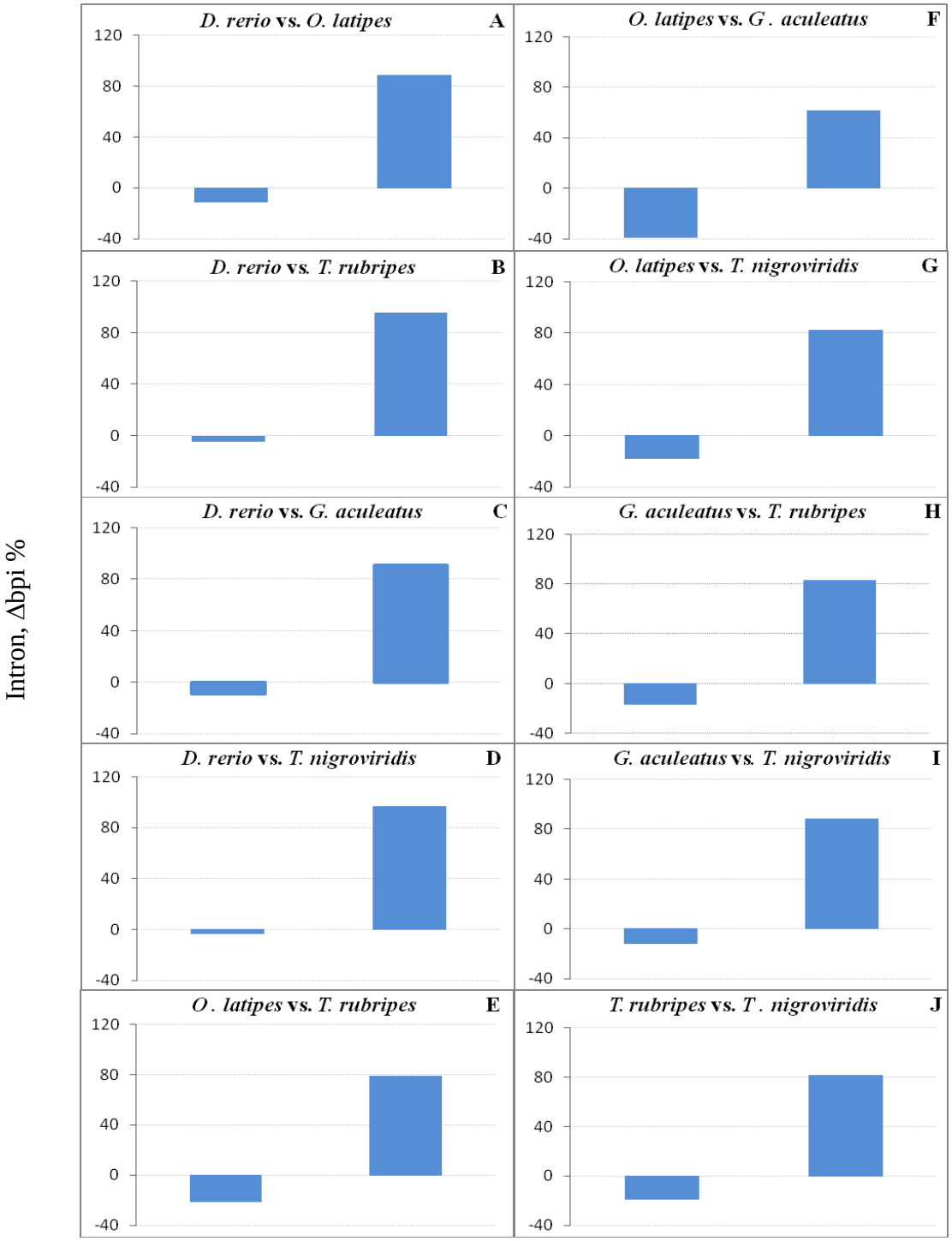


Figure 2.15: Histograms showing the percent of positive and negative values of Δbpi computed in each subset of orthologous intron pairs (panels A to J).

2.5 Discussion

In this chapter three different approaches were applied to undertake the problem of GC content variation and factors affecting it, among and within teleostean genomes (Chapter 2, section 2.1.2).

In the first section (Section 2.2) the effect of temperature and MR on the base composition among teleosts were discussed. Boltzmann correction was applied to metabolic rate values to disentangle the two parameters, in order to show their effect on GC content (GC%) independently. Indeed, considering the data on temperature (T°C) and temperature corrected metabolic rate, a negative correlation has been found $R^2 = 0.121$ and $p < 10^{-4}$ (Figure 2.4).

Data regarding GC% content and temperature corrected metabolic rate in different species of fish were grouped according to their habitat. Instead of performing simple correlation, this approach of dividing according to habitat has provided two benefits: firstly, to detect some outliers within each habitat, and secondly, to take into consideration the chemical-physical parameters of different habitat. Regarding first point, generally speaking, habitat are not delimited by strict boundaries, hence, although removing outliers within each habitat, still the assignment of a given species to the habitat could not be free of errors. Regarding the metabolic rate, species can easily and quickly adapt to different aquatic parameters like available oxygen, salinity, etc. While on contrary, the adaptation of the GC content to the environment is a very slow process. Indeed, in case of metabolic rate significant differences were found among different habitat (Figure 2.3, Panel B), whereas in case of GC content differences were significant in two main blocks one formed by Polar/Temperate and second by Subtropical/ Tropical (unfortunately, GC content among deep-water weren't available) (Figure 2.6). But overall an inverse relationship between GC content and temperature, further confirmed that the temperature dependence of the GC content, was not the expected one according to the thermal stability hypothesis (Bernardi, Olofson et al., 1985).

On one hand, there was an increasing trend of metabolic rate from tropical to the polar habitat, which appears to support the idea of a metabolic adaptation to low temperatures, i.e. to the hypothesis of Metabolic Cold Adaptation (Gaston and Chawn,

1999), a phenomenon still under debate (Clarke and Johnston, 1999; Gillooly, Allen et al., 2006), while on the other the average metabolic rate of deep-water fish was not significantly different from that of tropical fish (Figure 2.3, Panel B; Table 2.2), in spite of the fact that these two habitat are characterized by different average temperatures. This observation suggests that sources of variability, other than mass and temperature, bring to an increase in basal or routine metabolic rate. Probably, one such factor is level of oxygen content which is different in the various environments, and depends from both T° and Pressure (P) [deepness], according to the Henry's law i.e. $p = K_H c$, where p is the partial pressure and c is the concentration of the solute, K_H the Henry's law constant for oxygen (O_2) dissolved in water at 298 K 769.2 L·atm/mol.

Considering the close pattern observed for both metabolic rate and GC content among different habitat, plotting the two variables, a positive and significant correlation was found, $R^2 = 0.252$ and $p < 2.5 \times 10^{-3}$ (Figure 2.8).

Second section (Section 2.3) dealt with a detailed analysis of a possible factor affecting the above discussed results, i.e. methylation/deamination process of CpG doublet, well known to responsible of the genomic GC content variability. Indeed, a decrement of 5mC levels was observed from polar to tropical fish (Varriale and Bernardi, 2006). This result plus the report of the temperature dependence of the deamination process of 5mC (Josse, Kaiser et al., 1961), could explain the negative trend of the GC% among different habitat (Figure 2.6). Therefore the frequencies of CpG doublet and those of derivative doublets, i.e. TpG + CpA were checked in the intronic sequences of the available teleostean genomes completely sequenced. Plotting the frequencies of CpG and those of TpG + CpA (Figure 2.11), the former showed some fluctuation (according to the corresponding different genomic GC level), whereas the latter was rather constant among the genomes, in spite of the different environmental temperature.

The third section (Section 2.4) presents an overview ongoing arguments between two hypotheses, namely: ‘Genome design’ (Vinogradov, 2004) and ‘Transcription efficiency’ (Castillo-Davis, Mekhedov et al., 2002; Eisenberg and Levanon, 2003), proposed behind compactness of highly expressed genes.

The observation that "long genes are scarce in GC-rich isochores" was first reported by Duret and colleagues (1995), which was further shown by Versteeg and colleagues (2003), revealing an inverse correlation of intron length with GC content and gene expression in human genome. Arhondakis and collegues (2004) also reported a higher average expression level of the GC-rich genes as compared to the GC-poor ones. Hence in the light above results, linking the intron length with the GC content through the gene expression, the correlation between intron length (bpi) and GC content (GCi%) was investigated in the above cited five teleostean genomes, in order to shed light on the current energetic cost hypothesis (ascribed by Transcription efficiency).

First of all analyzing the whole data set of intronic sequences in each five species had revealed the distribution of GCi% which was in correspondence with overall genomic GC distribution in each case. The overall distribution of GCi% revealed (Figure 2.13)wide intergenomic spread of GCi content as well as narrow intragenomic distribution, portraying the shifting mode of evolution previously reported analyzing the isochore pattern found in the same species (Costantini, Auletta et al., 2007). The analysis of the link between GCi and bpi, restricted to the orthologous intron pairs found in each genome pairwise comparisons, showed a clear cut inverse relationship between GC% and bpi% in all pairwise genome comparisons (Figure 2.14 and Figure 2.15).

However, the relationship between intron length and GC content is a much debated subject since data about human and fly genomes (Haddrill, Charlesworth et al., 2005; Gazave, Marques-Bonet et al., 2007). Indeed, an extensive analysis within thirteen different eukaryotic species showed different pattern of two variables in different genomes. Precisely, a strong negative correlation was found in all mammalian genomes analyzed (Human, chimpanzee, Dog, Cow, Mouse and Rat), and on contrary a strong positive correlation was found in fly, rice, zebrafish and worm genomes (Zhu, Zhang et al., 2009). Unfortunately, the positive correlation between GC content and intron length in the zebrafish genome failed to be supported by our data based on the analysis of more than $13x10^3$ intronic sequences. The result was not unexpected. Indeed, intragenome correlations taking into account base compositional variability only hold within genomes showing a broad compositional heterogeneity, such as the case of high vertebrates, but

not in fairly homogeneous genomes (D'Onofrio and Bernardi, 1992). To surround the problem, comparative genome analyses should be performed in order to have clear cut pictures.

The preliminary analysis of intron length and GCi content in the five teleosts genome characterized by different genomic GC content, were found to be in good hold with the results reported for all vertebrate genomes so far analyzed (Duret, Mouchiroud et al., 1995; Versteeg, van Schaik et al., 2003; Arhondakis, Auletta et al., 2004; Zhu, Zhang et al., 2009), giving further support to the current hypothesis relating the intron length with the energetic cost of the transcriptional activity. The measure of the metabolic rate of the five fish will be a necessary complement of the above analysis in the frame of the metabolic rate hypothesis.

2.6 Conclusions

The three approaches described above were mainly intended to understand the main guiding factor(s). All the results presented here collectively strengthen the metabolic rate hypothesis as an important factor shaping the shifting mode of teleostean genome evolution. Data on metabolic rate and genomic GC of fish living in different habitat is supported by the analysis of the CpG doublets and highlights the role of temperature on genome evolution. Indeed, T° plays a crucial role, but not in the frame depicted by the thermal stability hypothesis proposed by Bernardi (2004), but as environmental parameter affecting the solvability of O_2 in the aquatic habitat. Furthermore, the inverse relationship between GCi and bpi was an interesting preliminary result supporting the hypothesis linking the intron length with the energetic cost.

2.7 References

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." Science **297**(5585): 1301-10.

Arhondakis, S., F. Auletta, G. Torelli and G. D'Onofrio (2004). "Base composition and expression level of human genes." Gene **325**: 165-169.

Barnett, V. and T. Lewis (1984). Outliers in statistical data: , John Wiley & Sons, Chichester.

Bernardi, G. (1990). "Compositional patterns in the nuclear genome of cold-blooded vertebrates." J Mol Evol **31**(4): 265-81.

Bernardi, G. (2004). Sructural and Evolutionary Genomics. Natural Selection in Genome Evolution. Amsterdam, Elsevier.

Bernardi, G., B. Olofson, J. Filipski, M. Zerial, et al. (1985). "The mosaic genome of warm-blooded vertebrates." Science **228**: 953-958.

Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature **321**: 209-203.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, et al. (2004). "Toward a metabolic theory of ecology." Evcology **85**(7): 1771-1789.

Bucciarelli, G., G. Bernardi and G. Bernardi (2002). "An ultracentrifugation analysis of two hundred fish genomes." Gene **295**(2): 153-62.

Carmel, L. and E. V. Koonin (2009). "A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes." Genome Biol Evol **1**: 382-90.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin, et al. (2002). "Selection for short introns in highly expressed genes." Nat Genet **31**(4): 415-8.

Chen, J., M. Sun, J. D. Rowley and L. D. Hurst (2005). "The small introns of antisense genes are better explained by selection for rapid transcription than by "genomic design"." Genetics **171**(4): 2151-5.

Christoffels, A., E. G. Koh, J. M. Chia, S. Brenner, et al. (2004). "Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes." Mol Biol Evol **21**(6): 1146-51.

Clarke, A. (2006). "Temperature and the metabolic theory of ecology." Functional Ecology **20**(2): 405-412.

Clarke, A. and N. M. Johnston (1999). "Scaling of metabolic rate with body mass and temperature in teleost fish." Journal of Animal Ecology **68**: 893-905.

Costantini, M., F. Auletta and G. Bernardi (2007). "Isochore patterns and gene distributions in fish genomes." Genomics **90**(3): 364-71.

Cruveiller, S., G. D'Onofrio and G. Bernardi (2000). "The compositional transition between the genomes of cold- and warm-blooded vertebrates: codon frequencies in orthologous genes." Gene **261**(1): 71-83.

D'Onofrio, G., K. Jabbari, H. Musto, F. Alvarez-Valin, et al. (1999). "Evolutionary genomics of vertebrates and its implications." Ann N Y Acad Sci **870**: 81-94.

Duret, L. and N. Galtier (2000). "The Covariation Between TpA Deficiency, CpG Deficiency, and G+C Content of Human Isochores Is Due to a Mathematical Artifact." Mol Biol Evol **17**(11): 1620-1625.

Duret, L. and D. Mouchiroud (1999). "Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis." Proc Natl Acad Sci U S A **96**(8): 4482-4487.

Duret, L., D. Mouchiroud and C. Gautier (1995). "Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores." J Mol Evol **40**(3): 308 - 317.

Eisenberg, E. and E. Y. Levanon (2003). "Human housekeeping genes are compact." Trends Genet **19**(7): 362-5.

Enquist, B. J., A. P. Allen, J. H. Brown, J. F. Gillooly, et al. (2007). "Biological scaling: does the exception prove the rule?" Nature **445**(7127): E9-10; discussion E10-1.

Fryxell, K. J. and E. Zuckerkandl (2000). "Cytosine deamination plays a primary role in the evolution of mammalian isochores." Mol Biol Evol **17**(9): 1371-83.

Gaston, K. J. and S. L. Chawn (1999). "Elevation and Climatic Tolerance: A Test Using Dung Beetles." OIKOS **86**(3): 584-590

Gazave, E., T. Marques-Bonet, O. Fernando, B. Charlesworth, et al. (2007). "Patterns and rates of intron divergence between humans and chimpanzees." Genome Biol **8**(2): R21.

Gillooly, J., A. Allen, V. Savage, E. Charnov, et al. (2006). "Response to Clarke and Fraser: effects of temperature on metbolic arte." Functional Ecology **20**: 400-4004.

Gillooly, J. F., A. P. Allen, G. B. West and J. H. Brown (2005). "The rate of DNA evolution: effects of body size and temperature on the molecular clock." Proc Natl Acad Sci U S A **102**(1): 140-5.

Gillooly, J. F., J. H. Brown, G. B. West, V. M. Savage, et al. (2001). "Effects of size and temperature on metabolic rate." Science **293**(5538): 2248-51.

Glazier, D. S. (2005). "Beyond the '3/4-power law': variation in the intra- and interspecific scaling of metabolic rate in animals." Biol Rev Camb Philos Soc **80**(4): 611-62.

Glazier, D. S. (2010). "A unifying explanation for diverse metabolic scaling in animals and plants." Biol Rev Camb Philos Soc **85**(1): 111-38.

Haddrill, P. R., B. Charlesworth, D. L. Halligan and P. Andolfatto (2005). "Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content." Genome Biol **6**(8): R67.

Hoegg, S., H. Brinkmann, J. S. Taylor and A. Meyer (2004). "Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish." J Mol Evol **59**(2): 190-203.

Huang, Y. F. and D. K. Niu (2008). "Evidence against the energetic cost hypothesis for the short introns in highly expressed genes." BMC Evol Biol **8**: 154.

Hurst, L. D., G. McVean and T. Moore (1996). "Imprinted genes have few and small introns." Nat Genet **12**(3): 234-7.

Jaillon, O., J. M. Aury, F. Brunet, J. L. Petit, et al. (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-57.

Josse, J., A. D. Kaiser and A. Kornberg (1961). "Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid." J Biol Chem **236**: 864-75.

Kasahara, M., K. Naruse, S. Sasaki, Y. Nakatani, et al. (2007). "The medaka draft genome and insights into vertebrate genome evolution." Nature **447**(7145): 714-9.

Killen, S. S., D. Atkinson and D. S. Glazier (2010). "The intraspecific scaling of metabolic rate with body mass in fishes depends on lifestyle and temperature." Ecol Lett **13**(2): 184-93.

Kleiber, M. (1932). "Body size and metabolic rate." Physiological reviews **27**: 511-541.

Kolokotrones, T., S. Van, E. J. Deeds and W. Fontana (2010). "Curvature in metabolic scaling." Nature **464**(7289): 753-6.

Li, S. W., L. Feng and D. K. Niu (2007). "Selection for the miniaturization of highly expressed genes." Biochem Biophys Res Commun **360**(3): 586-92.

Martin, A. P. and S. R. Palumbi (1993). "Body size, metabolic rate, generation time, and the molecular clock." Proc Natl Acad Sci U S A **90**(9): 4087-91.

Nelson, J. S. (2006). Fishes of the World John Wiley & Sons.

Rao, Y., Z. Wang, X. Chai, G. Wu, et al. (2010). "Selection for the compactness of highly expressed genes in Gallus gallus." Biology Direct **5**(1): 35.

Ravi, V. and B. Venkatesh (2008). "Rapidly evolving fish genomes and teleost diversity." Curr Opin Genet Dev **18**(6): 544-50.

Seoighe, C., C. Gehring and L. D. Hurst (2005). "Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction." PLoS Genet **1**(2): e13.

Shen, J. C., W. M. Rideout, 3rd and P. A. Jones (1994). "The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA." Nucleic Acids Res **22**(6): 972-6.

Stenoien, H. K. (2007). "Compact genes are highly expressed in the moss Physcomitrella patens." J Evol Biol **20**(3): 1223-9.

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

Vandepoele, K., W. De Vos, J. S. Taylor, A. Meyer, et al. (2004). "Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates." Proc Natl Acad Sci U S A **101**(6): 1638-43.

Varriale, A. and G. Bernardi (2006). "DNA methylation and body temperature in fishes." Gene **385**: 111-21.

Venkatesh, B. (2003). "Evolution and diversity of fish genomes." Curr Opin Genet Dev **13**(6): 588-92.

Versteeg, R., B. D. van Schaik, M. F. van Batenburg, M. Roos, et al. (2003). "The human transcriptome map reveals extremes in gene density, intron length, GC content,

and repeat pattern for domains of highly and weakly expressed genes." Genome Res **13**(9): 1998-2004.

Vinogradov, A. E. (2001). "Bendable genes of warm-blooded vertebrates." Mol Biol Evol **18**(12): 2195-200.

Vinogradov, A. E. (2004). "Compactness of human housekeeping genes: selection for economy or genomic design?" Trends Genet **20**(5): 248-53.

Vinogradov, A. E. (2005). "Noncoding DNA, isochores and gene expression: nucleosome formation potential." Nucleic Acids Res **33**(2): 559-63.

Vinogradov, A. E. and O. V. Anatskaya (2006). "Genome size and metabolic intensity in tetrapods: a tale of two lines." Proc Biol Sci **273**(1582): 27-32.

West, G. B., J. H. Brown and B. J. Enquist (1997). "A general model for the origin of allometric scaling laws in biology." Science **276**(5309): 122-6.

West, G. B., J. H. Brown and B. J. Enquist (1999). "The fourth dimension of life: fractal geometry and allometric scaling of organisms." Science **284**(5420): 1677-9.

Zhu, J., F. He, S. Hu and J. Yu (2008). "On the nature of human housekeeping genes." Trends Genet **24**(10): 481-4.

Zhu, L., Y. Zhang, W. Zhang, S. Yang, et al. (2009). "Patterns of exon-intron architecture variation of genes in eukaryotic genomes." BMC Genomics **10**: 47.

3 Functional Organization of Mammalian Genomes

3.1 Introduction

The most striking feature of the human genome was his mosaic structure (Bernardi, Olofsson et al., 1985). Comparing the density gradient profile of several vertebrates, major differences were observed in the compositional pattern during transition from poikilotherms to homeotherms. Genomes of amphibian/reptile showed a narrow compositional range, while the mammalian genomes are remarkably marked by the appearance of GC-rich isochores (Bernardi, 2004, for review). With the aim to study the nature of the forces driving the transition mode of vertebrate evolution, a different strategy was designed to analyze mammalian genomes. Applying a reverse approach, i.e. extracting the information about functional annotation of genes and then going back to the composition properties, firstly thirteen mammalian genomes were compared to highlight common conserved features and secondly, human genome was compared to amphibian and reptile in order to show genomic differences between poikilotherms and homeotherms. This approach got its background from a functional classification system based on orthologous relationships between genes, i.e. KOG database.

KOG Database

KOG (acronym for eu**K**aryotic **O**rthologous **G**roups) presents clusters of predicted orthologs for eukaryotic genomes (Tatusov, Natale et al., 2003). The database is a major update of the previously developed system for delineation of the database of **C**luster of **O**rthologous **G**roups of proteins (COG) representing a phylogenetic classification of proteins encoded in complete genomes of prokaryotes and unicellular eukaryotes (Tatusov, Galperin et al., 2000; Tatusov, Natale et al., 2001). KOGs include proteins from seven eukaryotic genomes: three animals (the nematode *C. elegans*, fruit fly *D. melanogaster* and *H. sapiens*), one plant (*A. thaliana)*, two yeasts (*S. cerevisiae* and *S. pombe*), and the intracellular microsporidian parasite *E. cuniculi*, with upcoming six new species mainly including *M. musculus* and *R. norvegicus*. Presence of few mammalian classifications in KOG, prompted us to analyze more genomes in order to have a better picture of base compositional pattern among mammalian genomes. Specifically KOG classification of human genes allowed performing comparative compositional analysis of different functional classes in several mammalian and two non-mammalian genomes. Regarding base composition, GC3 (the GC content at the third codon, or wobble, position) has been found to play a key role in understanding the genome organization. In fact, preliminary analysis of human genes grouped in functional classes according to the KOG database showed a biased distribution of the GC3 content that was significantly higher in the functional classes of genes involved in metabolic processes (D'Onofrio, Ghosh et al., 2007). Using orthology approach the analysis of the KOG categories and functional classes of genes was extended to thirteen completely sequenced mammalian genomes, as well as to amphibian (*X. tropicalis*) and reptile (*A. carolinensis*) genomes.

3.2 Results

3.2.1 Human KOG genes In the KOG database human genes were classified in 25 functional classes, denoted by capital letter in square brackets (Tatusov, Natale et al., 2001; Tatusov, Natale et al., 2003) and were grouped into three main categories: i) information storage and processing, represented by five functional classes; ii) cellular processes and signaling, represented by ten functional classes; and iii) metabolism, represented by eight functional classes. For sake of simplicity, theses three categories from this point on will be referred as Blue, Black, and Red respectively. The Blue category accounted for 22%, Black 42% and Red 16% of the whole dataset. In order to have a better compositional insight, the complete list of functional classes, corresponding number of genes and average GC3 content of each functional class were reported in Table 3.1. The average GC3 content of the human genome was very close to that of the KOG dataset, 58.5% and 58.4%, respectively). Within the human genome, the average GC3 of the three categories was found to be significantly different (Figure 3.1).

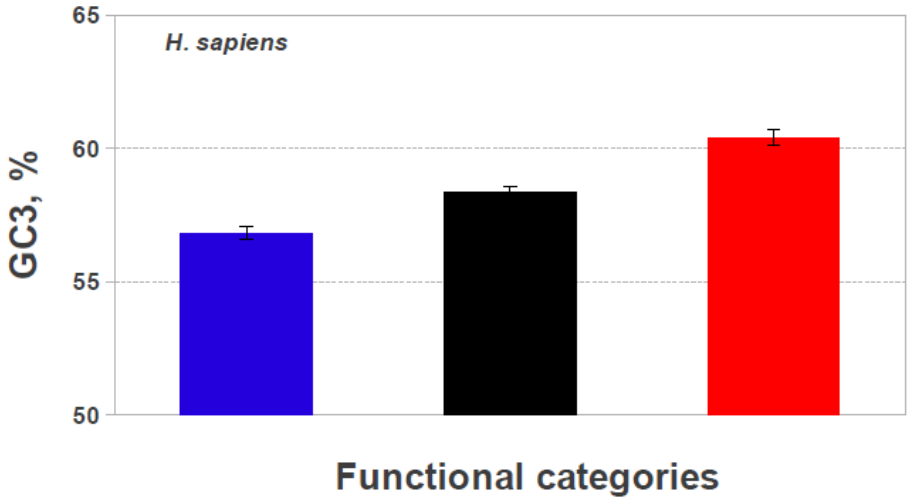


Figure 3.1: The histogram shows the average GC3 content in the three functional categories of the human genome**:** (i) information storage and processing (Blue bar); (ii) cellular processes and signaling (Black bar); (iii) metabolism (Red bar). For each histogram bar standard error is reported.

Determining the statistical significances of the differences observed between three categories using two sided Mann-Whitney test (see Appendix 1, section 6.5.2) showed, the GC3 of Red category was significantly higher than that of both Black and Blue ones ($p< 2.2x10^{-16}$ and $p< 4.2x10^{-8}$, respectively). In turn, the GC3 of the Black category was significantly higher than that of Blue one ($p<1.6x10^{-6}$). In short, in case of human genome the average GC3 content of the three functional categories showed the following trend: Blue<Black<Red (Figure 3.1).

3.2.2 Classification of vertebrate KOG genes

Using the best reciprocal hits approach, the genes of thirteen mammalian genomes (see Appendix 1, section 6.2) representing the following orders: **primates** (*G. gorilla* and *P. Pygmaeus*), **rodentia** (*M. musculus* and *S. tridecemilineatus*), **lagomorpha** (*O. cuniculus*), **artiodactyla** (*B. taurus*), **perissodactyla** (*E. caballus*), **chiroptera** (*P. vampyrus*), **cetacea** (*T. truncatus*), **proboscidea** (*L. africana*), **cingulata** (*D. novemcintctus*), **didelphimorphia** (*M. domestica*) and **monotremes** (*O. anatinus*) were classified in the KOG functional classes through the orthology with the human ones. The same approach was used to classify the genes of the species representing the order of **anura** (*X. tropicalis*) and **squamata** (*A. carolinensis*) genes. A gene of each species found to be orthologous with a specific human gene, acquired the same KOG classification. In other words, throughout orthology the KOG classification was extended to the genes of the 15 species so far analyzed. For each genome: the whole number of available genomic CDS, the subset of genes classified according to the KOG database, the amount of genes belonging to the Blue, Black and Red categories, as well as the corresponding GC3 content and the standard deviation, were reported in Table 3.2.

Table 3.1: Classification of Human Genes.

KOG classes	Categories	#	GC3
	INFORMATION STORAGE AND PROCESSING		
[A]	RNA processing and modification	600	0.517
[B]	Chromatin structure and dynamics	224	0.610
[J]	Translation, ribosomal structure and biogenesis	1237	0.545
[K]	Transcription	1137	0.619
[L]	Replication, recombination and repair	300	0.546
	CELLULAR PROCESS AND SIGNALING		
[D]	Cell cycle control, cell division, chromosome partitioning	267	0.552
[M]	Cell wall/membrane/envelope biogenesis	63	0.576
[N]	Cell motatility	26	0.586
[O]	Post transcriptional modification, protein turnover, Chaperons	1417	0.557
[T]	Signal transduction mechanism	2214	0.616
[U]	Intracellular trafficking, secretion and vesicular transport	685	0.571
[V]	Defense mechanism	1023	0.527
[W]	Extracellular structure	284	0.588
[Y]	Nuclear structure	17	0.534
[Z]	Cytoskeleton	801	0.638
	METABOLISM		
[C]	Energy production and conversion	403	0.576
[E]	Amino acid transport and metabolism	499	0.618
[F]	Nucleotide transport and metabolism	187	0.588
[G]	Carbohydrate transport and metabolism	469	0.606
[H]	Coenzyme transport and metabolism	102	0.563
[I]	Lipid transport and metabolism	410	0.595
[P]	Inorganic ion transport and metabolism	402	0.646
[Q]	Secondary metabolites biosynthesis, transport and catabolism	191	0.591
	POORLY CHARACTERIZED		
[R]	General function prediction only	1889	0.593
[S]	Function unknown	1171	0.568
	Total No. of genes and Average GC3	16118	0.561

(#) Number of genes

Table 3.2: Average GC3 content, standard deviation and number of genes in KOG's functional categories.

		KOG**			BLUE*			BLACK*			RED*		
Mammals	**Organism**	**GC3**	**s.d.**	**#**	**GC3**	**s.d.**	**#**	**GC3**	**s.d.**	**#**	**GC3**	**s.d.**	**#**
	H. sapiens	0.584	0.159	12942	0.568	0.154	3534	0.584	0.163	6745	0.604	0.155	2663
	G. gorilla	0.609	0.162	6268	0.593	0.166	1491	0.609	0.166	3357	0.626	0.148	1420
	P. pygmaeus	0.594	0.164	7455	0.583	0.167	1766	0.593	0.166	4012	0.611	0.154	1677
	M. musculus	0.606	0.114	7505	0.596	0.127	1780	0.605	0.112	4032	0.617	0.101	1693
	O. cuniculus	0.630	0.175	5413	0.609	0.179	1296	0.629	0.177	2867	0.653	0.164	1250
	S. tridecemilineatus	0.565	0.154	5455	0.542	0.157	1325	0.567	0.155	2900	0.584	0.144	1230
	B. taurus	0.630	0.167	7139	0.613	0.171	1706	0.632	0.169	3794	0.642	0.155	1639
	E. caballus	0.609	0.164	7102	0.594	0.170	1649	0.608	0.165	3840	0.626	0.153	1613
	P. vampyrus	0.605	0.164	6780	0.590	0.170	1638	0.607	0.165	3607	0.619	0.155	1535
	T. truncatus	0.618	0.167	6812	0.602	0.172	1635	0.617	0.169	3634	0.635	0.155	1543
	L. africana	0.583	0.152	5704	0.575	0.157	1413	0.582	0.153	3007	0.595	0.141	1284
	D. novemcinctus	0.585	0.180	5358	0.563	0.181	1310	0.585	0.182	2832	0.607	0.173	1216
	O. anatinus	0.648	0.166	5287	0.646	0.169	1319	0.645	0.166	2734	0.657	0.161	1234
	M. domestica	0.533	0.145	3598	0.529	0.153	1641	0.531	0.144	3598	0.542	0.137	1578
Reptile		0.539	0.159	5959	0.535	0.158	3063	0.546	0.1655	1498	0.539	0.153	1398
Amphibian		0.500	0.112	3584	0.499	0.112	1753	0.499	0.1174	961	0.501	0.108	870

(*) Blue, Black and Red refers to the gene classification of KOG database (see Appendix 1, 6.1.3).
(**) Genes orthologous to KOG human genes.
(#) Number of genes

3.2.3 Base composition of KOG genes Interestingly, the trend of the GC3 content observed in the human genome, *i.e.* Blue<Black<Red, was also found in all mammalian genomes (except *O. anatinus*), but not in the amphibian and reptile genomes (Figure 3.2).

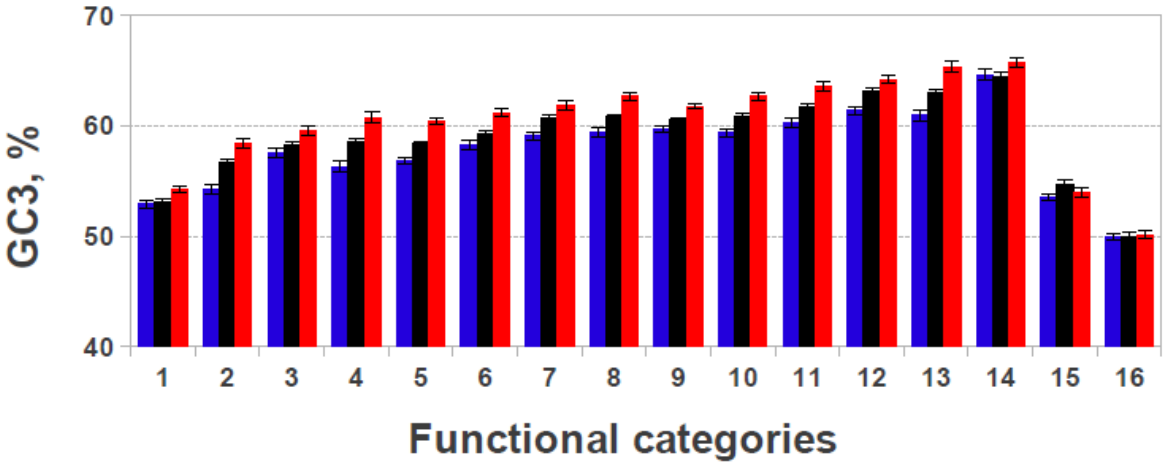


Figure 3.2: The histogram of the average GC3 content in the three functional categories in all analyzed genomes. Color codes as in Figure 3.1. (Standard error is reported). Genome legend: *M. domestica* (1), *S. tridecemilineatus* (2), *L. africana* (3), *D. novemcintctus* (4), *H. sapiens* (5), *P. pygmaeus* (6), *P. vampyrus* (7), *E. caballus* (8), *M. musculus* (9), *G. gorilla* (10), *T. truncatus* (11), *B. taurus* (12), *O. cuniculus* (13), *O. anatinus* (14), *A. carolinensis* (15), *X. tropicalis* (16).

For almost all mammals the average GC3 was significantly different among the three categories (Red, Blue and Black), with the exception for *B. taurus*, showing not highly significant differences between Red and Black categories (*p*-value = 0.1007), and for non-placental mammals (*O. anatinus* and *M. domestica*) the differences between Blue and Black categories were not significant at all (*p*-value of 0.7639 and 0.3252 respectively) (Table 3.3).

Table 3.3: *p-values* of the Mann-Whitney test among categories.

Class	Order	Species	Red vs. Blue	Red vs. Black	Black vs. Blue
Mammals	**Primates**	*H. sapiens*	hs	hs	hs
		G. gorilla	hs	0.0027	0.0012
		P. Pygmaeus	hs	0.0001	0.03
	Rodentia	*M. musculus*	hs	0.0004	0.0006
		S. tridecemmilineatus	hs	0.0022	hs
	Lagomorpha	*O. cuniculus*	hs	0.0001	0.0005
	Artiodactyla	*B. taurus*	hs	0.107	0.0002
	Perissodactyla	*E. caballus*	hs	0.0002	0.0023
	Chiroptera	*P. vampyrus*	hs	0.0276	0.0005
	Cetacea	*T. truncatus*	hs	0.0012	0.0028
	Proboscidea	*L. africana*	0.0002	0.0101	0.0916
	Cingulata	*D. novemcintctus*	hs	0.0001	0.0002
	Didelphimorphia	*M. domestica*	0.0015	0.005	0.3252
	Monotremes	*O. anatinus*	0.0958	0.0224	0.7639
Reptiles	**Squamata**	*A. cardiensis*	0.6258	0.1748	0.0625
Amphibians	**Anura**	*X. tropicalis*	0.4994	0.8865	0.5715

hs = Highly significant p-value lower than at least 10^{-5}

The null hypothesis is that the two variables have equal distributions

3.2.4 de Finetti's diagram

In order to compare across the different genomes the impact of the Blue<Black<Red relation, a descriptive analysis of the distribution of the functional categories over a GC3 range was performed. The results reported in the **de Finetti's** diagram (Figure 3.3) (details, Appendix 1, Section 6.5.3), showed the distance of a given point from a given side accounting for the frequency of a given category in one of the three GC3 ranges, namely Low, Medium and High. Clearly, short distances accounted for low frequencies.

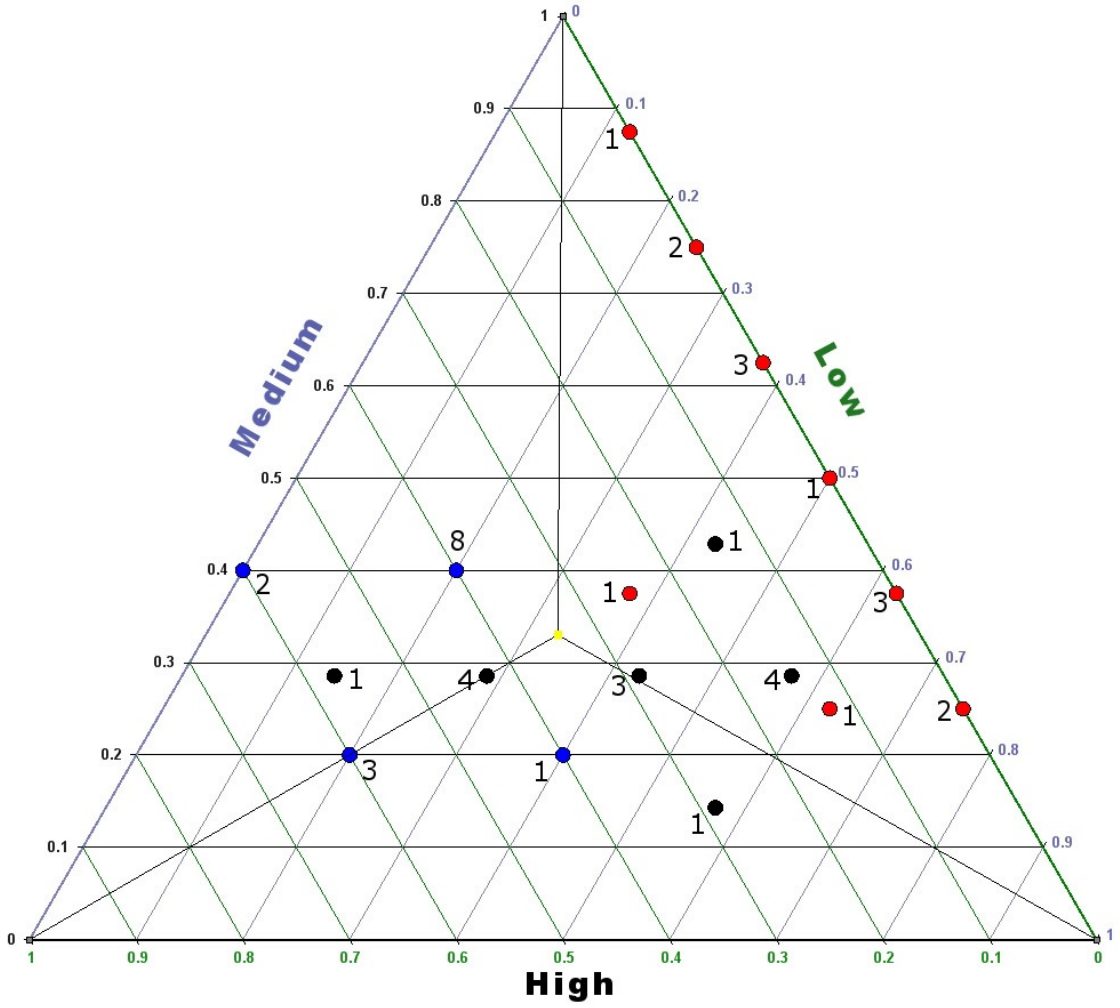


Figure 3.3: de Finetti's diagram shows the spatial distribution of the three functional categories: (i) information storage and processing (Blue dots); (ii) cellular processes and signaling (Black dots); (iii) metabolism (Red dots). Numbers close to dots refer to the occurrence of overlapping genomes.

Examining Figure 3.3, following points were observed**:** i) the Red category was rarely present in the lowest GC3 range, therefore confined to a restricted part of the space of the diagram (i.e. the upper right side of the diagram); ii) no overlap was observed between the spatial distribution of the Red categories with that of the Blue category; and iii) partial spatial overlap was observed between the Blue and Black categories which didn't show any specific distribution, but were rarely present at high GC3 content. The significance of assumption was tested performing 1000 class permutations and observing the diagram distribution of the Red classes. The probability to reproduce the spatial distribution of the Red category by chance confined in the upper right side of the diagram was estimated to be $p<1.76x10^{-2}$. On the contrary, the spatial distribution of the three categories was not significantly different from each other in frog (*X. tropicalis*) and lizard genomes (*A. carolinensis*) (data not shown).

3.2.5 The Butterfly plot

For a deeper insight into the functional organization of the genomes so far analyzed, within each genome the average GC3 of each functional class (Table 3.1, for definition) was investigated. More precisely, the difference between the GC3 content of each functional class against that of the corresponding genome was calculated (ΔGC3). By clustering the negative and positive values a so-called butterfly plot (name given for sake of simplicity) for each genome was obtained. An overview of the all butterfly plots was reported in Figure 3.4, where the color code of the functional classes corresponds to the three main functional categories and detailed representations of each genome were reported in supplementary materials (Supplementary materials Figure S4-S8). Although at first glance, the butterfly plots of the mammalian genomes appeared as an unbalance distribution of the bars (Figure 3.4), a close analysis of the butterfly-plots showed that the bars of the Blue and Black categories were mainly on the negative side of the histogram (i.e. showing an average GC3 content lower than the corresponding genomic one), whereas those of the Red category were mainly on the positive side.

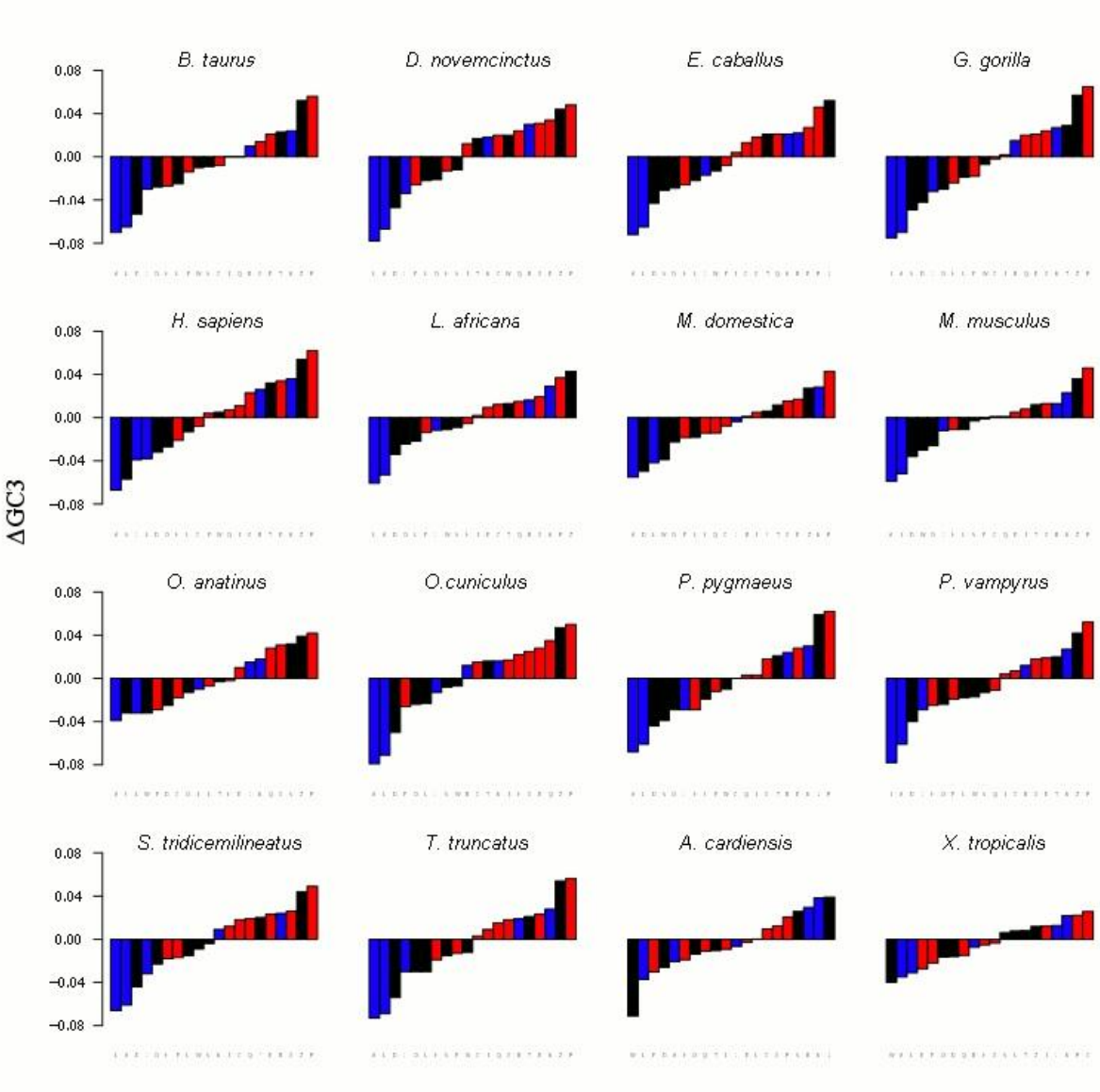


Figure 3.4: Butterfly plots: Histograms of the delta between average genomic GC3 content against that of each functional class within each genome. Color coding is according to the three main categories: (i) Information storage and processing (Blue); (ii) Cellular processes and signaling (Black); and (iii) metabolism (Red).

3.2.6 Mammalian vs. amphibian and reptile

In case of human genome, which could be considered as representative of all mammals so far analyzed, only two out of five blue classes (*i.e.* B and K) and three out of seven black classes (*i.e.* T and Z) were on the positive side of the butterfly plot (Figure 3.5, panel C). On contrary, six out of eight Red classes (*i.e.* F, C, I, G, Q, E and P) were in the positive side of the butterfly plot (Figure 3.5, panel C). This picture was recurrently found among mammalian genomes (Figure 3.4 and Suppl. Figure S4-S8). The percentage of occurrence was reported below each functional class. For example functional class B, for instance, was on the positive side in 88% of the mammalian genomes, whereas the classes G, K, E, Z and P were on the positive side in 100% of the cases (Figure 3.5, panel C). To determine which classes have a GC3 content significantly higher than of the whole genome a t-Student's test with Bonferroni's correction (α=0.05) was performed. The following functional classes: K (Blue), T and Z (Black) and G, E and P (Red), turned out to have an average GC3 content significantly higher than that of the whole genome (labeled by an asterisk in Figure 3.5, panel C). The butterfly plots of *X. tropicalis* and *A. carolinensis* were also reported (Figure 3.5, panel A and B, respectively). As already mentioned above (Figure 3.2), these two genomes clearly showed a different genome organization and, once again here, they were clearly differentiated from mammalian organization. Regarding the lizard genome only two functional classes, namely the K and the Z classes showed an average GC3 content significantly higher than the genomic one (Figure 3.5, panel B), while none of the frog functional classes turned out to be significantly different (Figure 3.5, panel A).

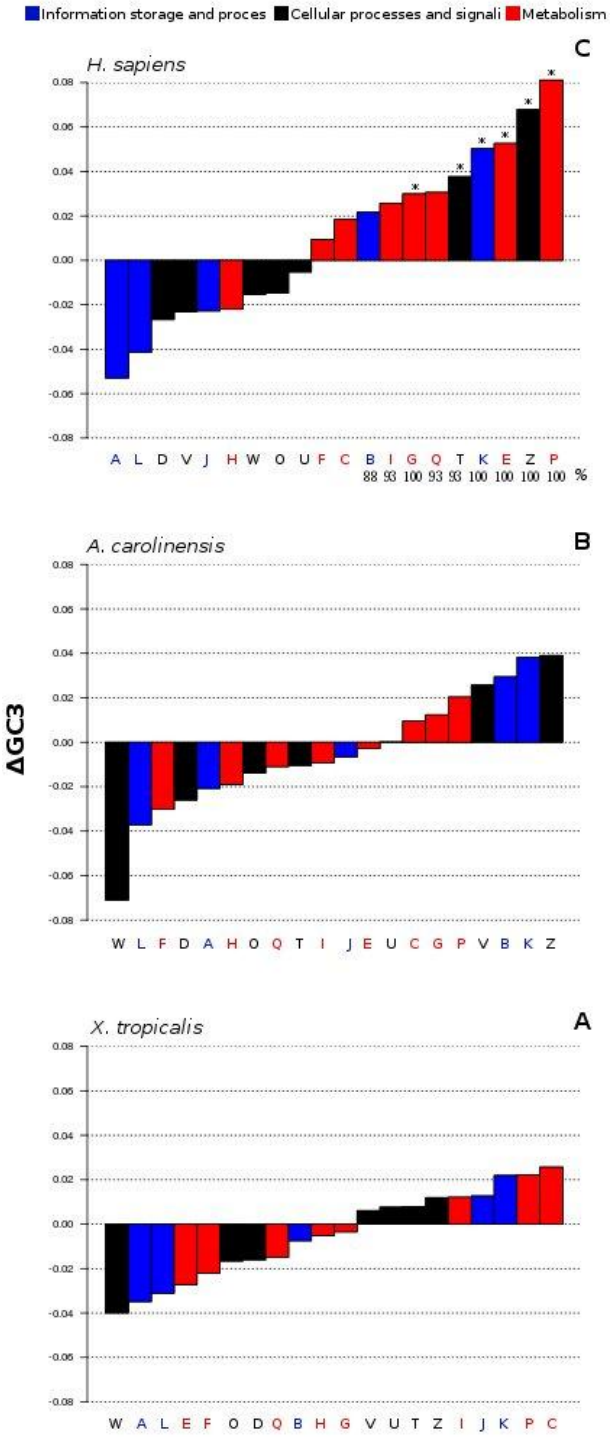


Figure 3.5: The butterfly plot of frog (panel A) lizard (panel B) and human functional classes (panel C). Color code of histogram bars as in Fig. Z.1. Asterisks show bars with an average GC3 content significantly higher than the genomic one (Bonferroni's test, α=0.05).

Finally with the aim to check which among the three functional categories underwent the highest GC3 increment during the transition from cold- to warm-blooded vertebrates, a comparative compositional analysis of the functional categories between human and frog (H/F), and between human and lizard (H/L) was performed. More precisely, the GC3 increment (ΔGC3) was investigated. In both comparisons positive values of ΔGC3 were observed in the human categories (Figure 3.6). Interestingly, the ΔGC3 increment in the Red category was the highest in both H/F and H/L comparisons. Indeed, ΔGC3 of the Red category was significantly higher than that of the Black ($p<4.72x10^{-4}$ and $p<1.14x10^{-2}$, respectively for the H/F and H/L comparison), and significantly higher than that of the Blue category ($p<4.56x10^{-5}$ and $p<2.27x10^{-16}$, respectively for the H/F and H/L comparison). Hence considering the GC3 content of each category in human vs. frog and lizard, the highest increment was found in the Red category (Figure 3.6) that corresponds to metabolism.

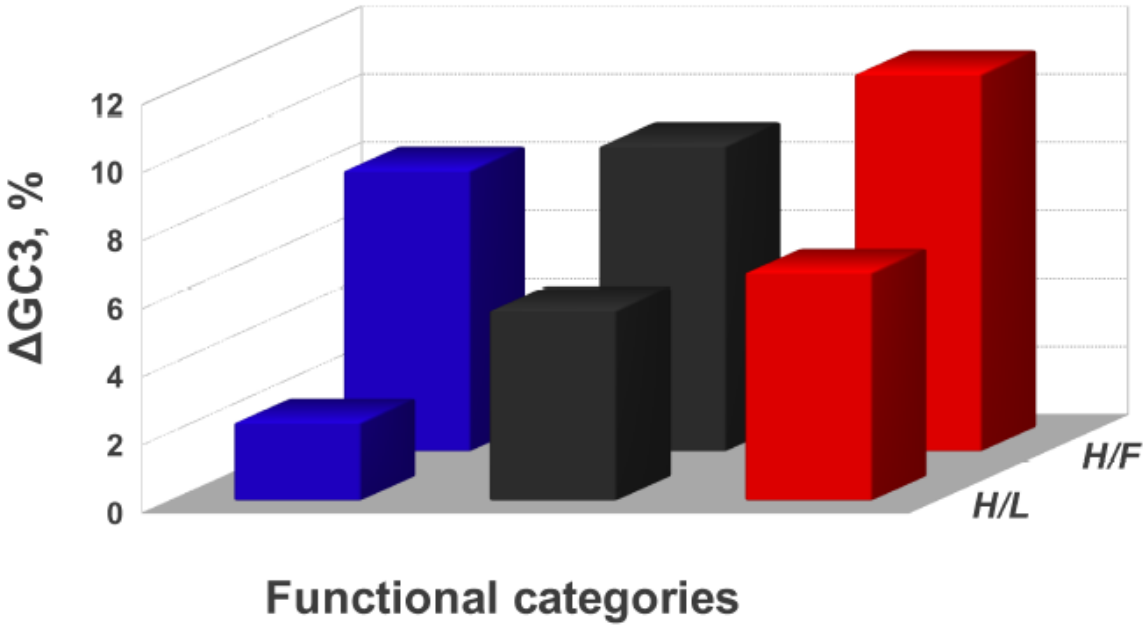


Figure 3.6: The histogram shows the average GC3 increment in the three functional categories comparing human *vs.* frog *(H/F)* and human *vs.* lizard *(H/L)*. Color code of histogram bars as in Figure 3.1.

3.2.7 Chromosomal bands

Human KOG genes were assigned to the different compositional band types (L1+, L1-, H3-, H3+ bands) previously identified in the human chromosomes (Figure 3.7). Genes were divided in two groups, i.*e.* positive and negative, according to the position of the corresponding functional class in the butterfly plot. More specifically, in the first group were included genes belonging to the KOG classes A, L, D, V, J, H, W, O and U, while in the second were included those belonging to the F, C, B, I, G, O, T, K, E, Z and P classes (Figure 3.5, panel C). Small differences were found between the two groups in the L1+ and L1- bands (about 2%), whereas in the H3- and H3+ bands the differences increased up to 5% and 7%, respectively. The probability to find genes of the positive group in the H3+ bands was significantly higher than in the L1+ bands (*z*-test one tail, $p<<10^{-2}$).

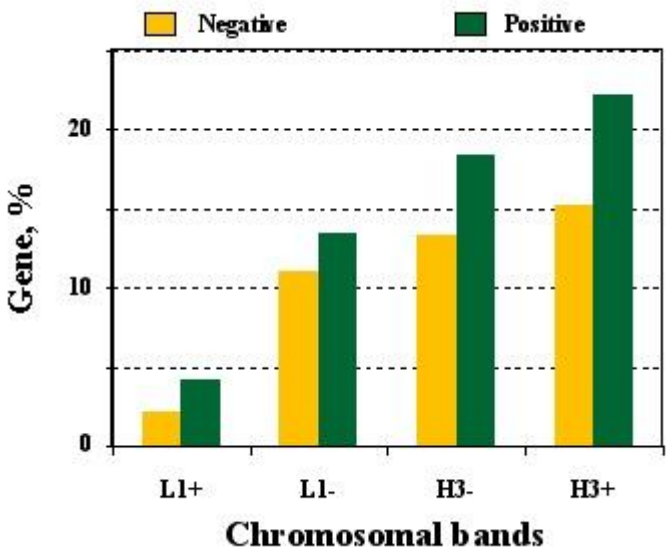


Figure 3.7: Histogram showing the gene distribution in the four types of human chromosomal bands. Positive and negative refers to the position of KOG functional classes in the butterfly plot of *H. sapiens* (Figure 3.5, panel C).

3.3 Discussion

Since the pioneering Ikemura's papers (Ikemura, 1981,1985), till nowadays the GC3 content, accounting for the base composition at the third, or wobble, position in a codon, has been generally associated mainly with the codon usage and with the tRNA content. Further studies, primarily performed on the human genome, showed that GC3 should be considered a keystone parameter to understand genome evolution. Indeed, GC3 turned out to be significantly correlated with the amino acid frequencies, *i.e.* GC1+2, as well as with the GC content of non-coding regions, *i.e.* introns and flanking regions (Aissani, D'Onofrio et al., 1991; D'Onofrio, Mouchiroud et al., 1991; D'Onofrio and Bernardi, 1992). Recent attempts to disregard the pivotal role of the GC3 parameter in understanding the genome organization (Elhaik, Landan et al., 2009), failed to take into consideration that "the use of indirect methods can lead to apparently conflicting conclusions" (Clay and Bernardi, 2010). Recently, the role of the GC3 parameter as genome marker was further confirmed by the unexpected finding of correlations with genome size and body mass of mammals (Romiguier, Ranwez et al., 2010). The subset of KOG human genes analyzed in the present paper followed, as expected, the well assessed rules first described in the 90's (Aissani, D'Onofrio et al., 1991; D'Onofrio, Mouchiroud et al., 1991; D'Onofrio and Bernardi, 1992). These rules held not only for the whole set of genes, but also when the genes were grouped in the KOG functional classes (see S9, for statistical reports).

In order to shed light on the debate around the evolutionary forces shaping the base composition among and within genomes (Chapter 1), instead of starting from the compositional properties of the genes and afterwards inferring biological properties, the opposite pathway was followed. Genes were first classified in the three functional categories [*i.e.* (i) information storage and processing (Blue); (ii) cellular processes and signaling (black); (iii) metabolism (Red)] and then the base compositional properties were analyzed. The results showed that within mammalian genomes the three functional categories were characterized by a different GC3 content, following the pattern Blue<Black<Red (Figure 3.2 and Figure 3.3). No pattern was found in the reptile and amphibian genomes (Figure 3.2).

Do current hypotheses could explain the above finding?

It is worth to bring to mind that the keystone of the biased gene conversion hypothesis (BGC) was the strong correlation between hot spot recombination sites and GC content, establishing a cause/effect link of the first over the second parameter (Eyre-Walker, 1993). Consequently, the genomic impact of the BGC would be an increment of the GC content detectable at non-synonymous sites, synonymous sites, flanking and intronic sequences (Duret and Galtier, 2009). Although remaining essentially a neutral process (Galtier, Piganeau et al., 2001; Duret and Galtier, 2009), BGC was reported to mimic perfectly natural selection. Thus, the compositional correlations holding in the human genome, including those reported in **S9**, could not be considered in favor of natural selection hypothesis. In the light of BGC, the Blue<Black<Red pattern found in all mammals could have been explained as the result of the star-like phylogeny of mammals (Kumar and Hedges, 1998). However, comparative genome analyses showed that hot spot recombination sites are "highly mobile" and therefore not phylogenetically related (Huang, Friedman et al., 2005). A result further supported by the studies conducted on the fast-evolving DNA-binding domain of PRDM9, identified as a major hotspot determinant of recombination. Indeed, the sequences and the number of PRDM9 domains were reported to vary a lot among species (reviewed in (Hochwagen and Marais, 2010)). Considering that BGC was reported to be a widespread process affecting eukaryotic genomes, the lack of the Blue<Black<Red pattern in both frog and lizard genomes at present stands unclear. Although BGC received experimental support from the analysis of the short sequences HARs and HACNSs in the human genome (Duret and Galtier, 2009), the hypothesis was unable to explain the base compositional variability among bacterial genomes (Hildebrand, Meyer et al., 2010).

An interesting alternative hypothesis to the BGC was recently proposed by Lemaitre and colleagues based on the analysis of the DNA breakpoint regions (BPR) (Lemaitre, Zaghloul et al., 2009). Very recently, indeed, a 3D analysis of BPR showed that "two loci distant in the human genome but adjacent in the mouse genome are significantly more often observed in close proximity in the human nucleus than expected"

(Veron, Lemaitre et al., 2011). The conservation of the Blue<Black<Red pattern among mammals, that started to diverge about 100 Mya (Kumar and Hedges, 1998), could probably be explained by the fact that 3D chromatin structure could be conserved over long evolutionary distances. The time of divergence between amniotes and amphibian and between mammals and lizard was estimated to be several orders of magnitude greater than that of mammalian radiation (340-370Mya and ~310Mya, respectively). Therefore explaining why the pattern was not conserved in reptiles and amphibians. However, according to the BPR hypothesis, evolutionary rearrangement breakages happen with a uniform propensity along the genome (Lemaitre, Zaghloul et al., 2009), leaving unexplained how the Blue<Black<Red pattern, absent in frog and lizard (Figure 3.6), could have been evolved in mammals. Moreover, as far as we know, no evidence has been produced to explain the base compositional variability among bacterial genomes in the light of the BPR hypothesis.

The critical query (Blue<Black<Red pattern) could be explained, on the contrary, by both thermal stability and metabolic rate hypotheses (Vinogradov, 2001; Bernardi, 2004, for review; Vinogradov, 2005). Indeed, *in-situ* hybridization experiments performed on both human and amphibian nuclei (*i.e. Rana esculenta*), showed a comparable chromatin organization (Saccone, Federico et al., 1999; Federico, Scavo et al., 2006). In both genomes, GC-poor regions were found in the compact, or closed, chromatin structures localized at the nuclear periphery, while GC-rich ones were found in an open chromatin structures localized in the internal region of the nuclei. According to the above reports, the different living temperature experienced by amphibians and mammals, could induce an increment of the GC content in mammals, in order to stabilize the open chromatin structures (Bernardi, 2004, for review). On the other hand, an increment of the metabolic rate, well known to be higher in mammals, should induce an increment of the GC content to increase DNA bendability, on one hand, and decrease nucleosome formation potential, on the other, due to an increment of the transcriptional activity (Vinogradov, 2001, 2005). To this regard it should be recalled that along human chromosome the GC content and the gene expression profiles showed a positive correlation (Versteeg, van Schaik et al., 2003).

Temperature and metabolic rate are well known to be strongly correlated (Kleiber, 1932). Therefore, disentangle the two variables would be unfortunately not an easy task in the light of present data, also considering that terrestrial animals are living in an environment where oxygen is not a limiting factor. The problem was tackled in the present paper analyzing the orthologous pairs of human/frog (H/F) and human/lizard (H/L) genes. In both cases, the highest ΔGC3 turned out to take place in the Red category that is the functional category grouping genes involved in metabolic processes (Figure 3.6). Although not resolving the dichotomy between temperature and metabolic rate, the result was congruent with the conclusion drawn out from the comparison of teleostean fish genomes (Uliano, Chaurasia et al., 2010; Chaurasia, Uliano et al., 2011).

The detailed investigation on the distribution of the KOG functional classes revealed that the Blue<Black<Red pattern was even more multifaceted (Figure 3.4). Indeed, in the positive side of the human butterfly plot, apart the majority of Red bars, the B and K blue bars, as well as the T and Z black bars were also observed (Figure 3.5, panel C).The above picture was not confined to the human genome, but commonly found in all mammals. Indeed, the B and T classes were in the positive side of the butterfly plot in the 93% of the cases, whereas the K and Z classes reached the 100% of the cases (Figure 3.5, panel C). The occurrence of the bars belonging to the Red category ranged from 86% of the Q class to 100% of the G, E and P classes. Needless to say, the pattern was not found in the frog and lizard genomes. All the considerations formerly drawn out in the light of the different evolutionary hypotheses regarding the Blue<Black<Red pattern, applied even more radically to the pattern of functional classes clustering in the positive side of all mammalian butterfly plots (Figure 3.4 and Figure 3.5, panel C).

The above result deserves a more detailed argumentation. As reported in Table 3.1, the genes belonging to the four classes were involved in the following task: Chromatin structure and dynamics (B), Transcription (K), Signal transduction mechanisms (T) and Cytoskeleton (Z). The fact that the GC3 content of genes belonging to the B and K classes was not surprising, since an increment of the metabolic rate affects transcription process and chromatin structure, as discussed above. More inscrutable was the result regarding the T and Z classes. Recently, an interesting paper was published on

the effect of estrogen exposure in mice brain, inducing an increment of the expression level of a discrete number of genes (Szego, Kekesi et al., 2009). Beside a 39% of genes involved in metabolic processes, 18% belonged to the Z class and 25% to the T class, whereas only 6% of the genes belonged to the category grouping genes involved in information storage and processing. In the light of Szego's and colleagues report, the high probability of genes belonging to the B, K, T, and Z class, and of course those involved in metabolic processes, to cluster in the H3+ chromosomal bands (Figure 3.7) was an interesting preliminary result, pointing towards further investigations on the link between genome organization and the physiological reaction to stressing stimuli increasing the metabolic rate. Interestingly, gene clusters for metabolic pathways have been reported also in plants, reviewed in (Osbourn, 2010).

3.4 Conclusions

All the different evolutionary hypotheses proposed till now as discussed above and as in Chapter 1, surely contribute, with different weight, to the compositional variability observed among and within organisms (Pozzoli, Menozzi et al., 2008). However few seem to fit with the very wide range from prokaryotes to eukaryotes (see section 1.2). In the frame of the adaptive forces, unfortunately, present data neither shed light in favor nor against the effect of temperature on the compositional transition from "cold to warm-blooded vertebrates" (Bernardi and Bernardi, 1990; D'Onofrio, Jabbari et al., 1999; D'Onofrio and Ghosh, 2005). On the contrary, all the results discussed above from the perspective of mammalian genomes were in congruence with the conclusion drawn out from the comparison of teleostean fish genomes (Chapter 2) (Uliano, Chaurasia et al., 2010; Chaurasia, Uliano et al., 2011). Hence metabolic rate hypothesis can explain both the transition and the shifting mode of evolution of vertebrate genomes (Vinogradov, 2001, 2005; Vinogradov and Anatskaya, 2006).

3.5 References

Aissani, B., G. D'Onofrio, D. Mouchiroud, K. Gardiner, et al. (1991). "The compositional properties of human genes." J Mol Evol **32**(6): 493-503.

Bernardi, G. (2004). Sructural and Evolutionary Genomics. Natural Selection in Genome Evolution. Amsterdam, Elsevier.

Bernardi, G. and G. Bernardi (1990). "Compositional transitions in the nuclear genomes of cold-blooded vertebrates." J. Mol. Evol. **31**: 282-293.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, et al. (1985). "The mosaic genome of warm-blooded vertebrates." Science **228**(4702): 953-8.

Chaurasia, A., E. Uliano, B. L, C. Agnisola, et al. (2011). "Does habitat affects the genomic GC content? A lesson from teleostean fish – a mini review." Book chapter : Fish Ecology.

Clay, O. K. and G. Bernardi (2010). "GC3 of genes can be used as a proxy for isochore base composition: a reply to Elhaik et al." Mol Biol Evol **28**(1): 21-3.

D'Onofrio, G. and G. Bernardi (1992). "A universal compositional correlation among codon position." Gene **110**: 81-88.

D'Onofrio, G. and T. C. Ghosh (2005). "The compositional transition of vertebrate genomes: an analysis of the secondary structure of the proteins encoded by human genes." Gene **345**(1): 27-33.

D'Onofrio, G., T. C. Ghosh and S. Saccone (2007). "Different functional classes of genes are characterized by different compositional properties." FEBS Lett **581**(30): 5819-24.

D'Onofrio, G., K. Jabbari, H. Musto, F. Alvarez-Valin, et al. (1999). "Evolutionary genomics of vertebrates and its implications." Ann N Y Acad Sci **870**: 81-94.

D'Onofrio, G., D. Mouchiroud, B. Aïssani, C. Gautier, et al. (1991). "Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins." J Mol Evol **32**(6): 504-10.

Duret, L. and N. Galtier (2009). "Biased gene conversion and the evolution of mammalian genomic landscapes." Annu Rev Genomics Hum Genet **10**: 285-311.

Duret, L. and N. Galtier (2009). "Comment on "Human-specific gain of function in a developmental enhancer"." Science **323**(5915): 714; author reply 714.

Elhaik, E., G. Landan and D. Graur (2009). "Can GC content at third-codon positions be used as a proxy for isochore composition?" Mol Biol Evol **26**(8): 1829-33.

Eyre-Walker, A. (1993). "Recombination and mammalian genome evolution." Proc. R. Soc. Lond. **B 252**: 237-243.

Federico, C., C. Scavo, C. D. Cantarella, S. Motta, et al. (2006). "Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates." Chromosoma **115**(2): 123-8.

Galtier, N., G. Piganeau, D. Mouchiroud and L. Duret (2001). "GC-content evolution in mammalian genomes: the biased gene conversion hypothesis." Genetics **159**(2): 907-911.

Hildebrand, F., A. Meyer and A. Eyre-Walker (2010). "Evidence of selection upon genomic GC-content in bacteria." PLoS Genet **6**(9).

Hochwagen, A. and G. A. Marais (2010). "Meiosis: a PRDM9 guide to the hotspots of recombination." Curr Biol **20**(6): R271-4.

Huang, S. W., R. Friedman, N. Yu, A. Yu, et al. (2005). "How strong is the mutagenicity of recombination in mammals?" Mol Biol Evol **22**(3): 426-31.

Ikemura, T. (1981). "Correlation between the abundance of E. coli transfer RNAs and the occurrence of the respective codons in its protein genes : a proposal for a synonymous codon choice that is optimal for the E. coli translation system." J. Mol. Biol. **158**: 573-597.

Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms." Mol Biol Evol **2**(1): 13-34.

Kleiber, M. (1932). "Body size and metabolic rate." Physiological reviews **27**: 511-541.

Kumar, S. and S. B. Hedges (1998). "A molecular time scale for vertebrate evolution." Nature **392**: 917-920.

Lemaitre, C., L. Zaghloul, M.-F. Sagot, C. Gautier, et al. (2009). "Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation." BMC Genomics **10**(1): 335.

Osbourn, A. (2010). "Gene clusters for secondary metabolic pathways: an emerging theme in plant biology." Plant Physiol **154**(2): 531-5.

Pozzoli, U., G. Menozzi, M. Fumagalli, M. Cereda, et al. (2008). "Both selective and neutral processes drive GC content evolution in the human genome." BMC Evol Biol **8**: 99.

Romiguier, J., V. Ranwez, E. J. Douzery and N. Galtier (2010). "Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes." Genome Res **20**(8): 1001-9.

Saccone, S., C. Federico, I. Solovei, M. F. Croquette, et al. (1999). "Identification of the gene-richest bands in human prometaphase chromosomes." Chromosome Res **7**(5): 379-86.

Szego, E. M., K. A. Kekesi, Z. Szabo, T. Janaky, et al. (2009). "Estrogen regulates cytoskeletal flexibility, cellular metabolism and synaptic proteins: A proteomic study." Psychoneuroendocrinology **35**(6): 807-19.

Tatusov, R. L., M. Y. Galperin, D. A. Natale and E. V. Koonin (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Res **28**(1): 33-6.

Tatusov, R. L., D. A. Natale, N. D. Fedorova, I. V. Garkavtsev, et al. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1): 22-8.

Tatusov, R. L., D. A. Natale, N. D. Fedorova, J. D. Jackson, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, et al. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1): 22-8.

Uliano, E., A. Chaurasia, L. Bernà, C. Agnisola, et al. (2010). "Metabolic rate and genomic GC. What we can learn from teleost fish." Marine Genomics **3**(1): 29-34

Veron, A., C. Lemaitre, C. Gautier, V. Lacroix, et al. (2011). "Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny." BMC Genomics **12**(1): 303.

Versteeg, R., B. D. van Schaik, M. F. van Batenburg, M. Roos, et al. (2003). "The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes." Genome Res **13**(9): 1998-2004.

Vinogradov, A. E. (2001). "Bendable genes of warm-blooded vertebrates." Mol Biol Evol **18**(12): 2195-200.

Vinogradov, A. E. (2005). "Noncoding DNA, isochores and gene expression: nucleosome formation potential." Nucleic Acids Res **33**(2): 559-63.

Vinogradov, A. E. and O. V. Anatskaya (2006). "Genome size and metabolic intensity in tetrapods: a tale of two lines." Proc Biol Sci **273**(1582): 27-32.

4 Compositional Study of Tunicate Genomes

4.1 Introduction

The *Ciona* genome

Among the ascidians, the most representative and studied organisms are those belonging to the genus of *Ciona*. Two species namely *C. intestinalis* and *C. savignyi* belong to tunicates, which is now considered as the sister group of vertebrates. Because of their key evolutionary position, they have been considered as fundamental organisms in explaining the origin of chordates and hence the origin of vertebrates. Apart from the evolutionary point of view another plus point is their relatively small genome size (160 mega bases pairs) that has made them (Simmen, Leitgeb et al., 1998; Adams, Celniker et al., 2000) an ideal experimental model to investigate mechanism behind molecular evolution.

The draft genome of *C. intestinalis* was published in 2002 (Dehal, Satou et al., 2002) and that of *C. savignyi* was published in 2005 (Vinson, Jaffe et al., 2005). The availability of two complete sequenced ascidian genomes *C. intestinalis* and *C. savignyi* has provided a deeper insight into the evolutionary origins of the vertebrates.

A comparative genome analysis pattern between *C. intestinalis* and *C. savignyi*, by intensively investigating the base composition pattern, has not only revealed a comprehensive picture of compositional pattern of two tunicate genomes but also offered a leap forward to argue on which among the proposed guiding force(s) could also hold in case of tunicates. Specifically keeping the focus towards stochastic forces, in absence of evidences for differences in habitat temperature between two *Ciona* species, the analysis was performed to test the role played by metabolic rate in shaping the tunicates' genome composition.

4.2 Results

4.2.1 GC content

The genomes of both tunicate *C. intestinalis* and *C. savignyi* have been reported to be AT rich (de Luca di Roseto, Bucciarelli et al., 2002; Dehal, Satou et al., 2002), but an exhaustive investigation on the base composition in different genomic regions was never carried out. Hence, the average GC level of the whole genome (GC_g), introns (GC_i), 5'- and 3'-flanking regions, i.e. 2kb up- and down-stream CDS (GC_f or, more precisely, $5'GC_f$ and $3'GC_f$, respectively), as well as that of coding sequences (GCcds), and that of each codon position (GC1, GC2 and GC3) was computed (Table 4.1 and Figure 4.1). Within each genome, the average GC content of the regions so far analyzed, showed a similar order of ranking, with GCi having the lowest and GC1 the highest value (Table 4.**1**). The most significant difference was found at the third codon positions. Indeed, *C. intestinalis* and C. savignyi showed a significantly different GC3 content, being 7% higher in *C. savignyi*.

Table 4.1: Different GC content in coding and non-coding regions of *C. intestinalis* and *C. savignyi*.

Species	All Genes	GCg	GCi	5' GC*	3'GC*	GC_{CDS}	GC1	GC2	GC3
C. intestinalis	19697	37.18%	34.44%	36.80%	36.59%	42.60%	48.87%	39.24%	39.81%
C. savignyi	**20143**	**38.67%**	**35.67%**	**37.17%**	**37.80%**	**45.42%**	**49.78%**	**39.65%**	**46.81%**
Orthologous Genes									
C. intestinalis	7747	--	34.23%	36.93%	36.80%	42.71%	49.31%	39.08%	39.70%
C. savignyi	**7747**	**--**	**35.48%**	**37.44%**	**37.97%**	**45.30%**	**50.03%**	**39.24%**	**46.60%**

(*) 2000 bp.

In all pair-wise comparisons the average GC content was higher in *C. savignyi* than in *C. intestinalis*, and the differences were statistically significant, (p-value < 10^{-10}, at least). The lowest delta was at GC2 (0.4%), whereas the highest was at GC3 (7.0%). A closer inspection of GC3 values clearly showed bell-shaped normal distributions, skewed towards high GC3 values (Figure 4.1). Interestingly, in both genomes, the lowest value of the GC3 range was around 20%, whereas the maximum was around 60% in *C.*

intestinalis and around 72% in *C. savignyi*, a picture that mimics the transition mode of evolution observed comparing cold- and warm-blooded vertebrates (D'Onofrio, Jabbari et al., 1999; Bernardi, 2004, for review). Restricting the analysis to a set of orthologous sequences, similar results were found (Table 4.1). Once more, the coding positions showed higher GC values in *C. savignyi*. The highest difference between the two *Ciona* was found at the third codon position (GC3).

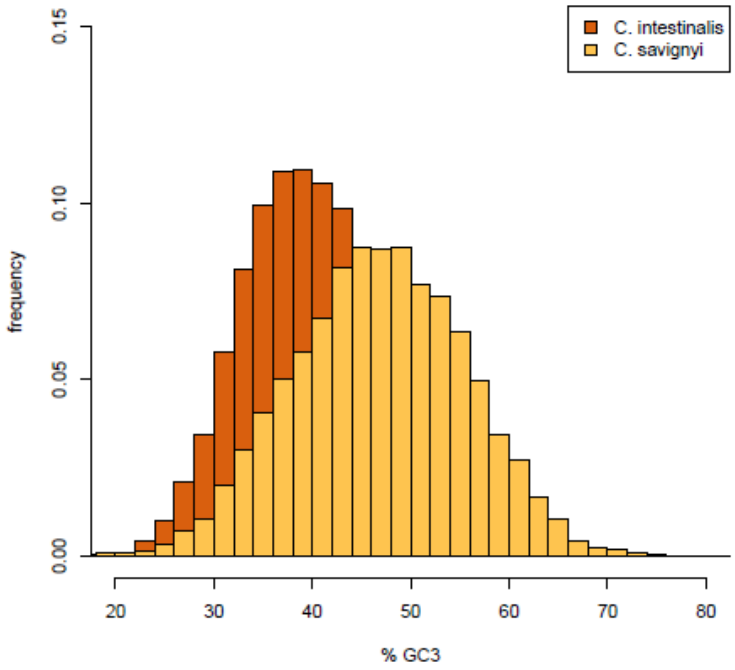


Figure 4.1: GC3 distribution of *C. intestinalis* and *C. savignyi*.

4.2.2 CpG doublet

The Frequency of CpG doublet as well as the doublet derivative TpG and CpA (see Chapter 2, section 2.3) were also calculated in both *Ciona* species with the rationale to check the whether the variation observed in GC content values could possibly be due the methylation/demination process of CpG doublet. The analysis was performed on both in the non coding (Table 4.2, upper panel) and also in coding regions (Table 4.2, lower panel). In the non-coding regions the frequencies of all the doublets analyzed between *C. intestinalis* and *C. savignyi*, showed either minor or no differences. More precisely: i) the

frequencies of the CpG doublets were 0.026 and 0.028 in *C. intestinalis* and *C. savignyi*, respectively; ii) those of TpG were 0.067 and 0.066, respectively; and iii) those of CpA were 0.095 and 0.090, respectively. Even in case of coding regions: i) the frequencies of TpG and CpA were practically identical, for former 0.076 and 0.074 and for later 0.075 and 0.076, respectively in *C. intestinalis* and *C. savignyi*, respectively.

Table 4.2: Di-nucleotides' frequencies of *C. intestinalis* and *C. savignyi* in non-coding and coding regions.

		CpG	TpG	CpA
Non-coding region	*C. intestinalis*	0.026	0.067	0.095
	C. savignyi	0.028	0.066	0.090
Coding region	*C. intestinalis*	0.036	0.076	0.075
	C. savignyi	0.044	0.074	0.076

4.2.3 Metabolic rate

Oxygen consumption of individual animals was calculated in a closed system, using a special setup that consists of a microelectrode connected to oxygen monitoring system (details in, Appendix 1, section 6.4.2). The linear regression of the total oxygen ($\mu g\text{-}O_2$) vs. time relationship gave the amount of oxygen consumed by the animal per unit time (usually 1 h). Dividing this value by the weight of the animal (wet or dry weight) the specific oxygen consumption ($mg\text{-}O_2\ h^{-1}\ kg^{-1}$) was calculated. The mean values of metabolic rate in *C. intestinalis* and in *C. savignyi* were found to be 236.9 and 323.13, respectively (Table 4.3).

Table 4.3: Table showing the different statistical measures of metabolic rate data in two tunicates.

	Count	Mean	Variance	Std. deviation	Std. error
C. intestinalis	15	236.9	4109.91	64.109	16.553
C. savignyi	14	323.13	33843.14	183.96	49.167

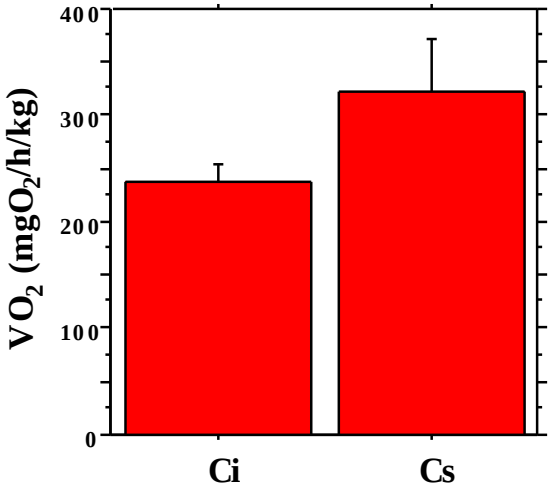


Figure 4.2: Bar-plot showing the average values of metabolic rate in C. intestinalis (Ci) and C. savignyi (Cs). Error bars represent the standard error.

Plotting the two means showed the *C. savignyi* showed higher values of metabolic rate than the *C. intestinalis* (Figure 4.2). The statistical significance of the differences observed was assessed by the t-Student's test, *p*-value = 0.0988, which is quite at the limit of significance (Table 4.4).

Table 4.4: Table showing the statistical test values of t-Student's test.

	Mean difference	DF	*t*-value	*p*-value
C. intestinalis, C. savignyi	-86.230	27	-1709	0.0988

As reported in Table 4.3, the difference between the variances in the two dataset was quite high, which could affect the *t*-test values. Applying the non-parametric Kolmogorov-Smirnov test the two distributions haven't showed significant differences. However, as per the expectation from correlation between GC% and metabolic rate (Uliano, Chaurasia et al., 2010; Chaurasia, Uliano et al., 2011), the preliminary analysis of two tunicate genomes also showed the same relationship, *C. savignyi* which is GC richer showed higher values of metabolic rate than the GC poor C. *intestinalis* (Figure 4.2, Table 4.1). More data on metabolic rate measurements in two tunicates could improve picture and provide a further confirmation to our preliminary results.

4.3 Discussion

Intra-genome comparison of genomic regions of *C. intestinalis* and *C. savignyi* clearly showed that GCcds was higher than GCg, GCi, 5'GC and 3'GC. In short, in both genomes the GC content was higher in coding than in non-coding regions, a feature first observed in the human genome (Aissani, D'Onofrio et al., 1991), afterwards observed in other eukaryotic genomes. The inter-genome comparison of the GC levels showed that *C. savignyi* was GC-richer, and this trend was found in all regions, coding and non-coding, so far analyzed. Interestingly, the highest increment of the GC level in *C. savignyi* was observed at the third codon positions (GC3) (about 7%, against an average of 2%). This observation has also been confirmed by the analysis of orthologous genes between two species.

Regarding CpG and its corresponding derivative doublets TpG and CpA in the non-coding region, the frequencies of the doublets showed either minor or no differences between the two species (Table 4.2). The frequencies of the CpG doublets in *C. intestinalis* and *C. savignyi* were 0.026 and 0.028, respectively. Even in coding regions the frequencies of doublets were practically identical in the two organisms (the frequencies of TpG were 0.076 and 0.074, respectively; and that of CpA were 0.075 and 0.076, respectively). These results showed that the differences observed at the GC level in two *Ciona* genome (*C. savignyi* was GC-richer than *C. intestinalis*) were not affected by the deamination of 5mC.

"*Which of the hypothesis could better fit with the data obtained from the compositional analysis of the ascidian genomes?*"

The selective pressure acting at synonymous codon positions and at intergenic/intervening sequences was generally accepted to be very similar (Graur and Li, 2000). Hence, the mutational bias could not completely explain a higher increment of the GC3 levels than those observed in non-coding regions.

Regarding BGC hypothesis, which is grounded on a significant correlation between GC content and recombination process (Eyre-Walker, 1993), the recombination rate were reported to be 25-49 kb/cM in *C. intestinalis* (Kano, Satoh et al., 2006) and 200

kb/cM in *C. savignyi* (Hill, Broman et al., 2008). From this it is expected that *C. intestinalis* should have to be GC-richer than *C. savignyi*. But the base compositional analysis showed contrary values i.e. *C. savignyi* is GC richer than *C. intestinalis*.

Regarding the thermal stability, biogeographical distribution of both *Ciona* species showed that *C. intestinalis* species is although a native of northern Europe (Monniot, Monniot et al., 1994) but is widely distributed from the Baltic Sea to the Mediterranean, along Atlantic coasts of North America, Atlantic and Pacific coasts of South America and also recorded at Hawaii, South Africa, Australia, New Zealand, and Japan and hence considered as cosmopolitan species. While *C. savignyi* is apparently restricted to North Pacific area, mainly in Japan and west cost of United States. But no evidences of different environmental temperature among two *Ciona* species have been reported.

Measurement of MR, i.e. O_2 consumption, in both the species, has provided initial proofs for the correlation between GC% and MR (Figure 4.2).

4.4 Conclusions

At present, among the proposed hypotheses discussed above namely, mutational bias; BGC and thermodynamic stability hypothesis, were unable to background the variations of base compositions in two tunicates. While only metabolic rate hypothesis have provided an understandable theoretical framework for the base composition differences observed between the two *Ciona* species. More measurements on the metabolic rate could further be of great support to the results.

4.5 References

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-95.

Aissani, B., G. D'Onofrio, D. Mouchiroud, K. Gardiner, et al. (1991). "The compositional properties of human genes." J Mol Evol **32**(6): 493-503.

Bernardi, G. (2004). Sructural and Evolutionary Genomics. Natural Selection in Genome Evolution. Amsterdam, Elsevier.

Chaurasia, A., E. Uliano, B. L, C. Agnisola, et al. (2011). "Does habitat affects the genomic GC content? A lesson from teleostean fish – a mini review." Book chapter : Fish Ecology.

D'Onofrio, G., K. Jabbari, H. Musto, F. Alvarez-Valin, et al. (1999). "Evolutionary genomics of vertebrates and its implications." Ann N Y Acad Sci **870**: 81-94.

de Luca di Roseto, G., G. Bucciarelli and G. Bernardi (2002). "An analysis of the genome of Ciona intestinalis." Gene **295**(2): 311-6.

Dehal, P., Y. Satou, R. K. Campbell, J. Chapman, et al. (2002). "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." Science **298**(5601): 2157-67.

Eyre-Walker, A. (1993). "Recombination and mammalian genome evolution." Proc. R. Soc. Lond. **B 252**: 237-243.

Graur, D. and W.-H. Li (2000). Fundamentals of molecular evolution. Mass, Sinauer, Sunderland.

Hill, M. M., K. W. Broman, E. Stupka, W. C. Smith, et al. (2008). "The C. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution." Genome Res.

Kano, S., N. Satoh and P. Sordino (2006). "Primary genetic linkage maps of the ascidian, Ciona intestinalis." Zoolog Sci **23**(1): 31-9.

Monniot, C., F. Monniot and oise (1994). "Additions to the Inventory of Eastern Tropical Atlantic Ascidians; Arrival of Cosmopolitan Species." Bulletin of Marine Science **54**(1): 71-93.

Simmen, M. W., S. Leitgeb, V. H. Clark, S. J. Jones, et al. (1998). "Gene number in an invertebrate chordate, Ciona intestinalis." Proc Natl Acad Sci U S A **95**(8): 4437-40.

Uliano, E., A. Chaurasia, L. Bernà, C. Agnisola, et al. (2010). "Metabolic rate and genomic GC. What we can learn from teleost fish." Marine Genomics **3**(1): 29-34

Vinson, J. P., D. B. Jaffe, K. O'Neill, E. K. Karlsson, et al. (2005). "Assembly of polymorphic genomes: algorithms and application to Ciona savignyi." Genome Res **15**(8): 1127-35.

5 Conclusions

The variation of base composition among genomes is an open question and still under debate within the neutralist/selectionist frame. Several factors have been proposed as main driving force, namely the mutational bias (Sueoka, 1962), the bias gene conversion (BGC) (Eyre-Walker, 1993; Galtier, Piganeau et al., 2001; Duret and Galtier, 2009); the thermal stability (Bernardi, 2004, for review), and the metabolic rate (Vinogradov, 2001, 2005). With the aim to highlight which of the aforementioned forces mainly influence the base compositional evolution, different approaches and strategies were applied to analyze the compositional pattern of genomes belonging to: (i) teleosts, (ii) mammals and (iii) tunicates. The thesis mainly tested the metabolic rate hypothesis and the results were discussed in the light on the pros and cons of all current evolutionary hypotheses.

First and foremost focus was on teleosts, a diverse group of fish covering wide range of habitat. The rationale of the choice started from the consideration that aquatic organism, different from terrestrial one, live in an environment where the available oxygen is a limiting factor, dictated by the Henry's law. Hence, the aim was to disentangle the oxygen consumption from the environmental temperature, and to check the role played by different factors in shaping genome structure and organization. On the final data set of 206 teleosts fish, data about mass specific routine metabolic rate temperature-corrected using the Boltzmann's factor (MR) and base composition of genomes (GC%) were examined and related to their major habitat: polar, temperate, sub-tropical, tropical and deep-water. Fish of the polar habitat showed the highest average MR, and that of both temperate and subtropical fish was significantly higher than that of tropical one. Regarding average GC%, polar and temperate fish both showed significantly higher values than those of sub-tropical and tropical fish. Crossing the two data sets and plotting the values of MR and GC%, a positive and significant correlation was found between the two variables (Uliano, Chaurasia et al., 2010).

In order to establish a possible effect of the environmental temperature on the GC content through the methylation/deamination process of cytosine, the frequencies of the CpG doublets and those of the derivative ones, TpG and CpA, were analyzed in intronic sequences of five completely sequenced fish genomes (namely: *D. rerio, O. latipes, T. rubripes, G. aculeatus* and *T. nigroviridis*). The results clearly showed that, in spite of the different temperatures and methylation levels among habitats, different CpG levels among fish were not paralleled by different levels of TpG and CpA. Hence, the low average GC content observed in tropical fish respect to polar and temperate ones could not be ascribed to the temperature dependence of the 5mC deamination process.

The intronic sequences of fish were further analyzed in the frame of the open question about factors affecting the intron length variation within and among genomes. Pairwise genome comparisons, using orthologous intron sequences, showed that the GC of introns (GCi%) and their length (bpi) were linked by an inverse relationship. The issue needs more data about the O_2 consumption of the species analyzed in order to provide a solid support to the metabolic rate hypothesis. Same consideration applies to the comparison between C. *intestinalis* and *C. savignyi*, where a positive correlation holds between the GC% and the metabolic rate.

Secondly, human genes were assigned to three large functional categories according to the KOG database: information storage and processing, cellular processes and signaling and metabolism. The GC3 level was significantly increasing from the former to the latter. This specific compositional pattern was found, as footprint, in all mammalian genomes, but not in frog and lizard ones. Comparative analysis of human versus both frog and lizard showed that genes involved in the metabolic processes underwent to the highest GC3 increment. Analyzing the KOG functional classes of genes, again a precise intra-genomic pattern was found in all mammals. Not only genes of metabolic pathways, but also genes involved in chromatin structure and dynamics, transcription, signal transduction mechanisms and cytoskeleton, showed GC3 levels higher than that of the whole genome. In the case of human, the genes of the

aforementioned functional classes showed a high probability to cluster in the GC-richest chromosomal bands.

All the different evolutionary hypotheses proposed till now surely contribute, with different weights, to the compositional variability observed among and within organisms (Pozzoli, Menozzi et al., 2008). Few, however, seems to fit with the very wide range from prokaryote to eukaryotes.

The analysis of single nucleotide polymorphism (SNPs) first in the human genome (Lander, Linton et al., 2001; Alvarez-Valin, Lamolle et al., 2002) and more recently among bacterial genomes (Hershberg and Petrov, 2010), showed the existence of an AT mutational pressure, therefore failing to give support to the Sueoka's hypothesis based on the mutational bias to explain the GC variability among genomes.

Regarding bias gene conversion hypothesis (BGC) a lack of correlation between recombination rates [cM/Mb] and GC% was found among bacteria (Hildebrand, Meyer et al., 2010) and chordates (Wataru Kai, 2011). In the frame of present results, BGC could hardly explains: (i) the opposite relationship between recombination rate and GC content found in *Ciona* species, since *C. savignyi* showed a lower recombination rate and a higher GC content than *C. intestinalis;* (ii) the specific compositional pattern of gene functional classes conserved in all mammalian genomes, in spite of the fact that the recombination hot-spot sites were reported to be not phylogenetically conserved, even in closely related species like chimpanzee and human (Ptak, Hinds et al., 2005).

Regarding adaptive factors, considering that: (i) teleosts living in polar and temperate habitat showed high genomic GC content than those living in subtropical and tropical habitat (Uliano, Chaurasia et al., 2010); (ii) the lack of correlation between CpG and the sum of its derivative doublet (TpG and CpA) excluded the temperature dependence of the deamination process as factor affecting the GC variability among fish (Chaurasia, Uliano et al., 2011); and (iii) although living in the same habitat compositional differences exists between *C. intestinalis* and *C. savignyi*, it is hard to see how the aforementioned results could fit into the thermodynamic stability hypothesis (Bernardi, 2004, for review).

Moreover, the same hypothesis states that the GC increment should stabilize not only DNA and RNA, but also proteins. However, an analysis of the lactate dehydrogenases (LDH) failed to support the hypothesis. Indeed, the heat denaturation profile of LDH was not correlated with habitat temperature (Fields and Somero, 1997).

It is worth to stress that the results of the present thesis do not deny, but rephrase the role of the environmental/body temperature on the genomic GC content. Environmental T°, indeed, if on one side doesn't affect the DNA stability through increments of hydrogen bonds (GC increment), on the other side plays an important eco-physiological role in the different habitat, at least affecting the O_2 availability in the aquatic environment. Nevertheless, the thermodynamic hypothesis could not be drawn out from the analysis of mammalian genomes. In this case, indeed, the data do not allow to exclude completely a direct link between increments of body temperature and increments of GC. However, at same time it is not possible to disregard the role played by the metabolism, since the genes that underwent the highest GC increment from "cold- to warm-blooded vertebrates" (Bernardi, Olofsson et al., 1985) were those involved in metabolic processes.

Considering that although the metabolic rate hypothesis is in the frame of the adaptive hypotheses, most probably there is no need to evoke the positive selection to account for the effect of MR on GC. Indeed, the shift of the negative selection threshold from habitat to habitat for the "best-fit GC content" was proposed to account for the genome compositional shift observed comparing teleostean fish living in different habitat (Uliano, Chaurasia et al., 2010). Natural (negative) selection has been also proposed to explain the great compositional heterogeneity and the appearance of GC-rich isochores of the human genome (Bernardi, 2007).

In conclusion, to give an answer to the starting question: *which hypothesis drives the base composition evolution among organisms*, certainly we can say that the metabolic rate doubtless plays not a minor role in the evolution of vertebrate and non-vertebrate genomes. Bearing in mind the works published on bacteria (Naya, Romero et al., 2002; Rocha and Danchin, 2002; Baudouin-Cornu, Schuerer et al., 2004; Musto, Naya et al.,

2006), most probably this hypothesis extends to all living organisms.

5.1 References

Alvarez-Valin, F., G. Lamolle and G. Bernardi (2002). "Isochores, GC3 and mutation biases in the human genome." Gene **300**(1-2): 161-8.

Baudouin-Cornu, P., K. Schuerer, P. Marliere and D. Thomas (2004). "Intimate evolution of proteins. Proteome atomic content correlates with genome base composition." J Biol Chem **279**(7): 5421-8.

Bernardi, G. (2004). Sructural and Evolutionary Genomics. Natural Selection in Genome Evolution. Amsterdam, Elsevier.

Bernardi, G. (2007). "The neoselectionist theory of genome evolution." Proc Natl Acad Sci U S A **104**(20): 8385-90.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, et al. (1985). "The mosaic genome of warm-blooded vertebrates." Science **228**(4702): 953-8.

Chaurasia, A., E. Uliano, B. L, C. Agnisola, et al. (2011). "Does habitat affects the genomic GC content? A lesson from teleostean fish – a mini review." Book chapter : Fish Ecology.

Duret, L. and N. Galtier (2009). "Biased gene conversion and the evolution of mammalian genomic landscapes." Annu Rev Genomics Hum Genet **10**: 285-311.

Eyre-Walker, A. (1993). "Recombination and mammalian genome evolution." Proc. R. Soc. Lond. **B 252**: 237-243.

Fields, P. and G. Somero (1997). "Amino acid sequence differences cannot fully explain interspecific variation in thermal sensitivities of gobiid fish A4-lactate dehydrogenases (A4-LDHs)." J Exp Biol **200**(Pt 13): 1839-50.

Galtier, N., G. Piganeau, D. Mouchiroud and L. Duret (2001). "GC-content evolution in mammalian genomes: the biased gene conversion hypothesis." Genetics **159**(2): 907-911.

Hershberg, R. and D. A. Petrov (2010). "Evidence that mutation is universally biased towards AT in bacteria." PLoS Genet **6**(9).

Hildebrand, F., A. Meyer and A. Eyre-Walker (2010). "Evidence of selection upon genomic GC-content in bacteria." PLoS Genet **6**(9).

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Musto, H., H. Naya, A. Zavala, H. Romero, et al. (2006). "Genomic GC level, optimal growth temperature, and genome size in prokaryotes." Biochem Biophys Res Commun **347**(1): 1-3.

Naya, H., H. Romero, A. Zavala, B. Alvarez, et al. (2002). "Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes." J Mol Evol **55**(3): 260-4.

Pozzoli, U., G. Menozzi, M. Fumagalli, M. Cereda, et al. (2008). "Both selective and neutral processes drive GC content evolution in the human genome." BMC Evol Biol **8**: 99.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, et al. (2005). "Fine-scale recombination patterns differ between chimpanzees and humans." Nat Genet **37**(4): 429-34.

Rocha, E. P. and A. Danchin (2002). "Base composition bias might result from competition for metabolic resources." Trends Genet **18**(6): 291-4.

Sueoka, N. (1962). "On the genetic basis of variatrion and heterogeneity of DNA base composition." Proc. Natl. Acad. Sci. USA **48**: 582-592.

Uliano, E., A. Chaurasia, L. Bernà, C. Agnisola, et al. (2010). "Metabolic rate and genomic GC. What we can learn from teleost fish." Marine Genomics **3**(1): 29-34

Vinogradov, A. E. (2001). "Bendable genes of warm-blooded vertebrates." Mol Biol Evol **18**(12): 2195-200.

Vinogradov, A. E. (2005). "Noncoding DNA, isochores and gene expression: nucleosome formation potential." Nucleic Acids Res **33**(2): 559-63.

6 Appendix – 1

Materials and Methods

6.1 Sequences

6.1.1 Coding sequences

Coding sequences of the genome assembly were retrieved from the ENSEMBL (http://ftp.ensembl.org) for all five fish namely:

1) ***D. rerio*** (Assembly: Zv7, Apr 2007, Ensembl Release: 48.7b);
2) ***O. latipes*** (Assembly: HdrR, Oct 2005, Ensembl Release 48.1d);
3) ***G. aculeatus*** (Assembly: BROAD S1, Feb 2006, Ensembl Release 48.1e);
4) ***T. nigroviridis*** (Assembly: TETRAODON 7, Apr 2003, Ensembl Release 48.1j);
5) ***T. rubripes*** (Assembly: FUGU 4.0, Jun 2005, Ensembl Release 48.4h).

Sequence data for ***C. intestinalis*** was retrieved from following:

1) JGI: Ensembl Release: 1.0, 2002 (http://genome.jgi-psf.org);
2) ENSEMBL: Ensembl Release: 48.2h (http://ftp.ensembl.org).

From the latter database coding sequences of the following species (in alphabetical order) were also retrieved: ***A. carolinensis***, ***B. taurus***, ***C. savignyi***, ***D. novemcinctus***, ***E. caballus***, ***G. gorilla***, ***L. africana***, ***M. domestica***, ***M. musculus***, ***O. anatinus***, ***O. cuniculus***, ***P. vampyrus***, ***P. pygmaeus***, ***R. norvegicus***, ***S. tridecemilineatus*** and ***T. truncates*** and data for ***X. laevis*** was retrieved from www.xenbase.org.

6.1.2 Non-Coding sequences and repetitive elements

Intronic sequences were retrieved from UCSC Genome browser (http://genome.ucsc.edu/), for all five fish namely:

1) *D. rerio* (Assembly: Apr 2007, Zv7/danRer5);
2) *O. latipes* (Assembly: Oct 2005, NIG/UT, MEDAKA 1/ oryLat2);
3) *G. aculeatus* (Assembly: Feb 2006, BROAD/gas Acu1);

4) *T. nigroviridis* (Assembly: Feb 2004, Genoscope 7.0/tetNig1);
5) *T. rubripes* (Assembly: Oct 2004 (JGI 4.2/ fr2).

Flanking regions (2000 bp flanking the transcript at 5'and 3' side) were retrieved respectively form Ensemble (www.ensembl.org) using Biomart tools.

In each genome the number of full length genes (i.e. CDS + introns) was: *D. rerio* 17085, *O. latipes* 13247, *G. aculeatus* 16101, *T. nigroviridis* 10898, *T. rubripes* 19123. Sequences containing characters indicating ambiguity in identification of certain bases, i.e. **N,** were discarded.

RepeatMasker (Version 3.1.9, http://repeatmasker.org) was used to mask the interspersed repeats and low complexity DNA sequences (Smit, R et al., 1996-2010). The overall number of sequences was basically unaffected by the masking process. The final data set analyzed consist of *D. rerio* (13521), *O. latipes* (5779), *G. aculeatus* (13854), *T. nigroviridis* (8905), and *T.rubripes* (15753).

6.1.3 Human KOG sequences and classification

Functional classification of human proteins was retrieved from KOG database (http://www.ncbi.nlm.nih.gov/COG/) (Tatusov, Natale et al., 2001; Tatusov, Natale et al., 2003). The corresponding coding sequences (CDS) were retrieved from NCBI (http://www.ncbi.nlm.nih.gov) using a batch entrez function. All the functional classes (denoted by capital letters in square brackets), were grouped into three large functional categories, namely: (i) information storage and processing; (ii) cellular processes and signaling; and (iii) metabolism. Proteins classified in more than one class were removed from further analysis. Genes with only predicted function [R] or unknown [S], representing about 19%, were removed from further analyses, as well as the three functional classes, namely [M], [N] and [Y] because were represented by less than a hundred sequences.

Human CDS from each KOG class were assigned to the different compositional band types (L1+, L1-, H3-, H3+ bands) previously identified in the human chromosomes (Federico, Andreozzi et al., 2000; Costantini, Clay et al., 2007).

6.2 Orthologs

Orthologous gene pairs were identified using a Perl script, which performs reciprocal Blastp (Altschul, Madden et al., 1997) and selects the Best Reciprocal Hits (BRH) i.e. when two genes, each in a different genome, find each other as the best hit in the other genome. This procedure of finding orthologs between two species (pairs orthologs), was used to perform following analyses:

1) For the comparative analysis in teleosts, orthologous genes for each possible pair between *D. rerio, O. latipes, G. aculeatus, T. nigroviridis* and *T.rubripes* were identified.
2) The orthology obtained from the coding sequences was then extended to non-coding. All the pair wise comparative analyses on non-coding sequences were performed between the orthologous gene pairs obtained from all the possible ten combinations among the five fish namely, *D. rerio* - *O. latipes* (2970); *D. rerio* - *G. aculeatus* (5801); *D. rerio* - *T. nigroviridis* (4540); *D. rerio* - *T. rubripes* (5430); *O. latipes* - *G. aculeatus* (3482); *O. latipes* - *T. nigroviridis* (2844); *O. latipes* - *T. rubripes* (3074); *G. aculeatus* - *T. nigroviridis* (5462); *G. aculeatus* - *T. rubripes* (6450); *T. rubripes* - *T. nigroviridis* (5430).
3) Comparisons of orthologous sequences between KOG-classified human genes and thirteen mammals (belonging to primates, rodents, laurasiatheria, afrotheria, xenarthra, marsupials and monotremes), one reptile and one amphibian were performed. Hence based on orthology with human genes, the same KOG classification was extended to the corresponding genes in all species analyzed.

4) For the comparative analysis in tunicates, 7747 orthologous pairs were identified between *C. intestinalis* and *C. savignyi*.

6.3 Base composition

6.3.1 Composition in Teleosts and tunicates

The genomic GC content of teleostean fish were retrieved from an extensive study (Bucciarelli, Bernardi et al., 2002; Varriale and Bernardi, 2006), reporting a comprehensive view of the compositional characteristics (average GC content) of more than 200 fish genomes. The data was then classified into five habitat classes representing 9 polar, 22 temperate, 48 subtropical and 70 tropical teleostean fish. Unfortunately, Genomic GC contents for fish living in deep-water habitat were not available.

Regarding base compositional analyses on five completely sequenced genomes, namely: *D. rerio, O. latipes*, *G. aculeatus*, *T. nigroviridis* and *T.rubripes* as well as for two tunicate species *C. intestinalis* and *C. savignyi*, CodonW (1.4.4) was used to calculate the molar ratio of guanine plus cytosine (GC) of the entire genome (GCg), as well as that of both non-coding and coding regions. More precisely, the GC content of introns (GCi), flaking region (5'GC and 3'GC), coding sequences (GCcds) and that of each coding position (GC1, GC2 and GC3) were calculated. The frequencies of the CpG di-nucleotide and those of the derivate doublets after the deamination process of the 5-methyl- cytosine (5mC), namely, TpG and CpA were also calculated.

Basic sequence information were retrieved by using Infoseq, an application of EMBOSS package (EMBOSS, Release 5.0; http://emboss.sourceforge.net/). Sequences with length less than hundred base pairs (bp) were excluded from further analysis.

6.3.2 GC3 of KOG classes

For thirteen mammalian, one reptile and one amphibian species (Chapter 3), the average GC3 level of each functional class was compared with that of the genome (i.e. the average of the GC3 level calculated using all the available sequences of the species), and statistical significance was assessed by the t-Student's test, with Bonferroni's correction ($\alpha = 0.05$) for multiple-comparisons. The data were showed as **Butterfly plot.**

A two-tale Mann-Whitney test was performed in order to test the statistical significance of the differences in GC3 content between the three main categories of genes: Blue, Black and Red.

6.4 Metabolic rate

6.4.1 Teleosts

Regarding teleostean fish, data about the metabolism as well as the information of taxonomic classification, geographical distribution, were downloaded principally from www.fishbase.org and from the available literature. Metabolic rate measurements obtained under any kind of stress was discarded (e.g. in hypoxia, feeding or starvation). In order to compare metabolic rates among organisms living in different environments, data about mass specific oxygen consumption (mg kg^{-1} h^{-1}) are commonly normalized to a standard temperature using the Q_{10} value. (Hegarty, 1973) first pointed out the misuse of this coefficient. As temperature dependence of metabolism is changing with actual temperature, the Boltzmann's factor appears to be a better procedure to correct for temperature effect on metabolism compared to correction via Q_{10} (Clarke and Johnston, 1999). More recently Gillooly and colleagues (Gillooly, Brown et al., 2001), and later on (Hodkinson, 2003), raised doubts about the use of Q_{10} value analyzing a large number of species covering a broad range of living temperatures. First, evaluating the temperature dependence of metabolism using Q_{10} values, an error up to 15% may be introduced (Gillooly, Brown et al., 2001). Second, Q_{10} is assumed to be temperature independent, while the temperature dependence of biological processes usually is not purely

exponential (Hodkinson, 2003). Moreover, the Q_{10} value, usually assumed to range between 2 and 3, can reach values significantly out of this range and has been shown to be species specific, therefore should be accurately determined (Karamushko, 2001). In order to avoid the above criticisms, the mass specific metabolic rate values obtained for each fish (expressed as milligrams of oxygen consumed per kilogram of wet weight per hour, $mgkg^{-1}h^{-1}$) was temperature-corrected utilizing the Boltzmann's factor (Gillooly, Brown et al., 2001) according to the following equation: $MR = MR_0 e^{-E/KT}$, where MR is the temperature-corrected mass specific metabolism, MR_0 is the metabolism at the temperature T expressed in K , E is the energy activation of metabolic processes ~ 0.65 eV, and k is the Boltzmann's constant equal to $8.62x10^{-5}$ eV K^{-1}.

Taking into consideration the experimental conditions described in fishbase database, the average MR values were calculated for: standard (S, in absence of physical activity), routine (R, in absence of constant swimming, but only spontaneous activity), and active (A, under constant swimming activity) conditions. As expected, values were increasing from standard to active conditions, i.e. $S < R < A$. Data from some species did not follow this order, consequently were removed from the dataset. Those species that present $R > A$ values: *Coregonus sardinella*, *Dorosoma cepedianum*, *Oreochromis mossambicus* and *Pleuronectes platessa*; those with $S > R$: *Gadus morhua*, *Ictalurus punctatus*, *Labeo capensis*, *Lipophrys pholis*, *Macrognathus aculeatus*, *Micropterus salmoides*, *Pseudopleuronectes americanus*, *Rhinogobiops nicholsii* and *Typhlogobius californiensis*. The final dataset consisted in 206 species that were classified in five habitats: polar, temperate, subtropical, tropical and deep-water.

The final dataset consisted in 206 species that were classified in five habitats: polar, temperate, subtropical, tropical and deep-water.

6.4.2 Tunicates

C. intestinalis is cosmopolitan and is widely distributed whereas *C. savignyi* is restricted to North Pacific area, mainly in Japan and west cost of United States. For comparative study of metabolic rates, the two *Ciona* species' samples should have to be under the same environmental and experimental conditions. Consequently, the metabolic rate calculations were performed on the *C. intestinalis* and *C. savignyi* collected from Santa Barbara Yacht harbor, Santa Barbara, California, US, in collaboration with Ascidian Stock Centre, MCD Biology, University of California Santa Barbara, California, United States. The final analysis was performed on a dataset of 14 samples from *C. savignyi* and 15 from *C. intestinalis*.

Regarding tunicates, oxygen consumption of individual animal was performed in a closed system. An oxygen microelectrode (YSI 5357 Micro Probe, USA) was set through the chamber cover to continuously record the sea water oxygen content. The microelectrode was connected to an Oxygen Monitor System (YSI 5300 A), whose output signal was acquired via an analogical-digital interface (Pico Technology Ltd, UK) connected to a PC for automated data acquisition with a specific software (Picolog Pico Technology Ltd., UK). Water in the chamber was fully aerated and continuously circulated to maintain uniform oxygen concentration. Before introducing animal into the chamber, the oxygen sensor was calibrated at 100% oxygen saturated water. After calibration with 100% saturated sea water, the chamber was closed and the fall in oxygen content was recorded. No more than 15-20% of oxygen content fall was allowed. **Calculations:** Atmospheric pressure during determination was measured and used to calculate pO_2 according to the equation:

$$pO_2 = (AP - SVP) \text{ X } 0.2096$$

where, *AP* is the atmospheric pressure (torr); and *SVP* the saturated vapor pressure of water at the temperature of measurement; 0.2096 = O_2% in the air. From the pO_2 value, the oxygen concentration, in mg l^{-1}, was calculated as: [O2] = $pO_2 \times \alpha$, where α (in mg-O_2

l^{-1} $torr^{-1}$) is the oxygen solubility in seawater at the temperature of measurement. Knowing the chamber volume, the total amount of oxygen (in μg-O_2) in the chamber as a function of time during the oxygen consumption measurement is determined. The linear regression of this total oxygen vs. time relationship gives the amount of oxygen consumed by the animal per unit time (usually 1 h). Dividing this value by the animal weight (wet or dry weight) gives the specific oxygen consumption (usually expressed in terms of mg-O_2 h^{-1} kg^{-1}). During the measurements average ambient temperature was 18.9°C, water temperature was 14.7°C, salinity was 34.8%.

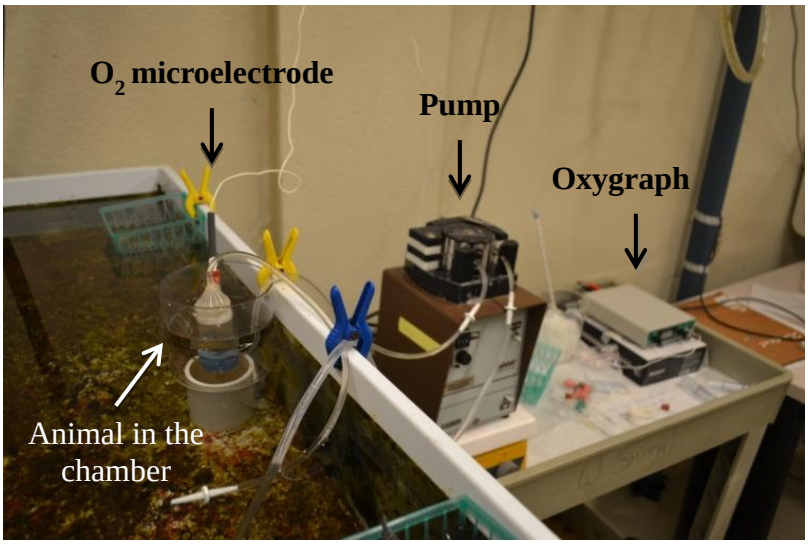


Figure 6.1: Picture showing the setup used during calculation of metabolic rate in two *Ciona* species.

6.5 Statistical Methods

6.5.1 Detection of Outliers

An outlier is defined as an observation that "appears" to be inconsistent with other observations in the data set. Because of a potentially large variance, outliers could be the outcome of sampling errors or errors during recording data. The requirement behind detecting the presence of outlier is that potentially they have strong influence on the estimates of the parameters and hence can produce misleading conclusions in test of hypotheses. Outliers were identified according to the procedure described in http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Outlier.htm and based on the analysis of average and standard deviation (Barnett and Lewis, 1994) that involves following steps:

1. Compute the mean ($\bar{x}$) and standard deviation (S) of the whole sample.
2. Set limits for the mean $\bar{x}$:

$$\bar{x} - k.\ S,\ \bar{x} + k.\ S$$

3. A typical value for k is 2.5.
4. Remove all sample values outside the limits.
5. Now, iterating through the algorithm, the sample set may reduce after removing the outliers by applying step 3. In most cases, we need to iterate through this algorithm several times until all outliers are removed.

6.5.2 Mann–Whitney U test

It also called the **Mann–Whitney–Wilcoxon** or **Wilcoxon rank-sum test** is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. Two important assumptions of the test are (i) the two samples under consideration are random, and are independent of each other (ii) the observations are numeric or ordinal (arranged in ranks). The test involves the calculation of a statistic, usually called U, whose distribution under the null hypothesis is known.

The combined set of data is first arranged in ascending order with tied scores receiving a rank equal to the average position of those scores in the ordered sequence.
The Mann-Whitney test statistic is then calculated using $U = n_1 n_2 + (n_1 (n_1 + 1)/ 2) – T$, where n_1 and n_2 are the sizes of the first and second samples respectively and T denote the sum of ranks for the first sample. Statistical significance of pair wise comparisons was assessed by the Mann-Whitney U test.

6.5.3 de Finetti's diagram

In order to assess the compositional/spatial distribution of the average GC3 in the three categories c (c= Blue, Black and Red) and to compare such behavior across different organisms $g = 1,\ldots,G$, the whole GC3 range $[a_g, b_g]$ of each organism g was split in three equal size intervals, corresponding to the levels denoted as Low, Medium and High, respectively. Then for each organism g and each category c we defined the vector $\left(c_g^L, c_g^M, c_g^H\right)$ containing the normalized occurrence for the corresponding functional classes in the three levels. Clearly, we have $c_g^i \geq 0$ and $\sum_{i\in\{Low,Medium,high\}} c_g^i = 1$ for all organisms g. Since each vector $\left(c_g^L, c_g^M, c_g^H\right)$ can be represented as a point (whose color corresponds to the category) in a de Finetti's diagram, each organism can be coded in the diagram using three colored points drawn in correspondence of its $\left\{\left(c_g^L, c_g^M, c_g^H\right)\right\}_{c\in\{blue,black,red\}}$ values. The de Finetti's diagram is a well known representation used in population genetics to show the range of genotype frequencies for which Hardy-Weinberg equilibrium is satisfied. Here we use it for comparing the GC3 compositional/spatial distribution between the categories in different organisms. Hence, to understand its meaning in our context we recall the Viviani's theorem that assures that sum of the distances from an internal point to the sides of an equilateral triangle equals the length of the triangle's altitude (that in our context is set to 1). According to such

results each $\left(c_g^L, c_g^M, c_g^H\right)$ value can be represented as a point inside the triangle and the distances to the corresponding side is equal to c_g^i. In practice the closer one point is to a particular side, the lesser such category is present in that genome at the level showed in that side. Additionally, by dividing the area of the triangle with the three triangle's altitudes and considering the centroid we can define the three sectors (identified by the Low, Medium, High line) of the triangle having different $\left(c_g^L, c_g^M, c_g^H\right)$ relational ordering and hence different GC3 abundance. Categories belonging to a given sector show minimal presence of that GC3 level with respect to the other levels (Figure 6.2). Intuitively, we observe that, in absence of association between the GC3 distribution and the functional categories, each configuration of the vector $\left(c_g^L, c_g^M, c_g^H\right)$ is equally likely and, as a consequence, the different sectors are expected to be equally represented.

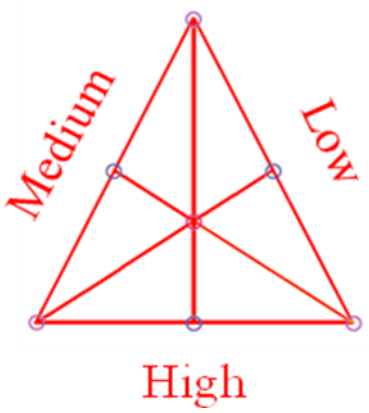


Figure 6.2: Representation of three sectors (Low, Medium, High line) of the triangle.

Discrepancy from such uniform distribution denotes a specific association. To measure such effect, first we observed that, due to the finite number of classes in each category only a finite number of configurations results attainable. Such configurations are invariants with respect to 2/3π rotations of the triangle. However, due to the fact that the range $[a_g, b_g]$ is organisms specific the observed $\left\{\left(c_g^L, c_g^M, c_g^H\right)\right\}_{c\in\{blue,black,red\}}$ are not independent since at least one class should be present in either the low and high levels,

hence the number of admissible configurations results less than the attainable. The $\left\{\left(c_g^L, c_g^M, c_g^H\right)\right\}_{c\in\{blue,black,red\}}$ for each organism g are shown in Figure. 6.1**,** where the value of each point represents the number of times the vector $\left(c_g^L, c_g^M, c_g^H\right)$ has been observed in the different organisms $g = 1,\ldots,G$. The de Finetti's diagram clearly showed that in the large majority mammalian genomes the Red category was confined to a restricted part of the space of the diagram (i.e., closed to the Low level line, denoting that in the large majority mammalian genomes the Red category: was rarely present in the lowest GC3 range. In order to test whether it was possible to obtain such configuration by chance, performed B class permutations among the categories and each time we counted k_i the occurrence of the Red class in the Low sector, then we estimated the p-value of the sector as $\frac{\sum_{i=1}^{B} I\left(k_i \geq k^*\right)}{B}$ where B denotes the number of permutation and k^* the observed occurrence of the Red class in the Low sector on our dataset.

7 Appendix - 2

Table 7.1: The fish genome projects registered in the NCBI database as of Sep'2011.

Organism Group	Genome Size (Mb)	Status	Release Date	List of Center/Consortium
Callorhinchus milii	0.647131	Assembly	1/5/2007	Institute of Molecular and Cell Biology, Singapore
Clarias fuscus	17	In Progress	NA	Institute of Bioinformatics, Anhui Normal University, China; Shanghai Sangon Biological Engineering Technology & Services Co. Ltd
Ctenopharyngodon idella	NA	In Progress	NA	Chinese Academy of Sciences and School of Biological Sciences, University of Hong Kong, Pokfulam Road, Hong Kong, SAR, China
Danio rerio	3112.62	Assembly	1/28/2005	Wellcome Trust Sanger Institute
Dicentrarchus labrax Adriatic clade, male 57	98.2281	Assembly	2/12/2010	European seabass sequencing consortium Max Planck Institute Max-Planck-Institute for Molecular Genetics, Ihnestr 63/73, D-14197 Berlin, Germany
Gadus morhua	930	In Progress	NA	Genofisk University of Oslo, Norwegian High-Throughput Sequencing Centre
Gasterosteus aculeatus	446.611	Assembly	2/1/2006	The Genome Assembly Team Broad Institute
Haplochromis burtoni	1000	In Progress	NA	Broad Institute
Labeotropheus fuelleborni Domwe Island	69.3117	Assembly	6/20/2008	Cichlid Genome Consortium School of Biology, Georgia Institute of Technology Joint Genome Institute
Lateolabrax japonicus	17	In Progress	NA	Institute of Bioinformatics, Anhui Normal University, China; Shanghai Sangon Biological Engineering Technology & Services
Leucoraja erinacea	3420	In Progress	NA	North East Cyber infrastructure Consortium Mount Desert Island Biological Laboratory
Maylandia zebra Mazinzi Reef	76.9832	Assembly	6/21/2008	Cichlid Genome Consortium School of Biology, Georgia Institute of Technology Joint Genome Institute
Mchenga conophoros	71.3915	Assembly	6/20/2008	Cichlid Genome Consortium School of Biology, Georgia Institute of Technology Joint Genome Institute
Nothobranchius furzeri	5.3	Assembly	3/17/2009	Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI) Kathrin Reichwald Dept. of Genome Analysis
Nothobranchius kuhntae	5.2	Assembly	3/17/2009	Leibniz Institute for Age Research - Fritz Lipmann Institute, Jena, Germany Dept. of Genome Analysis
Oreochromis niloticus	3000	Assembly	1/28/2011	Broad Institute Genome Assembly Team, Broad Institute Sequencing Platform Broad Institute
Oryzias latipes HNI	585.155	Assembly	5/7/2007	Medaka genome sequencing project University of Tokyo, Chiba , Japan
Oryzias latipes Hd-rR	700.37	Assembly	4/19/2006	Medaka genome sequencing project University of Tokyo, Chiba , Japan
Petromyzon marinus	NA	In Progress	9/30/2010	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine
Pundamilia nyererei	1000	In Progress	NA	Broad Institute
Rhamphochromis esox	69.8333	Assembly	6/23/2008	Cichlid Genome Consortium School of Biology, Georgia Institute of Technology Joint Genome Institute
Scophthalmus maximus	17	In Progress	NA	Institute of Bioinformatics, Anhui Normal University, China; Shanghai Sangon Biological Engineering Technology & Services Co. Ltd
Sebastes rubrivinctus	1000	In Progress	NA	University of Southern California
Takifugu rubripes	332.348	Assembly	8/22/2002	The Fugu Genome Sequencing Consortium DOE Joint Genome Institute, Institute of Molecular and Cell Biology, Singapore
Tetraodon nigroviridis	342.403	Assembly	5/14/2004	Genoscope

N.A.: Not available; Mb: millions of base pairs

7.2 Five completely sequenced genomes

The initial phase of teleost genomes sequencing mainly attempted to establish an ideal vertebrate model genome i.e. small-sized, less complexity, and highly homologous to human genome. Following sub-sections will describe in details the general characteristics of five completely sequenced teleost genomes considered in this study, as well as the annotation of their genomes. These five principal teleosts are particularly studied at the genetic and genomic levels and are subjected to whole-genome sequencing projects. This has allowed to perform different analyses and helped obtaining interesting results in the present thesis, particularly those related to understand biological function of particular sets of genes.

7.2.1 Pufferfish

The Japanese pufferfish (*Takifugu rubripes*) genome project was initiated in 1989 and was the second vertebrate genome sequenced after human (Aparicio, Chapman et al., 2002). The genome is about one-eighth the size of the human genome but contains approximately same number of genes and is the smallest known vertebrate genome (~390Mb). This difference is primarily due the fact that non-exonic regions (intronic and intergenic sequences) are generally shorter in the pufferfish than in humans, because of a relative paucity of repetitive sequences (less than 10% of repeat sequences) (Brenner, Elgar et al., 1993). Although human and fugu lineages diverged over 450 million years ago, the vast majority of genes identified in human have a counterpart in fugu.

Hence the compactness and a similar repertoire of genes to humans, makes it a useful 'reference' genome for identifying genes and other functional elements such as regulatory elements in human and other vertebrate genomes, and for understanding the structure and evolution of vertebrate genomes.

A 'draft' sequence of the fugu genome was determined by the International Fugu Genome Consortium in 2002, using the 'whole-genome shotgun' sequencing strategy (Aparicio, Chapman et al., 2002). **Fourth assembly, v4:** This assembly is based on

~8.7X coverage of the genome, and includes 7,213 scaffolds, covering approximately 95% of the non-repetitive fraction of the genome (http://www.fugu-sg.org/).

The green spotted pufferfish (*Tetraodon nigroviridis*), a closely-related species have a similar compact genome (~385Mb). *Tetraodon nigroviridis* is found in rivers and streams of Southeast Asia (Indonesia, Indochina, Malaysia, the Philippines), as well as in estuaries and mangrove swamps, and even occasionally in the sea; it is therefore not strictly limited to fresh water. The sequencing of the tetraodon genome at a depth of about 8X, carried out as collaboration between Genoscope and the Broad Institute of MIT and Harvard, was finished in 2002, with the production of an assembly covering 90% of the euchromatic region of the genome. Its gene assemblage is also very similar to that of mammals such as humans and mice. Importantly, hundreds of putative novel human genes have been discovered by comparing the pufferfish and human genome sequences using EXOFISH (**Exo**n **Fi**nding by **S**equence **H**omology), a tool for comparative genomics (Roest Crollius, Jaillon et al., 2000). In addition, *T. nigroviridis* genome allowed reconstructing many of the chromosome rearrangements which led to the modern human karyotype (Jaillon, Aury et al., 2004). **Assembly-V7** consists of 27918 annotated genes (http://www.genoscope.cns.fr/-externe/tetraodon/).

7.2.2 Zebrafish

Zebrafish (*Danio rerio*) is a small **tropical fresh-water** fish, taxonomically classified as a member of the Cyprinidae family. Its natural geographic distribution includes rivers of northern India, northern Pakistan, Nepal, and Bhutan in South Asia. During the past few decades, zebrafish has emerged as one of the most important model organism for genetic studies including vertebrate development, developmental biology, and a wide variety of human congenital and genetic diseases. Short generation time, producing large clutches of eggs, and external fertilization and embryo development are some of the features that make the zebrafish experimentally amenable.

The sequencing of zebrafish genome began in 2001. The genome ~3113Mb in size proved extremely valuable for functional and comparative genomics studies and becomes the most well-established teleost model for studying gene functions. **Seventh assembly, Zv7**: The July 2007 zebrafish (*Danio rerio*) Zv7 assembly was produced by "The Wellcome Trust Sanger Institute" (http://www.sanger.ac.uk/Projects/D_rerio), in collaboration with the Max Planck Institute for Developmental Biology in Tübingen, Germany, and the Netherlands Institute for Developmental Biology (Hubrecht Laboratory), Utrecht, The Netherlands. Seventh assembly comprises a total sequence length of 1,440,582,308 bp in 5,036 fragments. The main zebrafish genomic resources available are namely: ZFIN (Zebrafish Information Network) (http://zfin.org/cgi-bin/webdriver?MIval=aa-ZDB_home.apg); Tubingen zebrafish lab (http://www.eb.tuebingen.mpg.de/departments/3-genetics/zebrafish); Johnson Laboratory (http://www.genetics.wustl.edu/fish_lab/).

7.2.3 Medaka

Medaka (*Oryzias latipes*) is a small freshwater fish native to eastern Asia, primarily Japan, Korea, and China. It is an egg laying, oviparous fish that can survive under a wide range of environmental conditions. Being easy to breed and having a short generation time of 2-3 months, medaka has being extensively used in both genetic and environmental research.

Medaka has a diploid chromosome number of 48 and an estimated genome size of 800-1000 Mb. It exhibits a nuclear DNA content of 2.2pg (Lamatsch, Steinlein et al., 2000). About half the size of the zebrafish genome, medaka is also an excellent model for studying vertebrate development and is another emerging teleost model ideal for functional genomics studies (Wittbrodt, Shima et al., 2002). There are currently two whole genome shotgun assemblies of the medaka genome available, one of strain HNI, and one of strain **Hd-rR**. (http://utgenome.org/medaka/).

7.2.4 Stickleback

Stickleback (*Gasterosteus aculeatus*) is a species of freshwater fish that have undergone a dramatic evolutionary radiation since the last Ice Age. Ancestral marine sticklebacks populated the newly created lakes and subsequently adapted to different environments. A number of sub-species have recently evolved multiple changes in their anatomical and physiological traits. Stickleback species are therefore a good model with which to study adaptive evolution.

Current assembly **gasAcu1.0**, released July, 2007 was produced by The Broad Institute. They have sequenced a freshwater three-spine stickleback to 9x coverage and have also provided 1xIllumina coverage of sticklebacks from ten marine populations and from ten additional freshwater populations.

Table 7.2: Physiological and environmental parameters of five teleosts.

Teleost	Max. Length(cm)	Environment	pH range	Climate	Temperature Range(°C)
Danio rerio	3-4	Freshwater	6.0 - 8.0	Tropical	18 - 24
Oryzias latipes	3.2	Freshwater	7.0 - 8.0	Subtropical	18 - 24
Gasterosteus aculeatus	11	Marine; freshwater	NA	Temperate	4 - 20
Tetraodon nigroviridis	17	Freshwater	8.0 - 8.0	Tropical	24 - 28
Takifugu rubripes	80	Marine; freshwater	NA	Temperate	NA

Fish biodiversity is important to humans at the economical, ecological and cultural points of view, and its maintenance is an important challenge for the next generations. The availability of the whole genome drafts from other teleost species are increasingly becomes available, opening wider perspectives for comparative studies for a better understating of vertebrate genomes, particularly the structure and organization of the genes and genomes relevant to evolution.

7.3 References

Altschul, S., T. Madden, A. Schaffer, J. Zhang, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**: 3389 - 3402.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." Science **297**(5585): 1301-10.

Barnett, V. and T. Lewis (1994). Outliers in statistical data: , John Wiley & Sons, Chichester.

Brenner, S., G. Elgar, R. Sandford, A. Macrae, et al. (1993). "Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome." Nature **366**(6452): 265-8.

Bucciarelli, G., G. Bernardi and G. Bernardi (2002). "An ultracentrifugation analysis of two hundred fish genomes." Gene **295**(2): 153-62.

Clarke, A. and N. M. Johnston (1999). "Scaling of metabolic rate with body mass and temperature in teleost fish." Journal of Animal Ecology **68**: 893-905.

Costantini, M., O. Clay, C. Federico, S. Saccone, et al. (2007). "Human chromosomal bands: nested structure, high-definition map and molecular basis." Chromosoma **116**(1): 29-40.

Federico, C., L. Andreozzi, S. Saccone and G. Bernardi (2000). "Gene density in the Giemsa bands of human chromosomes." Chromosome Res **8**(8): 737-46.

Gillooly, J. F., J. H. Brown, G. B. West, V. M. Savage, et al. (2001). "Effects of size and temperature on metabolic rate." Science **293**(5538): 2248-51.

Hegarty, T. W. (1973). "Temperature coefficient (Q_{10}), seed germination and other biological processes." Nature **243**: 305-306.

Hodkinson, I. D. (2003). "Metabolic cold adaptation in arthropods: a smaller-scale perspective." Functional Ecology **17**: 562-572.

Jaillon, O., J. M. Aury, F. Brunet, J. L. Petit, et al. (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-57.

Karamushko, L. I. (2001). "Metabolic Adaptation of Fish at High Latitudes." Doklady Biological Sciences **379**: 359-361.

Lamatsch, D. K., C. Steinlein, M. Schmid and M. Schartl (2000). "Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: detection of triploid Poecilia formosa." Cytometry **39**(2): 91-5.

Roest Crollius, H., O. Jaillon, A. Bernot, C. Dasilva, et al. (2000). "Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence." Nat Genet **25**(2): 235-238.

Smit, A., H. R and G. P (1996-2010). "RepeatMasker Open-3.0."

Tatusov, R. L., D. A. Natale, N. D. Fedorova, I. V. Garkavtsev, et al. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1): 22-8.

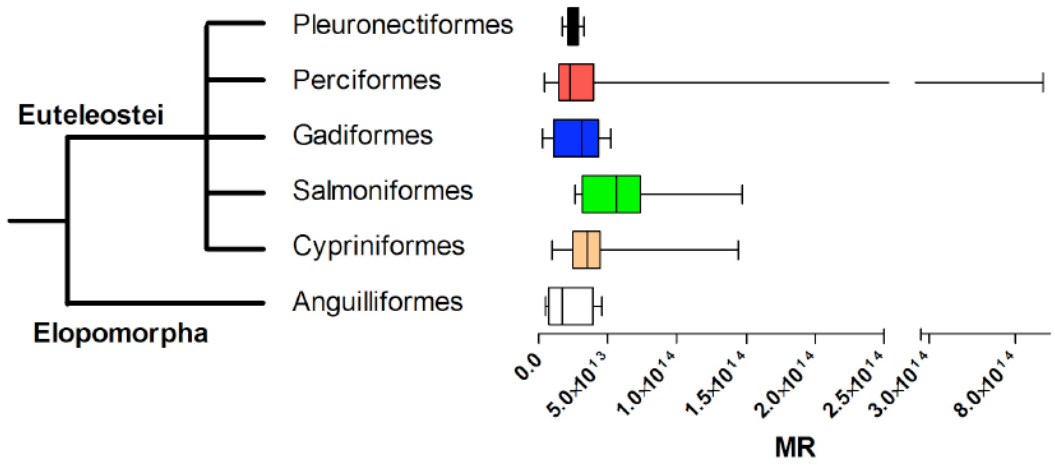
Tatusov, R. L., D. A. Natale, N. D. Fedorova, J. D. Jackson, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Varriale, A. and G. Bernardi (2006). "DNA methylation in reptiles." Gene **385**: 122-7.

Venkatesh, B. (2003). "Evolution and diversity of fish genomes." Curr Opin Genet Dev **13**(6): 588-92.

Wittbrodt, J., A. Shima and M. Schartl (2002). "Medaka--a model organism from the far East." Nat Rev Genet **3**(1): 53-64.

8 Supplementary materials



S1: Box-plot of metabolic rate, corrected to Boltzmann factor (MR), super imposed on a working phylogeny according to Clarke and Johnston (1999).

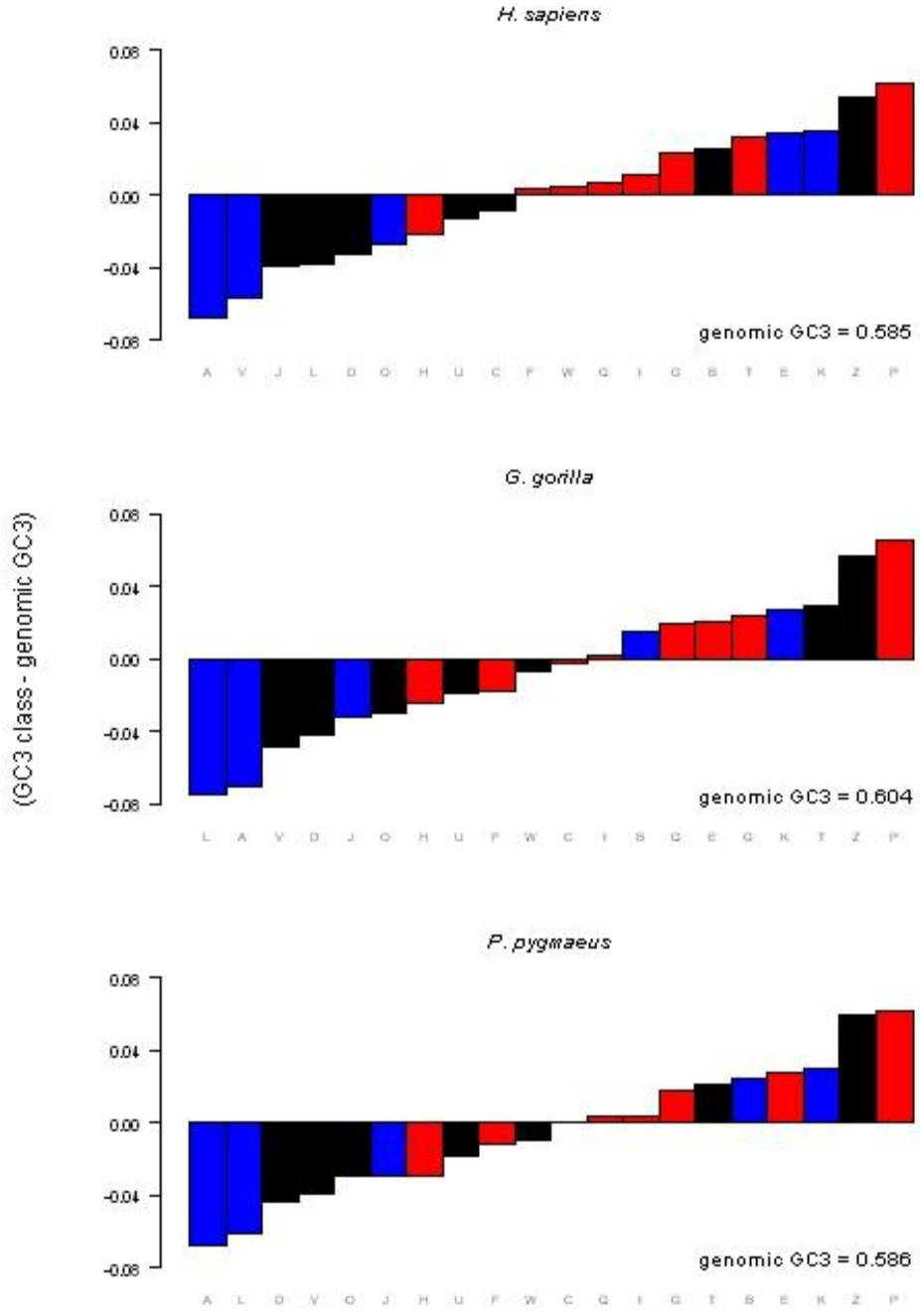
Table S2: Average GC content of orthologous intron pairs (GCi) in teleosts*

Species	*D. rerio*	*O. latipes*	*T. rubripes*	*G. aculeatus*	*T. nigroviridis*
D. rerio		0.37	0.36	0.36	0.36
O. latipes	0.40		0.4	0.39	0.4
T. rubripes	0.44	0.44		0.44	0.44
G. aculeatus	0.43	0.43	0.43		0.43
T. nigroviridis	0.47	0.46	0.47	0.47	

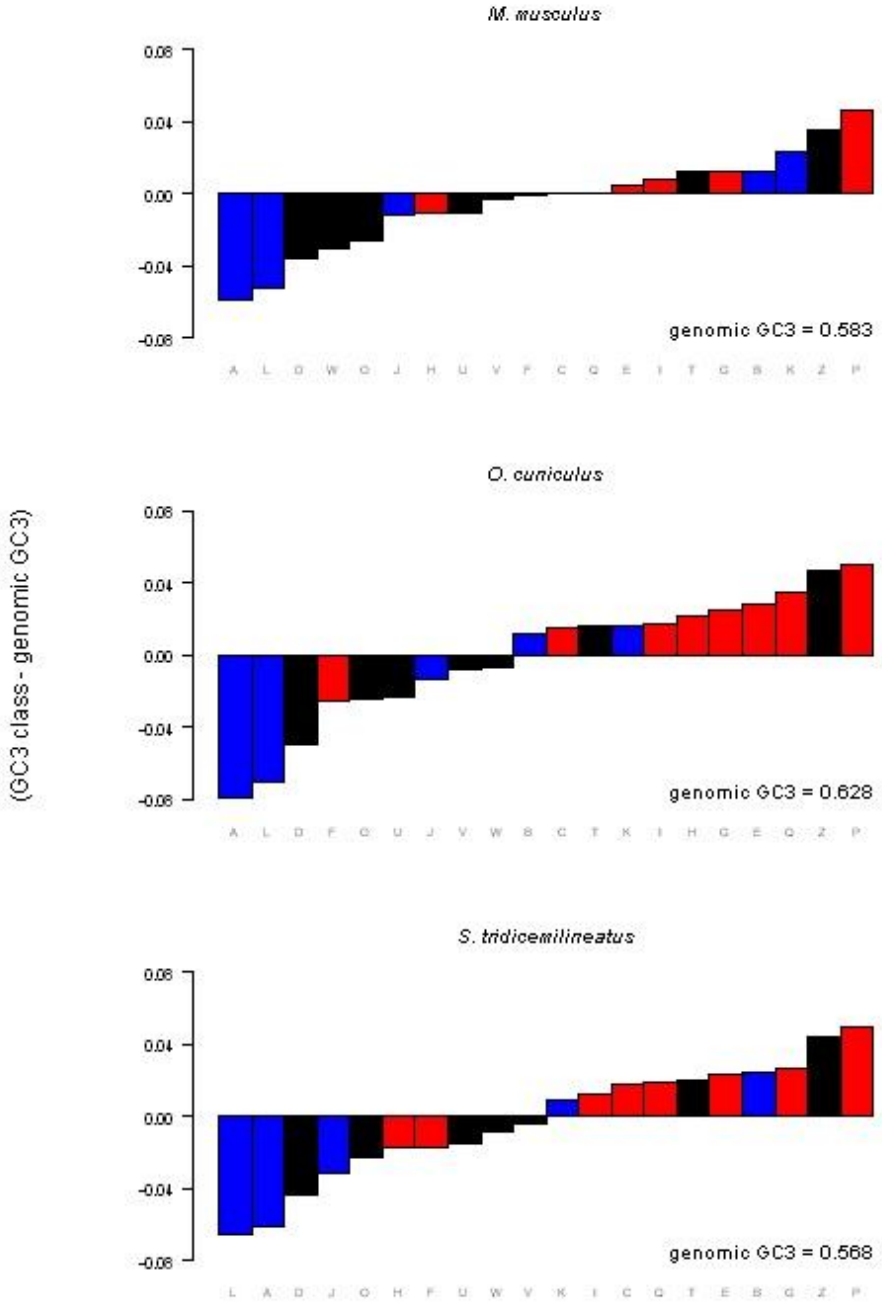
Table S3: Average intron length of orthologous pairs (bpi) in teleosts*

Species	*D. rerio*	*O. latipes*	*T. rubripes*	*G. aculeatus*	*T. nigroviridis*
D. rerio		13535	19936	18277	18737
O. latipes	3488		3543	3423	3444
T. rubripes	4212	2527		3923	3817
G. aculeatus	5298	3414	5405		4968
T. nigroviridis	2901	2077	3001	2810	

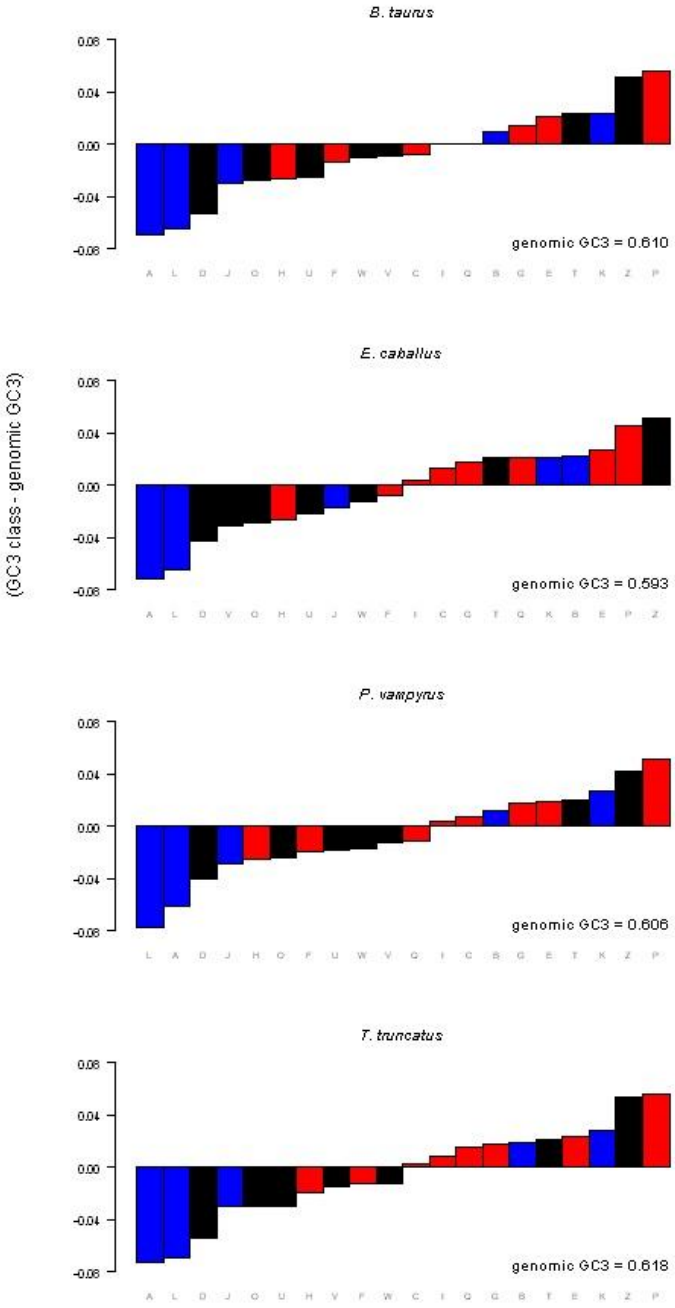
***Reading Table S2 and S3:** The values should be read along the rows. For example orthologous pair genes between *D. rerio and O. latipes*, the average GCi% of *D. rerio* is 0.37 and that of *O.* latipes 0.40. Similarly the average bpi% is 13535 and 3488, respectively.



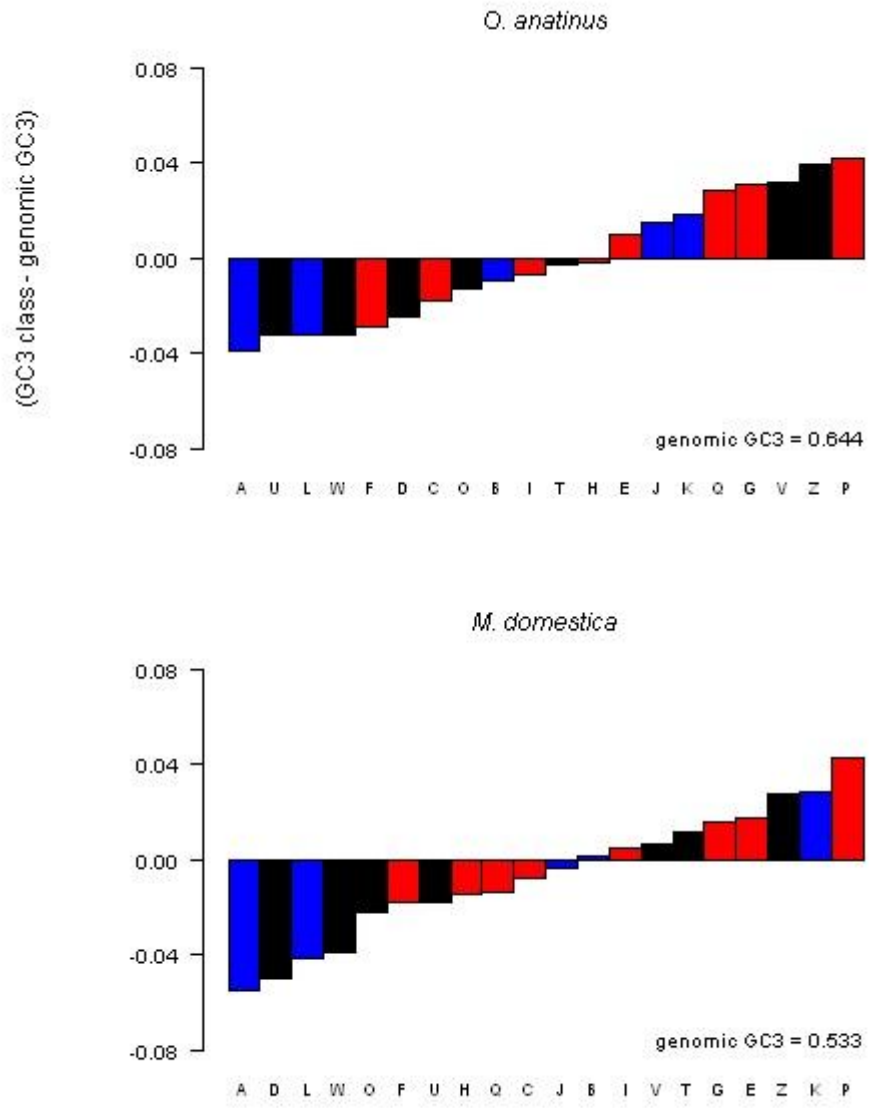
S4: Histograms of delta between average genomic GC3 levels against that of in each functional class in *H. sapiens*, *G. gorilla* and *P. Pygmaeus*.



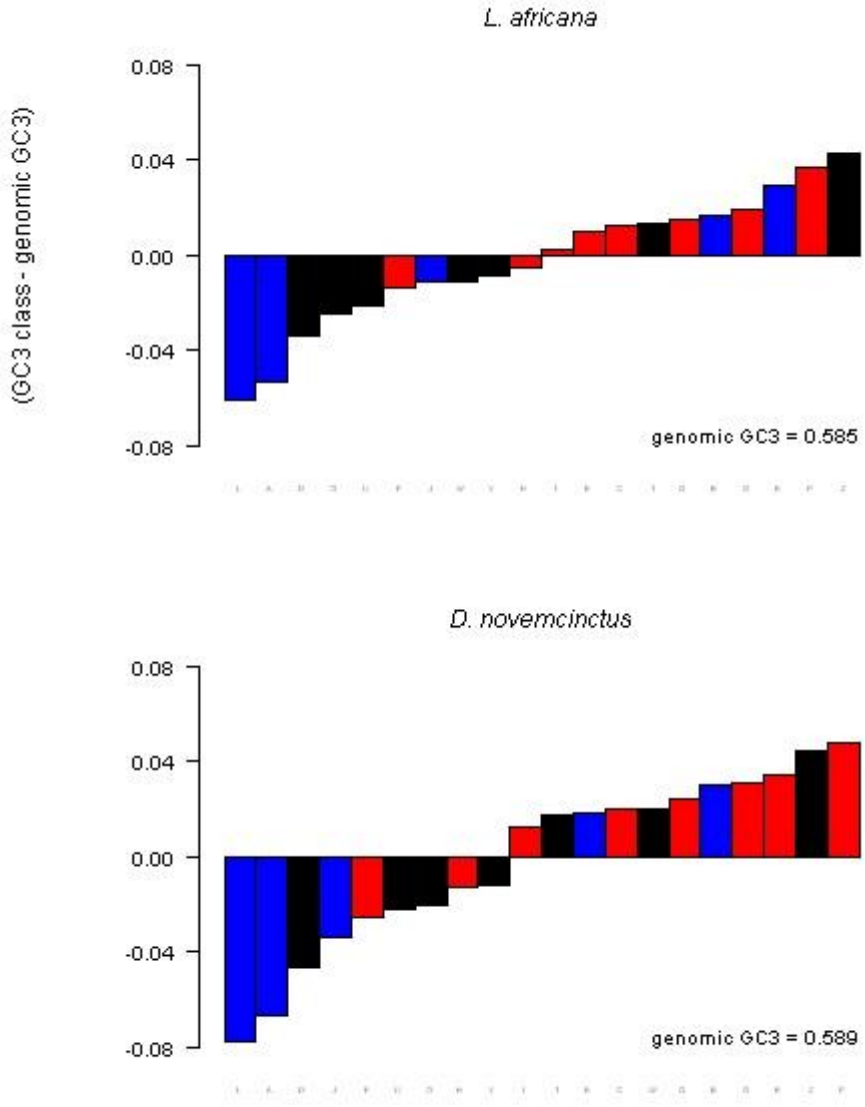
S5: Histograms of delta between average genomic GC3 levels against that of in each functional class in *M. musculus, O. cuniculus* and *S. tridecemilineatus*.



S6: Histograms of delta between average genomic GC3 levels against that of in each functional class in *B. taurus*, *E. caballus*, *P. vampyrus* and *T. truncatus*.



S7: Histograms of delta between average genomic GC3 levels against that of in each functional class in *O. anatinus* and *M. domestica*.



S8: Histograms of delta between average genomic GC3 levels against that of in each functional class in *L. africana* and *D. novemcintctus*.

KOG genes

Regression Coefficient of GC12 vs. GC3, %

|R| 0.589 - R Squared 0.347

	Coefficient	Std. Error	t-Value	P-Value
Intercept	36.099	0.141	255.186	<0.0001
GC3%	0.214	0.002	91.374	<0.0001

Regression Coefficient of GCi vs. GC3, %

|R| 0.805 - R Squared 0.648

	Coefficient	Std. Error	t-Value	P-Value
Intercept	0.241	0.002	127.281	<0.0001
GC3%	0.384	0.003	125.330	<0.0001

Regression Coefficient of GC vs. GC3, %

|R| 0.761 - R Squared 0.579

	Coefficient	Std. Error	t-Value	P-Value
Intercept	0.209	0.002	85.904	<0.0001
GC3%	0.428	0.004	108.463	<0.0001

Regression Coefficient of 5'-Flanking vs. GC3, %

|R| 0.533 - R Squared 0.284

	Coefficient	Std. Error	t-Value	P-Value
Intercept	0.352	0.003	119.758	<0.0001
GC3%	0.276	0.005	58.187	<0.0001

S9: Statistical report for KOG classified genes.

KOG Functional Classes

Regression Coefficient of GC12 vs. GC3, %

|R| 0.469 - R Squared 0.220

	Coefficient	Std. Error	t-Value	P-Value
Intercept	39.616	3.817	10.379	<0.0001
GC3%	0.223	0.065	2.314	<0.0320

Regression Coefficient of GCi vs. GC3, %

|R| 0.819 - R Squared 0.671

	Coefficient	Std. Error	t-Value	P-Value
Intercept	0.338	0.022	15.560	<0.0001
GC3%	0.223	0.037	6.061	<0.0001

Regression Coefficient of GC vs. GC3, %

|R| 0.678 - R Squared 0.460

	Coefficient	Std. Error	t-Value	P-Value
Intercept	0.351	0.029	12.293	<0.0001
GC3%	0.189	0.048	3.913	<0.0001

Regression Coefficient of 5'-Flanking vs. GC3, %

|R| 0.298 - R Squared 0.089

	Coefficient	Std. Error	t-Value	P-Value
Intercept	0.461	0.041	11.176	<0.0001
GC3%	0.092	0.070	1.325	<0.2016

S9: Statistical report for genes grouped in the KOG functional classes.

List of Publications

1. **Metabolic rate and genomic GC. What we can learn from teleost fish.**

   Uliano. E[a,], **Chaurasia. A[b]**, Bernà. L[b], Agnisola. C[a] and D'Onofrio. G.[b].

   Marine Genomics, Volume **3**, Issue 1, March 2010, 29-34.

   <doi:10.1016/j.margen.2010.02.001>

   [a] Department of Biological Sciences, University of Naples Federico II, Via Mezzocannone, 8 - 80134 Naples, Italy.
   [b] Laboratory of Animal Evolution and Physiology, Stazione Zoologica A. Dohrn, Villa Comunale, 80121 Naples, Italy.

2. **Does habitat affect genome features? A lesson from teleostean genome.**

   Chaurasia. A[b], Uliano. E[a,], Bernà. L[b], Agnisola. C[a] and D'Onofrio. G.[b].

   In Press, Fish Ecology, Nova Science Publishers, Inc. Hauppauge, NY 11788 USA.

   [a] Department of Biological Sciences, University of Naples Federico II, Via Mezzocannone, 8 - 80134 Naples, Italy.
   [b] Laboratory of Animal Evolution and Physiology, Stazione Zoologica A. Dohrn, Villa Comunale, 80121 Naples, Italy.

3. **The footprint of metabolism in the organization of mammalian genomes.**

   Bernà. L[a], **Chaurasia. A[a]**, Angelini. C[b], Federico. C[C], Saccone, S[C] and D'Onofrio. G.[a].

   Submitted to BMC Genomics**,** July, 2011.

   [a] Laboratory of Animal Physiology and Evolution.Stazione Zoologica A. Dohrn.Villa Comunale, 8012, Naples Italy.
   [b] Istituto per le Applicazioni del Calcolo "Mauro Picone", IAC-CNR, Via Pietro Castellino, 111 - 80131 Napoli – Italy.
   [c] Università di Catania. Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Via Androne,81- 95124 ,Catania- Italy.