Comparative genome analyses of deuterostomes: Metabolism and base composities of tunicates, fish and mammals genomes

Thesis

UNRESTRICTED

Comparative genome analyses of Deuterostomes

Metabolism and base compositions of tunicates, fish and mammals genomes

Luisa Berná Zanotta

Bachelor of Science in Biochemistry, School of Sciences, Universidad de la República,

Montevideo - Uruguay

Doctor of Philosophy

Sponsoring establishment

Stazione Zoologica A. Dohrn

Napoli, Italy

September 2010

DATE OF SUBMISSION: 30 SEP[T] 2010

DATE OF AWARD: 11 JAN 2011

Director of studies:

Dr. Giuseppe D'Onofrio

Laboratory of Animal Physiology and evolution - Evolutionary Genomics -

Stazione Zoologica A. Dohrn. Naples, Italy.

External Supervisor:

Dr. Pietro Liò

Computer Laboratory, University of Cambridge, U.K.

Examination pannel:

Dr. Édouard Yeramian

Structural Bioinformatics group, Pasteur Institute, Paris, France.

Dr. Paolo Sordino

Laboratory of Cellular and Developmental Biology, Stazione Zoologica A. Dohrn. Naples Italy.

To Rodrigo,

Abstract

The hypothesis that the metabolic rate could affect the base composition of genomes, i.e. the GC content, was tested along the phylogenetic line of deuterostomes, in the classes of tunicates, fish and mammals, by different approaches dictated by available data.

Significant differences were found between the GC content of the completely sequenced genomes of *C. intestinalis* and *C. savignyi*. Interestingly, the increment was higher in *C. savignyi* in coding and non-coding regions, and mainly at the third codon positions (7% GC3). Methylation process and substitution rate were checked if responsible of the observed differences. The frequencies of CpG, CpA and TpG were no different between the two tunicates, and (even if tunicates showed a substitution rate 50% faster than that of vertebrates) the base composition was not directly affected. Multivariate analysis on codon usages showed that: i) in both species selection acts on codon usage, shaping translationally preferred codons between highly and lowly expressed genes; and ii) gene expression level affects the GC content of the two tunicate genomes.

Data on the metabolic rate and the average genomic GC content were available for more than 200 teleostean species. Analyzing data according to the habitats, significant differences were found, with polar fish characterized by the highest metabolic rate and by a high GC content. A significant correlation was found between the two variables.

Finally, intra-genome analyses of functional classes in more than ten mammalian genomes, confirmed the link between metabolism and GC content. In each mammalian genome, indeed, the average GC3 level was higher in genes involved in "metabolic processes". The same genome organization was not found in amphibians and reptiles, characterized by lower metabolic rates than mammals.

The various approaches converged towards a coherent picture, highlighting the importance of metabolic rate as a factor shaping the base composition of genomes.

Acknowledgments

I want to thank my director of studies Dr. Giuseppe D'Onofrio with his endless enthusiasm for science, for discussing hypotheses, for supporting me and growing with me. I would also like to thank my supervisor Dr. Pietro Liò for his suggestions and support during my PhD studies.

Moreover I am so grateful to Dr. Fernando Alvarez-Valin, who not only drove me in this experience, but also has taught me the importance of enjoying the small things of life.

I would also like to thank all the people that helped me on my daily work, specially to Guillermo Lamolle, for coming to Naples and sing tangos with me and for innumerable graceful chats, but also for his help on performing the majority of the scripts I have used! Also to Ankita Chaurasia, Nicola's Arrambide, Diego Castagna, Miguel Ponce de Leon, Hector Romero, Hector Musto, Laurent Dubroca, and Giuseppe Bianco.

La cosa più importante in questi tre anni passati nella bellissima e a volte tragica Napoli non è stato fare un dottorato, ma avere avuto l'oportunità di conoscere molte persone con le quali ho condiviso tanti momenti indimenticabile, ho stretto legami e relazioni che so, dureranno per sempre. Pertanto, vorrei ringraziare Agostina, per avermi aperto le porte a voi, per essere sempre state assieme, per la sua gioia e la sua forza, per essere stata come una sorella. A Cecilia per essere cresciute assieme, per la nostra affinità, per esserci stata sempre, per me e per tutti, per avermi accettata e capita, per le risate assieme e il "brbr" che a volte condividiamo. A Simo, per vedere tutto da un'altra prospettiva, per essersi aperta a me, per fare parte della "casa della morte" e andare avanti. A Benedetto, per le tante serate passate insieme, per l'ironia, per le nostre discussioni, per avermi fatto capire il bello di porsi delle domande, per la sua curiosità, per le grappe insieme e le interminabili risate durante le sere del "trio dinamico". A Laurent, per il suo spirito, per rendere tutto più divertente, per le iniziative, per la sua disponibilità, e ovviamente per i tanti momenti passati insieme, ed ancora, le sere del "trio dinamico". Ad Andrea, per la sua volontà di fare, la

sua apertura, per esserci sempre stata. A Giuseppe, per il suo affetto, il suo sorriso e la simpatia e per le pizze a casa sua! A Rosario, per farmi ridere, per il migliore caffè preparatp al terzo piano, per aver saputo superare gli ostacoli ed andare avanti. A Carmen, per muovere tutti, per le feste insieme, per non fermarsi mai. A Vasco, credo di non aver mai conosciuto una persona così attenta con gli altri, aperta e buona. Per aver condiviso questi due anni insieme e per essere sempre stato presente in tutti momenti difficili e anche quelli buoni, "frate'". A Enrique, per la sua gioia e semplicità. A Pedro, per la sua cucina e le sue canzoni! A Luciana agripina, per essere la persona più aperta, con una voglia intrinseca di fare ed una gioia contagiosa. A Maurizio, per la sua simpatia e le discussione aperte. E grazie a tutti coloro (altementi scriverei un'altra tesi di ringraziamenti) che hanno condiviso con me, momenti di confronto e momenti di compagnia, sofferenze e gioie, grazie! A Francesca, Bruno, Daniele, Mariano, Sylvie, Dorris, Chiara, Maria Letizia, Isabella, Diego, Ferdinando, Raffaele, Gianluca, Paolo, Gabriele, Monia, Valeria.

No solo estoy agradecida, estoy orgullosa de haber tenido a Rodrigo a mi lado, quien no solo me ha acompaado en este proyecto, sino que de alguna forma lo ha hecho posible. Gracias.

Muchas personas están además implicadas al hecho que haya concluido con esta etapa. Me gustaría dedicar esta tesis especialmente a mi abuela, que me ha enseado tantas cosas que ningún doctorado lograría estar al nivel. A mis padres que me han apoyado siempre y me han dado todas las libertades para crecer. A mis hermanas, Leonor y Lucía incondicionales y hermosas. A mis sobrinas por su alegría, sus sonrisas y por hacerme tíaluisi. A toda mi familia, a mis primos y tíos, presentes y ausentes. A Pao, sin moñ o pero objetivamente y subjetivamente, la mejor. A Kari, no sólo por su practicidad y su forma de salir adelante, también por su forma de ver las cosas y crecer conmigo. A Fabián quién es responsable que haya elegido este camino, y en él a todos los que lo han compartido conmigo, Victoria, Silvia, y todo el "cuarto" piso y aledaos, o sea Leonel , yuyo, el pantera, Miguel , Musto, Rúben, Adriana, Natalia y Hugo. Obviamente entre ellos, el pelo, a quien le estoy profundamente agradecida, por confiar y creer en mi, por abrirme las puertas, por conocerme y tener razón.

Contents

List of Abbreviations

3'GC	average of the molar ratio of guanine and cytosine at 5' of the CDS, 10
5'GC	average of the molar ratio of guanine and cytosine at 5' of the CDS, 10
AT	molar ratio of adenine + thymine in DNA, 4
C3	molar ratio of cytosine at the third codon position, 13
CDS	CoDing Sequence, region of nucleotides that corresponds to the sequence of amino acids in the predicted protein, 10
CoA	Correspondence analysis, it is an ordination technique that identifies the major trends in the variation of the data and distributes genes along continuous axes in accordance with these trends, 14
CpG	dinucleotide that are the target of methylation process is shorthand for cytosine and guanine separated by a phosphate, which links the two nucleosides together in DNA, 8
EST	Expressed Sequence Tag, 8
G3	molar ratio of guanine at the third codon position, 13
GC1	molar ratio of guanine and cytosine at the first codon position, 10
GC2	molar ratio of guanine and cytosine at the second codon position, 10
GC3	molar ratio of guanine and cytosine at the third codon position, 8
GCcds	average of the molar ratio of guanine + cytosine of coding sequences, 10
GCg	average of the molar ratio of guanine + cytosine of the whole genome, 10
GCi	average of the molar ratio of guanine and cytosine of intron sequences, 10

JTT Jones, Taylor, Thornton moldel. JTT distances correct for multiple substitutions based on the model of amino acid substitution described as substitution-rate matrix, 30

kb kilobase equal to 1000 base pairs of DNA, 5

KOG EuKaryotic clusters of Orthologous Groups of proteins, 59

R^2 coefficient of determination, 10

RBH Reciprocal Best Hits, 29

RSCU Relative Synonymous Codon Usage, 8

TpA dinucleotides where a thymine nucleotide occurs next to a adenine nucleotide in the linear sequence of bases along its length, 21

List of Figures

List of Tables

List of Supplementary figures

1 | *General Introduction*

"Since Zuckerkandl and Pauling established DNA as a document of evolutionary history, many approaches have been devised to transform the information enclosed in genomes into knowledge of their history. Comparative genomics and phylogenetic have emerged as the keys to many biological questions, such as the identification of functional sequences in complete genomes, the principles of genome architecture, the reconstruction of life's evolution and the inference of ancestral phenotypes and population characteristics" (Boussau and Daubin, 2010)

1.1 Surrounding Deuterostomes

The superphylum Deuterostomes is a subtaxon of Bilateria, traditionally distinguished by different developmental features from Protostomes. One of the features that distinguish the two superphyla is the mouth opening. While in Protostomes the site of gastrulation initiation (the blastopore) becomes the mouth, in Deuterostomes becomes the anus (Nielsen, 2001), thus the name from the Greek "second mouth" or "other mouth". The delimitation of the two groups is still matter of debate. For example the "historical" question about the phylogenetic position of chaetognaths and lophophorates, reclassified as Protostomes (Telford and Holland, 1993; Halanych et al., 1995)) or the appearance of new phylum (Bourlat et al., 2006). Nowadays Deuterostomes are considered to be composed by four phyla: Echinoderms (sea stars, sea cucumbers, sea urchins, etc.), Hemichordates (acorn worms and pterobranchs), Chordates (Winchell et al., 2002) and the recently added Xenoturbellida (Bourlat et al., 2006). These phyla arose from a common ancestor estimated to have lived more than 550 million years ago.

The phylum Chordata diverged into three subphyla: Cephalochordates (represented by nearly 29 species), Tunicates (or urochordates, represented by more than 3000 species) and Vertebrates (represented by almost 53000 species) (Hedges, 2002; Vienne and Pontarotti, 2006). These three groups share several characteristics that are always found in the larval forms or in the early embryo, but may be absent in the adult (Putnam et al., 2008):

– i – The notochord, what they are named after, which is a rod that supports the nerve cord.

– ii – The neural tube, a bundle of nerve fibres which connect the brain with the muscles and organs.

– iii – The gill or pharyngeal slits, which are openings that connect the inside of the throat to the outside of the neck, usually close or developed into other structures in vertebrates.

– iv – The post-anal tail, a muscular tail that extends backwards behind the anus.

– v – The endostyle, a longitudinal ciliated groove on the ventral wall of the pharynx which produces mucus to gather food particles.

More than 140 years ago, the Russian embryologist A. Kowalevsky first recognized that tunicates have vertebrate-like characteristics and speculated that vertebrates may have evolved from tunicate-like ancestor (cited in Holland and Gibson-Brown (2003)). Interestingly, Darwin supported the same point of view (cited in Caestro et al. (2003)), and new phylogenetic studies about chordates revealed that tunicates were, indeed, the sister species of vertebrates (Bourlat et al., 2006; Delsuc et al., 2006, 2008; Dunn et al., 2008; Putnam et al., 2008). The crucial phylogenetic position of tunicates prompted extensive studies in order to understand the evolutionary origin of vertebrates, and, in particular, the mechanisms of vertebrate development (Meinertzhagen and Okamura, 2001; Davidson and Levine, 2003; Satoh, 2003).

Tunicates are traditionally divided in three classes: the great majority of species belong to the class Ascidiacea (commonly known as "sea squirt"), whereas a low number of species belong to the classes Thaliacea and Appendicularia, both exclusively consisting of planktonic species. However, their precise evolutionary relationships are still matter of controversy since a recently pub-

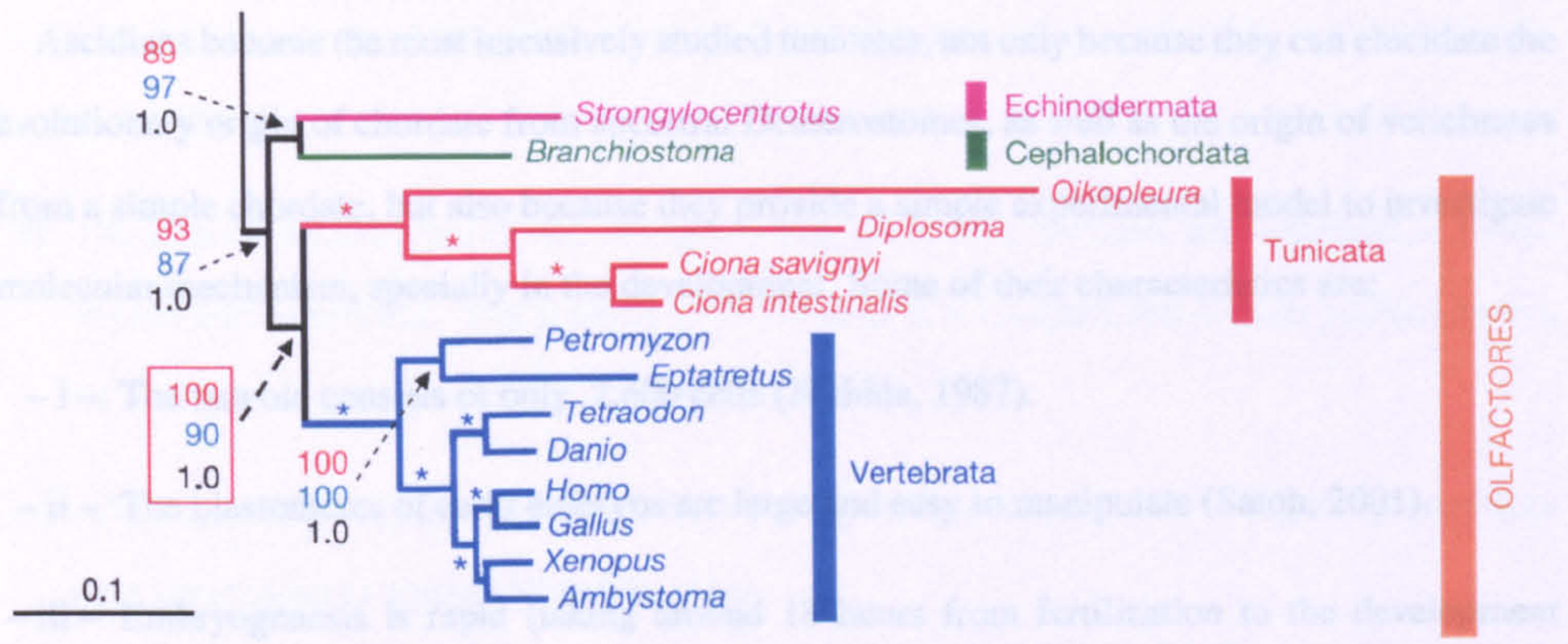


Figure 1.1: Chordate phylogeny, from Delsuc et al. (2006)

lished phylogeny based on the 18S rRNA suggesting a somewhat different point of view (Tsagkogeorga et al., 2009). Indeed, the new tree constructed using more data and taking into account phylogenetic reconstruction artefacts (namely the long-branch attraction due to the high rates of evolution of tunicates that can lead to homoplasy) supported the division of tunicates into three major clades: Appendicularia, Stolidobranchia, and Phlebobranchia + Thaliacea + Aplousobranchia (Tsagkogeorga et al., 2009). The former being peculiar because, departing from the common developmental program of metamorphosis, retains, indeed, the larval characteristics for their entire lifespan. Consequently this class is of pivotal interest from an evolutionary developmental point of view (Wada, 1998). The latter contains the ascidian family, that is of extreme importance as experimental model, specially in development studies. More recently their importance grew because two genomes were completely sequenced.

1.2 Ascidians

Ascidians are sessile marine animals widely distributed and represented by more than 2000 species. They are simple filter feeders, living solitary or in developing colonies. Regarding reproduction they are hermaphroditic with a spawing triggered by light following a period of dark. The adults have a relatively simple and symmetrical body plan and are characterized by a tough outer "tunic" made of the polysaccharide tunicin covering completely the surface of solitary individuals or forming a common matrix in colonial species (Harrison and Ruppert, 1997).

Ascidians become the most intensively studied tunicates, not only because they can elucidate the evolutionary origin of chordate from ancestral Deuterostomes, as well as the origin of vertebrates from a simple chordate, but also because they provide a simple experimental model to investigate molecular mechanism, specially in the development. Some of their characteristics are:

– i – The tadpole consists of only 2,600 cells (Nishida, 1987).

– ii – The blastomeres of early embryos are large and easy to manipulate (Satoh, 2001).

– iii – Embryogenesis is rapid (taking around 18 hours from fertilization to the development of a free-swimming tadpole, at 18 °C) and the entire life cycle takes less than 3 months (Sordino et al., 2001; Nakatani et al., 1999).

– iv – Novel functions of developmental genes can be determined by mis-expressing or over-expressing regulatory genes that encode transcription factors and signaling molecules (Takahashi et al., 1999; Imai et al., 2002).

– v – Transgenic DNA can be introduced into developing embryos using simple electroporation methods (Corbo et al., 2001; Gregorio and Levine, 2002)

1.3 The Ciona genomes

Among the ascidians, the most representative and studied organisms are those belonging to the genus of Ciona. Apart from the evolutionary and experimental advantages already mentioned, another bonus of this species is the relatively small genome size, estimated in only 160 million bases pair, similar to that of invertebrates (i.e. 180 Mb *Drosophila melanogaster*), and about 20 times smaller than the human genome (Adams et al., 2000; Simmen et al., 1998). The draft genome of *Ciona intestinalis* published in 2002 provided new insights into the evolutionary origins of the vertebrates (Dehal et al., 2002). Some important features of the Ciona genome are:

i) The genome contains 14 chromosomes (Shoguchi et al., 2006) and is notably AT rich, with an average AT content of 65% (Dehal et al., 2002).

ii) A total of 15.500 (±3.700) protein-coding genes were estimated (Simmen et al., 1998), this number is comparable with the number of genes described in *D. melanogaster* (13.600 pro-

teins (Adams et al., 2000)), as well as in *Caenorhabditis elegans* (19.000 proteins (Consortium, 1998)).

iii) The general organization of the protein-coding genes is in between that observed in Protostomes and vertebrates. The genome of *C. intestinalis* is compact and densely packed having on average one gene per 6.8 kb (Dehal et al., 2002).

iv) Almost 60% of the Ciona gene models have a detectable protostome homolog and presumably they correspond to ancient bilaterian genes.

v) 16% lack a clear protostome homolog, yet possess a recognizable vertebrate counterpart. These genes might arose in the modern deuterostome before the chordate divergence.

vi) 20% have no clear homolog in fly, worm, pufferfish, or human.

vii) Some genes are found in multiple copies in *C. intestinalis* but present in single copies in vertebrates.

viii) Several genes that are conserved in other animals appear to be missing in *C. intestinalis* (Dehal et al., 2002). Most of the genes that were lost by Ciona are observed in the genomes of invertebrates and also remain in the genomes of vertebrates. This observation indicates that these losses are lineage specific and hence led to the proposal that the genome of Ciona is not representative of the ancestral chordate genome (Hughes and Friedman, 2005).

After the publication of the draft of *C. intestinalis* genome, another Ciona genome was published, that of *C. savignyi* (Vinson et al., 2005; Small et al., 2007). The morphology of *C. intestinalis* and *C. savignyi* is very similar, and cases of misidentification, indeed, have been reported (Hoshino and Nishikawa (1985), and references therein). In fact, the adult of the two species can only be distinguished by minor features such as the presence of an endostylar appendage in *C. intestinalis*, absent in *C. savignyi*; by the location of the pharingeo-epicardic openings and by the test, that is "soft gelatinous" in one species, and "hard and cartilaginous" in the other (Hoshino and Nishikawa, 1985). Although no hybrids of the two species have been recognized in nature, they can be experimentally obtained by fertilizing either species' dechorionated eggs with heterologous sperm from the other. These hybrids are able to develop up to the tadpole stage (Byrd and

Lambert, 2000). While morphology and the possibility of producing hybrids suggest that these two Ciona species are close relatives, the comparisons using sequence data indicate that they are rather divergent (Johnson et al., 2004). Even more, inside the *C. intestinalis* species group, the divergence is quite remarkable and cryptic species were found (Caputi et al., 2007; Iannelli et al., 2007; Suzuki et al., 2005).

1.4 Thesis objectives

The aim of the present thesis is to shed light on which forces are driving the base composition evolution at the genome level. To this goal comparative genome analyses of Deuterostome were conducted.

It will be worth to recall shortly the current neutral and selective hypotheses proposed to explain the base composition variation at the genome level. Several hypotheses were drawn out: the bias gene conversion (BGC) (Eyre-Walker, 1993; Galtier et al., 2001; Duret and Galtier, 2009), the breakpoints distribution (Lemaitre et al., 2009); the thermal stability (Bernardi (2007) for a review), and the metabolic rate (Vinogradov, 2001, 2005).

The BGC hypothesis was grounded on a significant correlation between GC content and recombination process (Eyre-Walker, 1993). Theoretically the GC content variability is seen as a consequence of a random mutation/fixation process due to the synergy of recombination events with a biased DNA replication/repair system (Duret and Galtier, 2009).

Recently, Lemaitre and colleagues proposed a new model based on the analysis of DNA fragile-points, *i.e.* DNA breakpoints. A biased DNA repair system could lead to a GC increment, and, indeed, in the GC-rich regions of the human genome the authors found a highest occurrence of DNA breakpoints (Lemaitre et al., 2009).

According to the thermodynamic stability hypothesis, it is expected that the increment of environmental or body temperature produce a GC increment, stabilizing DNA, RNA and proteins (Bernardi, 2007). The hypothesis was based, on one hand, on the fact that DNA and RNA structure are stabilized by the triple hydrogen bonds of the GC pair; and on the other, on the fact that proteins are stabilized by the increasing occurrence of hydrophobic amino acids, mainly encoded by GC-rich codons (D'onofrio et al., 1999).

The metabolic rate hypothesis was based on the observation that two DNA features, bendability and nucleosome formation, are significantly correlated with the GC content. In turn, both bendability and nucleosome formation are affected by the increment of the transcriptional activity (Vinogradov, 2001, 2005).

Three different strategies and approaches were acted upon deuterostomes organisms with the aim to disclose among the forces proposed by the current evolutionary hypotheses to drive the base composition of genomes.

Regarding tunicates, a comparative analysis of base composition and substitution rate between Ciona intestinalis and C. savignyi were carried out. Specifically Chapter 2 ""Compositional study of Ciona genomes" describe the differences on base composition that these genome present and the analyses performed to understand the possible causes of these observations. On the other hand, Chapter 3 "High rate of evolution in urochordates" describe the characterization of the rate of molecular divergence of ascidian species and other tunicates, specifically that of Oikopleura dioica, trying to understand the evolutionary and biological reasons behind it.

Regarding teleosts, metabolic rate and genomic GC content were studied for more than 200 fish living in different habitats, the results are discussed in the light of the mentioned theories about genome evolution in Chapter 4 "Comparative analyses of base composition in vertebrate genomes".

Finally, regarding mammals, Chapter 4 also described the analyses of functional classes in more than ten mammalian genomes, showing the link between metabolism and GC content. Their genome organization was compared with amphibian and reptile genomes.

The overall results presented here favored the metabolic rate hypothesis as one of the possible forces driving compositional genome evolution.

2 | *Compositional Study of Ciona Genomes*

2.1 Introduction

C. intestinalis and *C. savignyi* belong to tunicates, now considered the sister group of vertebrates. Thanks to this key evolutionary position, they have been considered as fundamental organisms to clarify the origins of chordates and hence the origins of vertebrates from simpler organisms. This chapter presents a comparative analysis of their genomic features, particularly focused on the comparative analyses of base composition in different genome compartments. The GC level was higher in *C. savignyi* than in *C. intestinalis*. All regions so far compared showed the same trend. 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*.

Since CpG di-nucleotide (sites of DNA where a cytosine nucleotide occurs next to a guanine nucleotide) was pointed out as possible factor affecting the variation of the GC content, and since methylation process was reported to take place effectively in ascidians, a detailed analysis was performed on both coding and non-coding regions, as well as on the three reading frames, to assess the role, if any, played by this hyper-variable doublet. Multivariate statistical analysis of relative synonymous codon usage (RSCU) was also performed in both species, and revealed that: i) in *C. intestinalis*, the first axis was strongly correlated with the base composition at the third codon positions, whereas, on the second axis, putatively highly and lowly expressed genes were strongly clustered at the opposite extremes, an observation supported by the ESTs distribution along the axis2; ii) in *C. savignyi*, the first axis was strongly correlated with the base composition

at the third codon positions, and a correlation with gene expression level was observed on both first and second axis, again confirmed by the ESTs distribution along the axes. Finally, on the first axis a correlation with the GC3 increment was also found in *C. savignyi*, but not in *C. intestinalis*.

The results are discussed in the light of current theories about genome evolution.

2.2 Results

2.2.1 GC content

The genomes of both ascidians *C. intestinalis* and *C. savignyi* have been reported to be AT rich (Dehal et al., 2002; Roseto et al., 2002; Singh et al., 2009). However, an exhaustive investigation on the base composition in different genomic regions was never carried out. Hence, the average GC level of the whole genome (GCg), introns (GCi), 5'- and 3'-flanking regions, *i.e.* 2kb up- and down-stream CDS (GCf or, more precisely, 5'GC and 3'GC, respectively), as well as that of coding sequences (GCcds), and that of each codon position (GC1, GC2 and GC3) was computed (Table 2.1 and Figure 2.1). Within each genome, the average GC content of the regions so far analysed showed a similar ranking order, where GCi was the lowest and GC1 the highest value (Table 2.1).

Table 2.1: 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}	GC_1	GC_2	GC_3
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%
	Orth. genes								
C.intestinalis	7747		34.23%	36.93%	36.76%	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 pb.

In all pairwise 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 that of GC2 (0.4%), whereas the highest was that of GC3 (7.0%). A closer inspection of GC3 values clearly showed a bell-shaped normal distributions, skewed towards high GC3 values (Figure 2.2). 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 et al. (1999); Bernardi (2004), for a review).

Restricting the analysis to a set of orthologous sequences, similar results were found (Table 2.1). Once more, the coding positions showed higher GC values in *C. savignyi*. The highest difference between the two Cionas was found at the third codon position. The scatter plots of

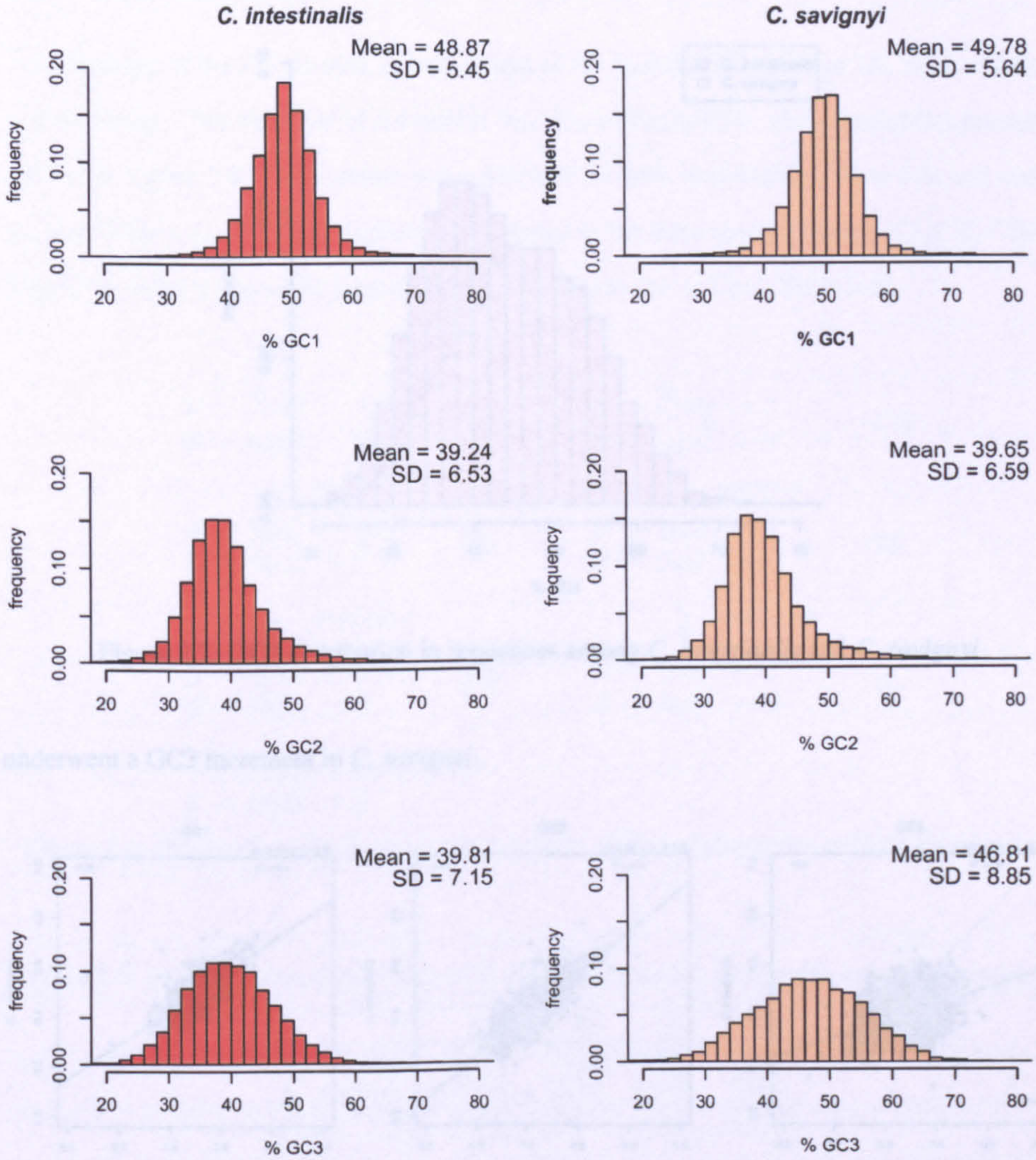


Figure 2.1: GC1, GC2 and GC3 distributions in sequences of *C. intestinalis* and *C. savignyi*

orthologous genes for each codon position, as well as the coefficients of determination were reported in Figure 2.3. In the case of GC1 and GC2, the coefficients of determination (R^2) were highly significant, 0.54 and 0.73, respectively, and intercepts and slopes of the regression lines were not different from the diagonals (slope = 1). In the case of GC3, the points were more scattered around the regression line, showing a R^2 value of 0.19. As much as 80% of the points were under the diagonal (5686 vs. 1422). In summary, the great majority of the orthologous genes

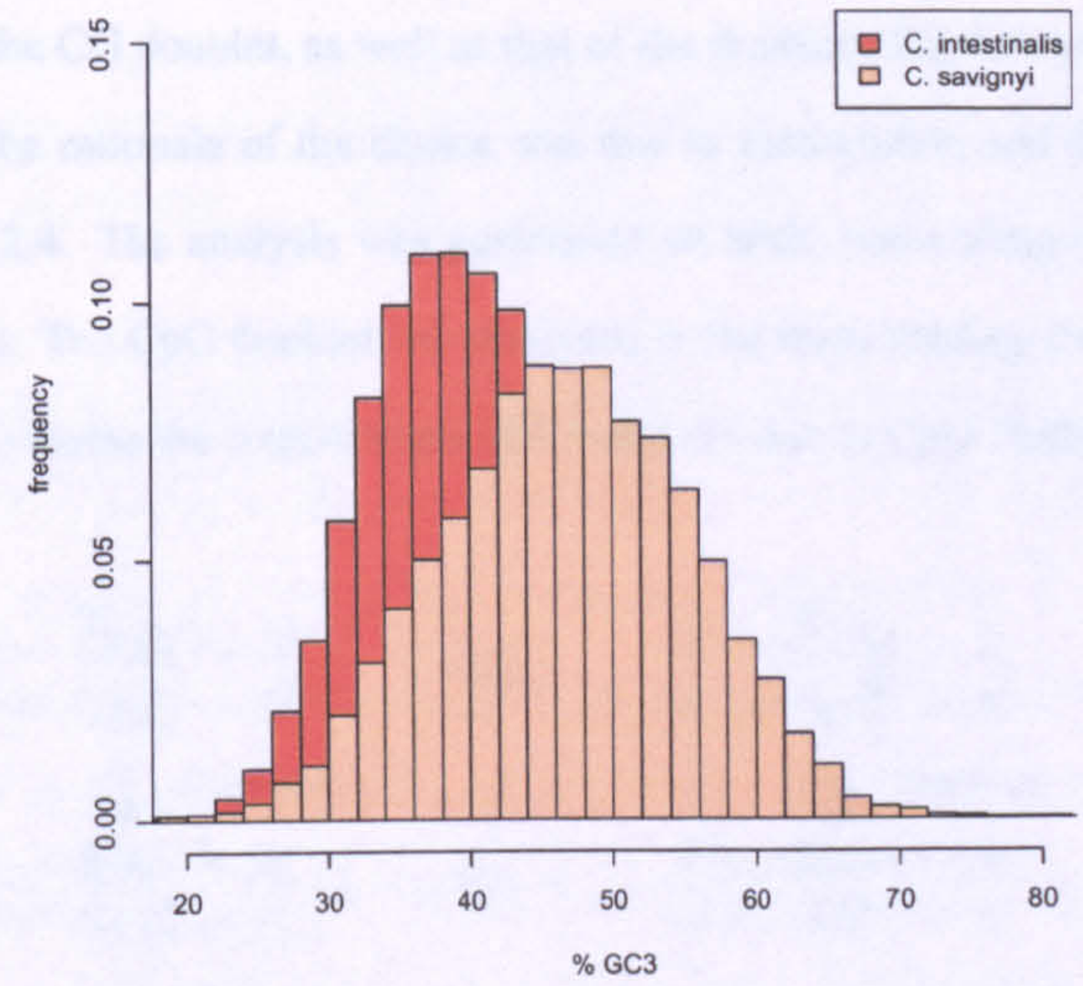


Figure 2.2: GC3 distribution in sequences among *C. intestinalis* and *C. savignyi*

underwent a GC3 increment in *C. savignyi*.

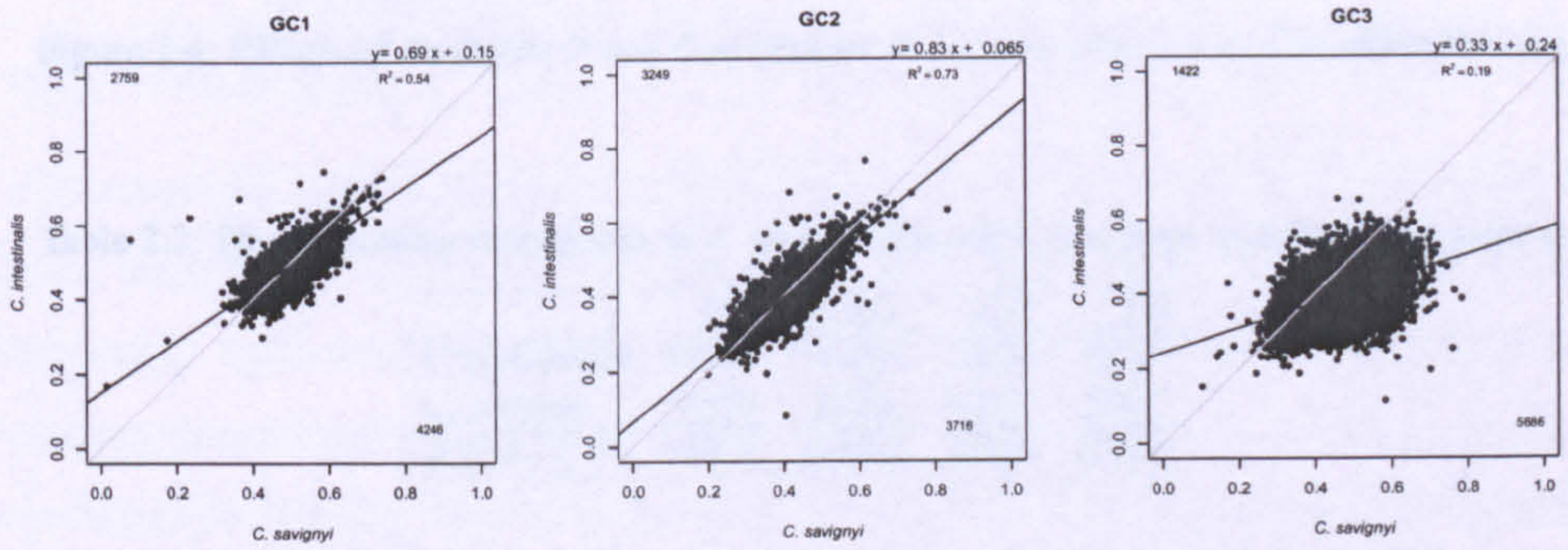


Figure 2.3: Scatter-plot of GC1, GC2 and GC3 for orthologous sequences of *C. intestinalis* and *C. savignyi*

2.2.2 CpG dinucleotides

The frequency of the CG doublet, as well as that of the doublets TG, CA and TA, were calculated in both Cionas. The rationale of the choice was due to methylation and deamination processes, showed in Figure 2.4. The analysis was performed on both, non-coding (Table 2.2) and coding regions (Table 2.3). The CpG doublet was analysed in the three reading frames (C1pG2, C2pG3, C3pG1) in order to assess the total amount of C3 and G3 due to CpG (Table 2.4).

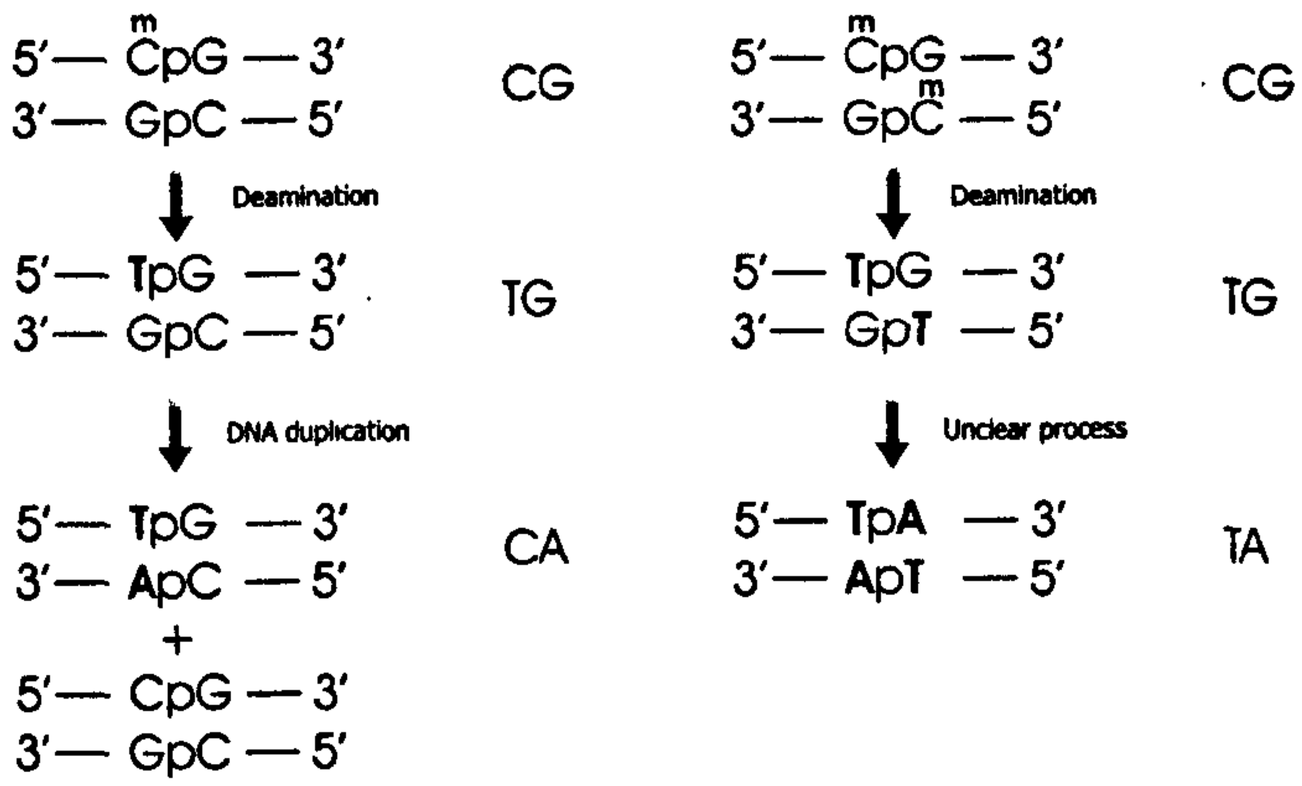


Figure 2.4: Effects of methylation and deamination of 5-methylcytosine in CpG dinucleotides

Table 2.2: Di-nucleotides frequencies of *C. intestinalis* and *C. savignyi* in non-coding regions

	CG	TG	CA	TA
C. intestinalis	0.026	0.067	0.064	0.095
C. savignyi	0.028	0.066	0.064	0.090
Δ *C. s - C. i*	0.003	-0.001	0.000	-0.005

Table 2.3: Di-nucleotides frequencies of *C. intestinalis* and *C. savignyi* in coding regions

	CG	TG	CA	TA
C. intestinalis	0.036	0.076	0.075	0.052
C. savignyi	0.044	0.074	0.076	0.045
Δ C.s - C.i	0.008	-0.002	0.001	-0.007

In non-coding regions, the frequencies of doublets showed either minor or no differences. More precisely: i) the frequencies of the CG doublets were 0.026/0.028 in *C. intestinalis* and *C.*

Table 2.4: CpG frequencies in different codon positions, G3 and C3 of *C. intestinalis* and *C. savignyi*

	C1pG2	C2pG3	C3pG1	G3	C3
C. intestinalis	0.030	0.039	0.038	0.201	0.196
C. savignyi	0.036	0.050	0.046	0.230	0.236

savignyi, respectively; ii) those of TG were 0.067/0.066, respectively; iii) those of CA were 0.064 in both Cionas; and iv) those of TA were 0.095/0.090, respectively. In coding regions: i) the frequencies of TG and CA were practically identical, 0.076/0.074 and 0.075/0.076, respectively in the two organisms; and ii) the frequencies of CG and TA were mirroring each other, being the former 0.036/0.044 and the latter 0.052/0.045, in *C. intestinalis* and *C. savignyi*, respectively. The frequencies of the CG doublets in the three reading frames were always higher in *C. savignyi*. The highest delta was that of C2pG3 (1.1%), whereas the delta of C3pG1 and C1pG2 were, respectively 0.8% and 0.6%. In spite of the different increments, the contribution of C3pG1 and C2pG3 to the whole genomic amount of C3 and G3 (20.1%/23.0% and 19.6%/23.6%, for *C. intestinalis* and *C. savignyi*, respectively), was of the same order of magnitude, *i.e.* ~20%.

2.2.3 Codon usage

In the current scientific literature it is reported that GC3 levels can be affected by a biased codon usage, which, in turn can be affected by a biased amino acid frequency. Since by nature codon usage varies according to many factors, it is necessary to analyse this data with multivariate statistical techniques (*i.e.* Correspondence analysis, CoA). In order to make a comparison between the two ascidians genomes the CoA for each species was run on a set of orthologous genes, using Relative Synonymous Codon Usage (RSCU), see Chapter 6 for more details.

2.2.3.1 CoA analysis on *C. intestinalis* genes

The first four axes of CoA on RSCU values accounted for 11.32%, 4.96%, 4.29%, 3.83% of the total variance. The position of each sequence on the plane defined by the first two axes were presented in Figure 2.5. Highlighting ribosomal proteins a biased distribution was observed (red point in Figure 2.5). The first axis was strongly correlated with the GC3 level of each gene

($R^2 = 0.82$), as it can clearly be seen at the right side of Figure 2.5.

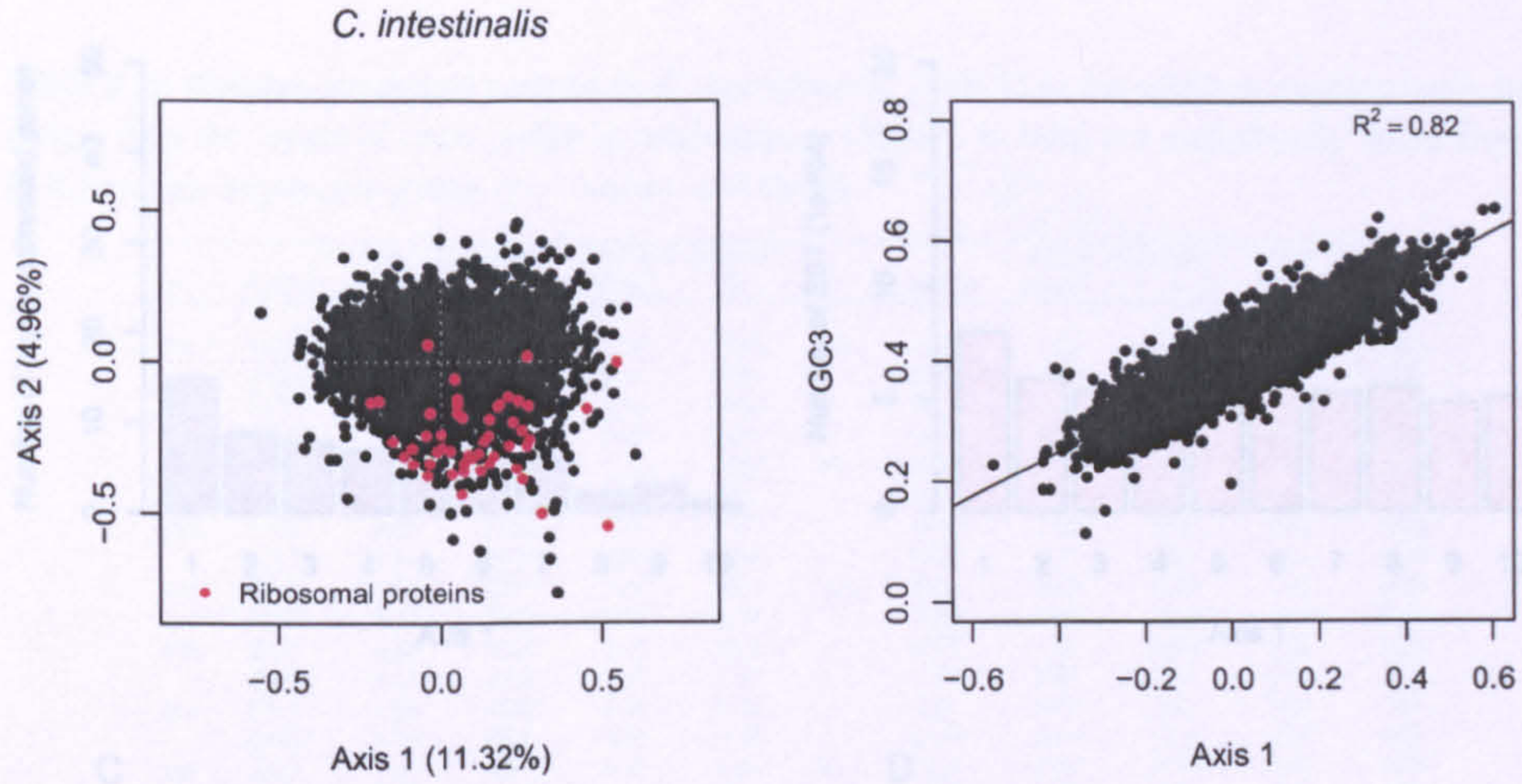


Figure 2.5: Left: Distribution of *C. intestinalis* genes on the plane defined by the two main axes of the correspondence analysis of the RSCU values, within parenthesis the amount of variance accounted for each axis. Right: Correlation between GC3 and axis1 on CoA of RSCU values

Discrimination between highly and lowly expressed genes was checked on both first and second axes. On the former, no significant discrimination was observed (Figure 2.6, A). On the contrary, sorting highly expressed housekeeping genes (*i.e.* ribosomal proteins, histones and elongation factors), according to their positions along axis2, a strong clustering in the top 10% of the distribution was found (Figure 2.6, C). More precisely, 72% of the 68 putative highly expressed proteins were found in the first decile.

In order to further support the above observation, the distribution of the ESTs (covering a total of 6218 genes) genes was checked along the same axis (Figure 2.6, panel D). In the first decile, where the majority of ribosomal proteins and other highly expressed housekeeping genes were located, genes characterized by high number of ESTs were found, confirming that axis2 of CoA on RSCU was related to the gene expression level. The total frequency of pyrimidine (Thymine + Cytosine) at the third codon position (Y) of each gene was also determined and a significant correlation was found with axis2 ($R^2 = 0.19$; $p - value < 0.0001$). It is worth to mention that the same features (i.e. GC3, gene expression, number of EST, frequency of pyrimidine) were also checked at the third and fourth axes. For these axes, which accounted for equivalent variability

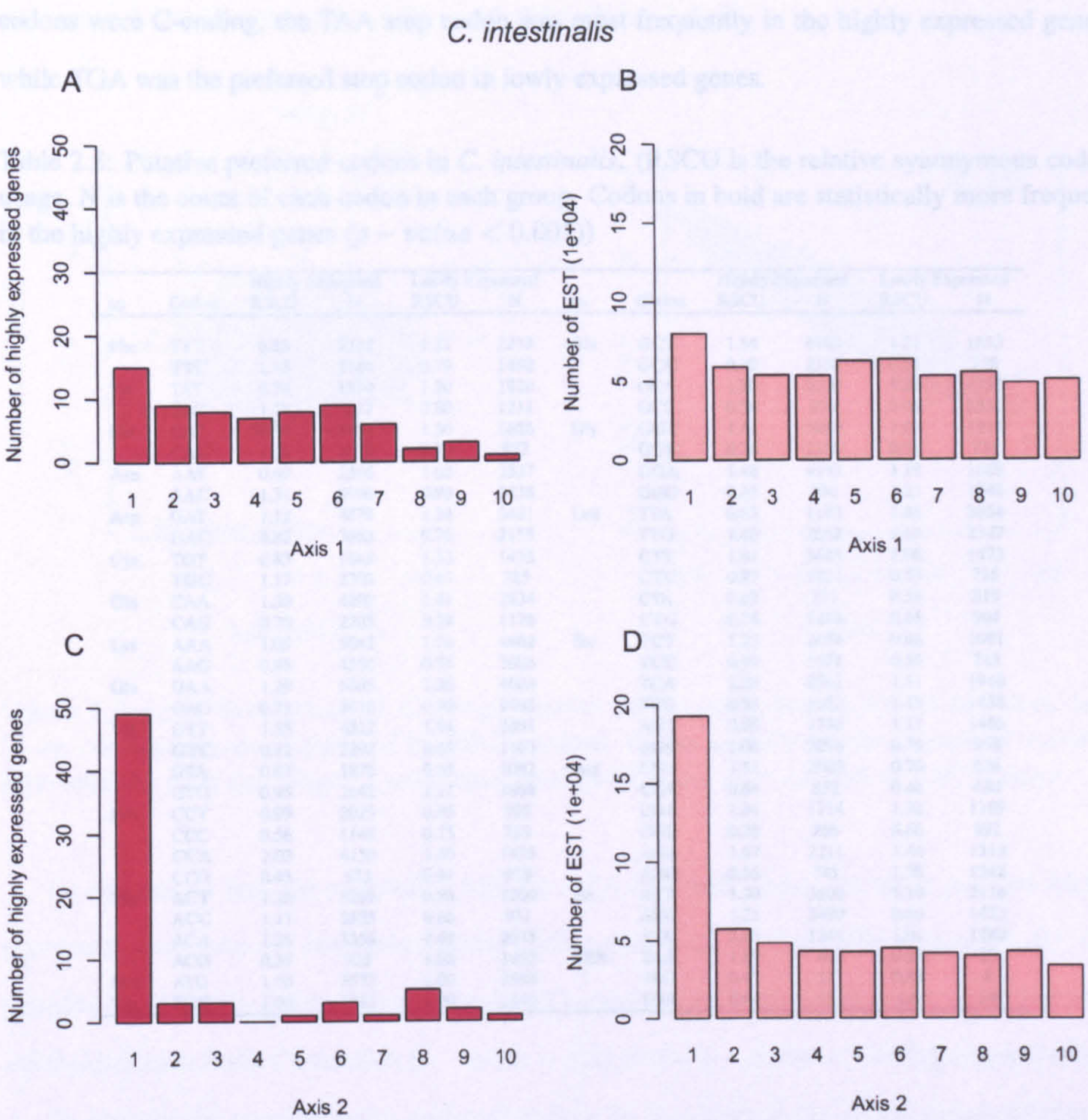


Figure 2.6: axis1 and 2 (CoA of RSCU values) was divided into 10 parts, each one containing the same number of genes of *C. intestinalis*. A: distribution of highly expressed genes along the axis1. B: distribution of expressed sequences tags (ESTs) along the axis1, C: distribution of highly expressed genes along the axis2, D: distribution of expressed sequences tags (ESTs) along the axis2

than axis2, no significant discrimination was observed for any parameter so far analysed.

The codon usage patterns of sequences clustering at both extremes of the second axis (representing respectively 5% of all data set) was compared in order to identify the translationally preferred codons. The significance of the difference was tested by a χ^2 test, and as much as 29 codons could be considered as putative preferred codons, corresponding to 17 amino acids (Table 2.5). For glutamic acid (Glu) no preferred codon was found. Interestingly, 50% of the preferred

codons were C-ending, the TAA stop codon was most frequently in the highly expressed genes, while TGA was the preferred stop codon in lowly expressed genes.

Table 2.5: Putative preferred codons in *C. intestinalis*. (RSCU is the relative synonymous codon usage, N is the count of each codon in each group. Codons in bold are statistically more frequent in the highly expressed genes ($p - value < 0.001$))

		Highly Expressed		*Lowly Expressed*				*Highly Expressed*		*Lowly Expressed*	
aa	**Codon**	**RSCU**	**N**	**RSCU**	**N**	**aa**	**Codon**	**RSCU**	**N**	**RSCU**	**N**
Phe	**TTT**	**0.85**	**2332**	**1.21**	**2273**	**Ala**	**GCT**	**1.54**	**4193**	**1.21**	**1653**
	TTC	**1.15**	**3184**	**0.79**	**1490**		**GCC**	**0.79**	**2158**	**0.55**	**758**
Tyr	**TAT**	**0.72**	**1814**	**1.20**	**1820**		**GCA**	**1.30**	**3539**	**1.28**	**1758**
	TAC	**1.28**	**3217**	**0.80**	**1211**		**GCG**	**0.36**	**973**	**0.96**	**1310**
His	**CAT**	**0.94**	**1493**	**1.30**	**1626**	**Gly**	**GGT**	**1.41**	**3984**	**1.07**	**1515**
	CAC	**1.06**	**1672**	**0.70**	**873**		**GGC**	**0.76**	**2158**	**0.51**	**731**
Asn	**AAT**	**0.69**	**2596**	**1.02**	**2537**		**GGA**	**1.48**	**4193**	**1.19**	**1688**
	AAC	**1.31**	**4906**	**0.98**	**2438**		**GGG**	**0.35**	**980**	**1.23**	**1748**
Asp	**GAT**	**1.12**	**4879**	**1.24**	**3481**	**Leu**	**TTA**	**0.63**	**1193**	**1.48**	**2054**
	GAC	**0.88**	**3863**	**0.76**	**2155**		**TTG**	**1.40**	**2652**	**1.69**	**2347**
Cys	**TGT**	**0.83**	**1648**	**1.33**	**1455**		**CTT**	**1.81**	**3443**	**1.06**	**1473**
	TGC	**1.17**	**2302**	**0.67**	**725**		**CTC**	**0.97**	**1831**	**0.52**	**715**
Gln	**CAA**	**1.30**	**4109**	**1.41**	**2834**		**CTA**	**0.42**	**791**	**0.59**	**819**
	CAG	**0.70**	**2205**	**0.59**	**1178**		**CTG**	**0.78**	**1474**	**0.65**	**904**
Lys	**AAA**	**1.05**	**5042**	**1.26**	**4462**	**Ser**	**TCT**	**1.25**	**2474**	**0.86**	**1091**
	AAG	**0.95**	**4550**	**0.74**	**2606**		**TCC**	**0.99**	**1971**	**0.56**	**713**
Glu	**GAA**	**1.29**	**6563**	**1.30**	**4604**		**TCA**	**1.29**	**2561**	**1.51**	**1918**
	GAG	**0.71**	**3618**	**0.70**	**2463**		**TCG**	**0.53**	**1052**	**1.15**	**1458**
Val	**GTT**	**1.55**	**4322**	**1.59**	**2691**		**AGT**	**0.88**	**1736**	**1.17**	**1486**
	GTC	**0.82**	**2292**	**0.65**	**1105**		**AGC**	**1.06**	**2096**	**0.76**	**968**
	GTA	**0.67**	**1876**	**0.65**	**1092**	**Arg**	**CGT**	**1.51**	**2002**	**0.70**	**626**
	GTG	**0.95**	**2642**	**1.11**	**1866**		**CGC**	**0.64**	**852**	**0.48**	**430**
Pro	**CCT**	**0.99**	**2025**	**0.95**	**998**		**CGA**	**1.34**	**1774**	**1.32**	**1189**
	CCC	**0.56**	**1148**	**0.75**	**789**		**CGG**	**0.28**	**366**	**0.66**	**592**
	CCA	**2.03**	**4150**	**1.40**	**1475**		**AGA**	**1.67**	**2211**	**1.46**	**1313**
	CCG	**0.43**	**873**	**0.91**	**959**		**AGG**	**0.56**	**745**	**1.38**	**1242**
Thr	**ACT**	**1.26**	**3299**	**0.86**	**1209**	**Ile**	**ATT**	**1.30**	**3600**	**1.19**	**2116**
	ACC	**1.11**	**2895**	**0.66**	**931**		**ATC**	**1.25**	**3480**	**0.80**	**1423**
	ACA	**1.28**	**3354**	**1.48**	**2075**		**ATA**	**0.45**	**1244**	**1.01**	**1788**
	ACG	**0.35**	**928**	**1.00**	**1402**	**TER**	**TAA**	**1.89**	**44**	**0.87**	**13**
Met	**ATG**	**1.00**	**3573**	**1.00**	**2594**		**TAG**	**0.47**	**11**	**0.53**	**8**
Trp	**TGG**	**1.00**	**1743**	**1.00**	**1149**		**TGA**	**0.64**	**15**	**1.6**	**24**

2.2.3.2 CoA analyses on *C. savignyi* genes

The first four axes of CoA on RSCU values accounted for 14.48%, 3.88%, 3.77% and 3.61% of the total variance. The position of the genes on the plane defined by the two main axes was reported in Figure 2.7. As observed in *C. intestinalis*, a biased distribution of ribosomal proteins (red points) was also found. The third and fourth axes showed no correlations with the parameter so far analyzed. On the contrary, the first axis was strongly correlated with the GC3 level ($R^2 = 0.92$), as expected (Figure 2.7, right side). Differently from *C. intestinalis*, a biased distribution of putative highly expressed genes was found on axis1 (Spearman Rank Correlation $p - value < 0.0150$), although a strong clustering was not observed (Figure 2.8, A). The observation was supported by the distribution of the ESTs of 2741 genes, significantly correlated with the first axis

(Spearman Rank Correlation $p - value < 0.0093$, Figure 2.8, B).

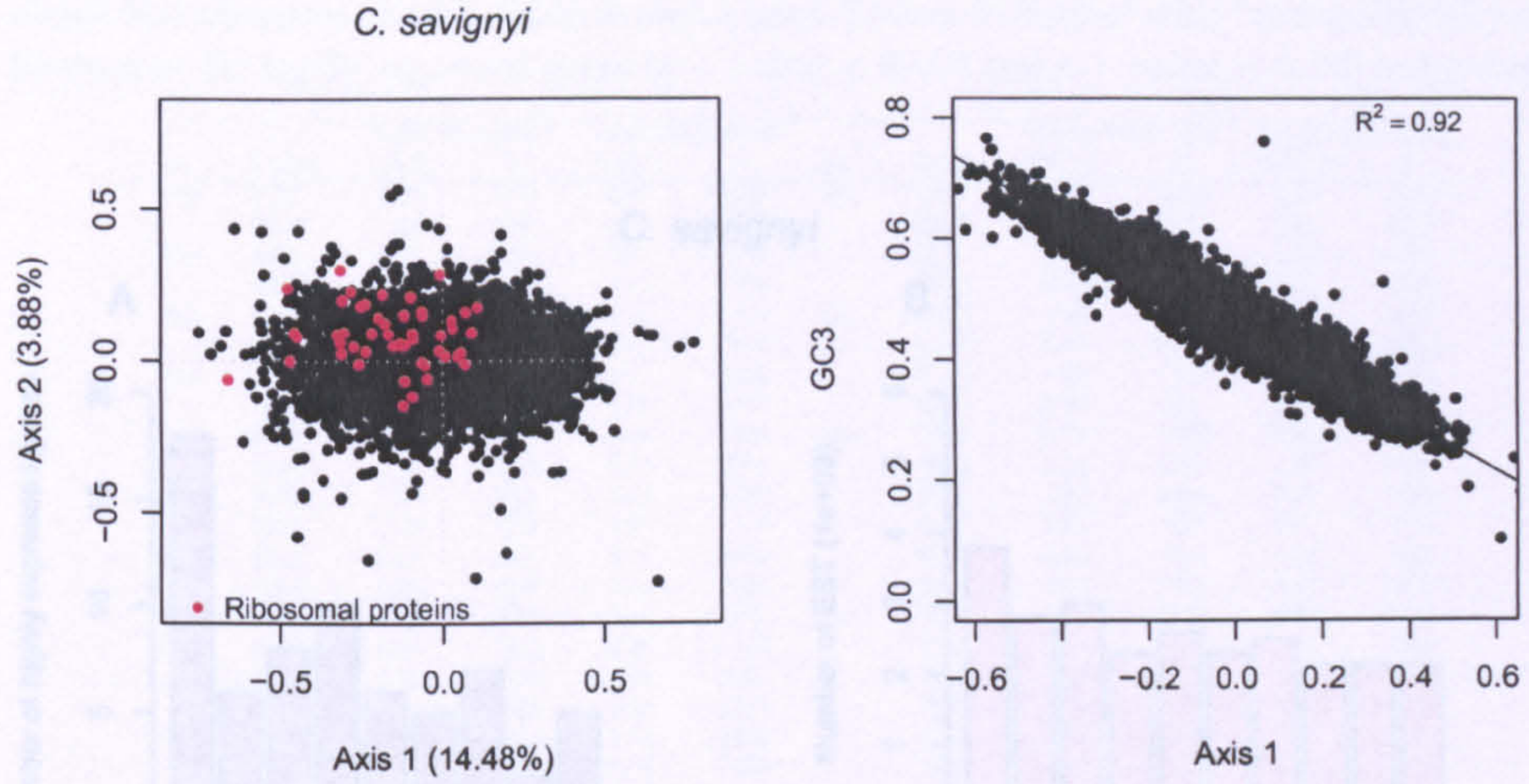


Figure 2.7: Left: Distribution of *C. savignyi* genes on the plane defined by the two main axes of the correspondence analysis of the RSCU values, within parenthesis the amount of variance accounted for each axis. Right: Correlation between GC3 and axis1 of CoA of RSCU values

On the second axis, no correlation with the GC3 level was found, but again putative highly expressed genes were not randomly distributed (Figure 2.8, C), neither ESTs (Figure 2.8, D). The distribution of highly expressed genes and ESTs values were both significantly correlated with axis2 (Spearman Rank Correlation $p - value < 0.0110$ and $p - value < 0.0104$, respectively).

To identify the translationally preferred codons in *C. savignyi*, the codon usage patterns of the sequences located at the opposite extremes of axis2 (350 genes, 5%,) were analysed. A χ^2 test was performed in order to determine statistical significances (Table 2.6). As much as 25 codons showed a frequency significantly higher among the putatively highly expressed genes. The majority of the preferred codons (40%) were C-ending, and TAA was the stop codon mostly used by highly expressed genes, as observed in *C. intestinalis*. Only three amino acids, namely phenylalanine (Phe), glutamine (Gln) and lysine (Lys) showed no preferred codons.

Finally, the delta GC3 between *C. intestinalis* and *C. savignyi* was plotted against the four axes of the CoA. Interestingly, a significant correlation was found on the first axis of *C. savignyi* ($R^2 = 0.4355$), whereas on the same axis of *C. intestinalis* no significant correlation was found ($R^2 = 0.0919$) (Figure 2.9).

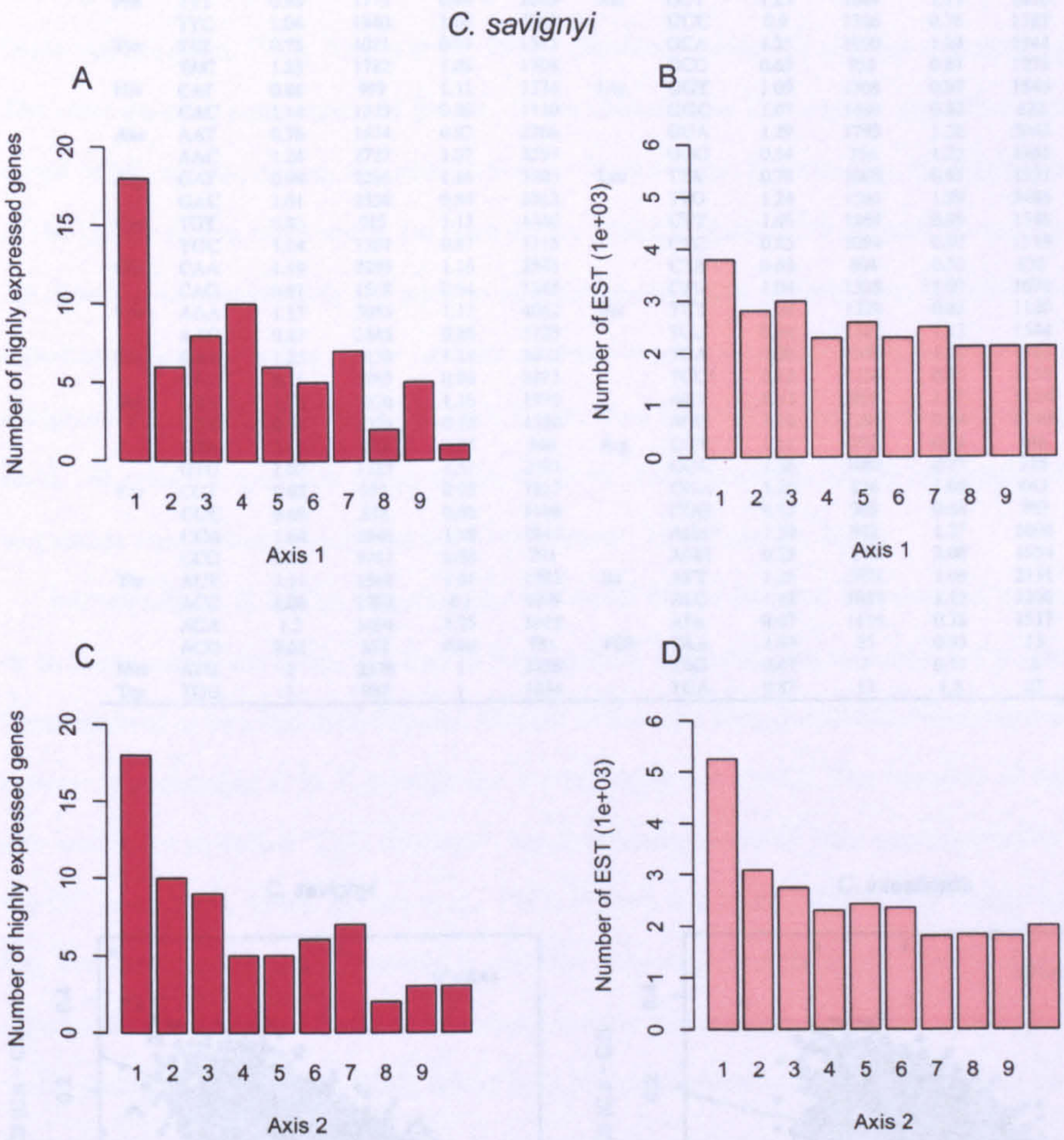


Figure 2.8: Axis 1 and 2 (CoA on RSCU values) were divided into 10 parts, each one containing the same number of genes. A: distribution of highly expressed genes along the axis1. B: distribution of expressed sequences tags (ESTs) along the axis1, C: distribution of highly expressed genes along the axis2, D: distribution of expressed sequences tags (ESTs) along the axis2

Table 2.6: Putative preferred codons in *C. savignyi*. (RSCU is the relative synonymous codon usage, N is the count of each codon in each group. Codons in bold or with * are statistically more frequent in the highly expressed genes ($p - value < 0.001$ and $p - value < 0.05$) respectively)

		Highly Expressed		Lowly Expressed				Highly Expressed		Lowly Expressed	
aa	Codon	RSCU	N	RSCU	N	aa	Codon	RSCU	N	RSCU	N
Phe	TTT	0.96	1775	0.94	2005	Ala	GCT	1.23	1869	1.19	1874
	TTC	1.04	1940	1.06	2243		GCC	0.9	1366	0.76	1187
Tyr	TAT	0.75	1071	0.94	1513		GCA	1.25	1900	1.24	1944
	TAC	1.25	1782	1.06	1704		GCG	0.63	955	0.81	1278
His	CAT	0.86	999	1.11	1374	Gly	GGT	1.09	1508	0.97	1545
	CAC	1.14	1333	0.89	1110		GGC	1.07	1488	0.52	822
Asn	AAT	0.76	1674	0.93	2266		GGA	1.29	1793	1.28	2043
	AAC	1.24	2727	1.07	2597		GGG	0.54	754	1.23	1951
Asp	GAT	0.99	2285	1.16	3205	Leu	TTA	0.78	1000	0.85	1337
	GAC	1.01	2338	0.84	2313		TTG	1.24	1593	1.59	2486
Cys	TGT	0.86	915	1.13	1446		CTT	1.46	1869	0.99	1548
	TGC	1.14	1207	0.87	1115		CTC	0.85	1094	0.97	1519
Gln	CAA	1.19	2295	1.16	2541		CTA	0.63	804	0.53	835
	CAG	0.81	1568	0.84	1845		CTG	1.04	1338	1.07	1678
Lys	AAA	1.13	3053	1.12	4012	Ser	TCT	1.02	1229	0.82	1120
	AAG	0.87	2345	0.88	3123		TCC	0.98	1181	1.13	1544
Glu	GAA	1.25	3138	1.14	3803		TCA	1.09	1310	1.29	1763
	GAG	0.75	1885	0.86	2873		TCG*	0.92	1103	0.85	1153
Val	GTT	1.28	1826	1.16	1978		AGT	0.82	985	1.07	1454
	GTC	0.98	1394	0.78	1330		AGC	1.16	1396	0.84	1149
	GTA	0.68	973	0.55	946	Arg	CGT	1.11	772	0.44	346
	GTG	1.07	1523	1.51	2593		CGC	1.56	1088	0.27	215
Pro	CCT	0.82	971	0.93	1157		CGA	1.26	876	1.06	843
	CCC	0.69	818	0.96	1198		CGG	0.43	303	0.88	703
	CCA	1.64	1946	1.48	1841		AGA*	1.39	972	1.27	1006
	CCG	0.85	1011	0.63	791		AGG	0.25	177	2.08	1654
Thr	ACT	1.11	1569	1.01	1563	Ile	ATT	1.21	2021	1.09	2131
	ACC	1.06	1502	1.1	1699		ATC	1.12	1857	1.13	2206
	ACA	1.2	1694	1.25	1919		ATA	0.67	1114	0.78	1517
	ACG	0.62	882	0.64	981	TER	TAA	1.67	25	0.93	13
Met	ATG	1	2330	1	2758		TAG	0.47	7	0.57	8
Trp	TGG	1	987	1	1334		TGA	0.87	13	1.5	21

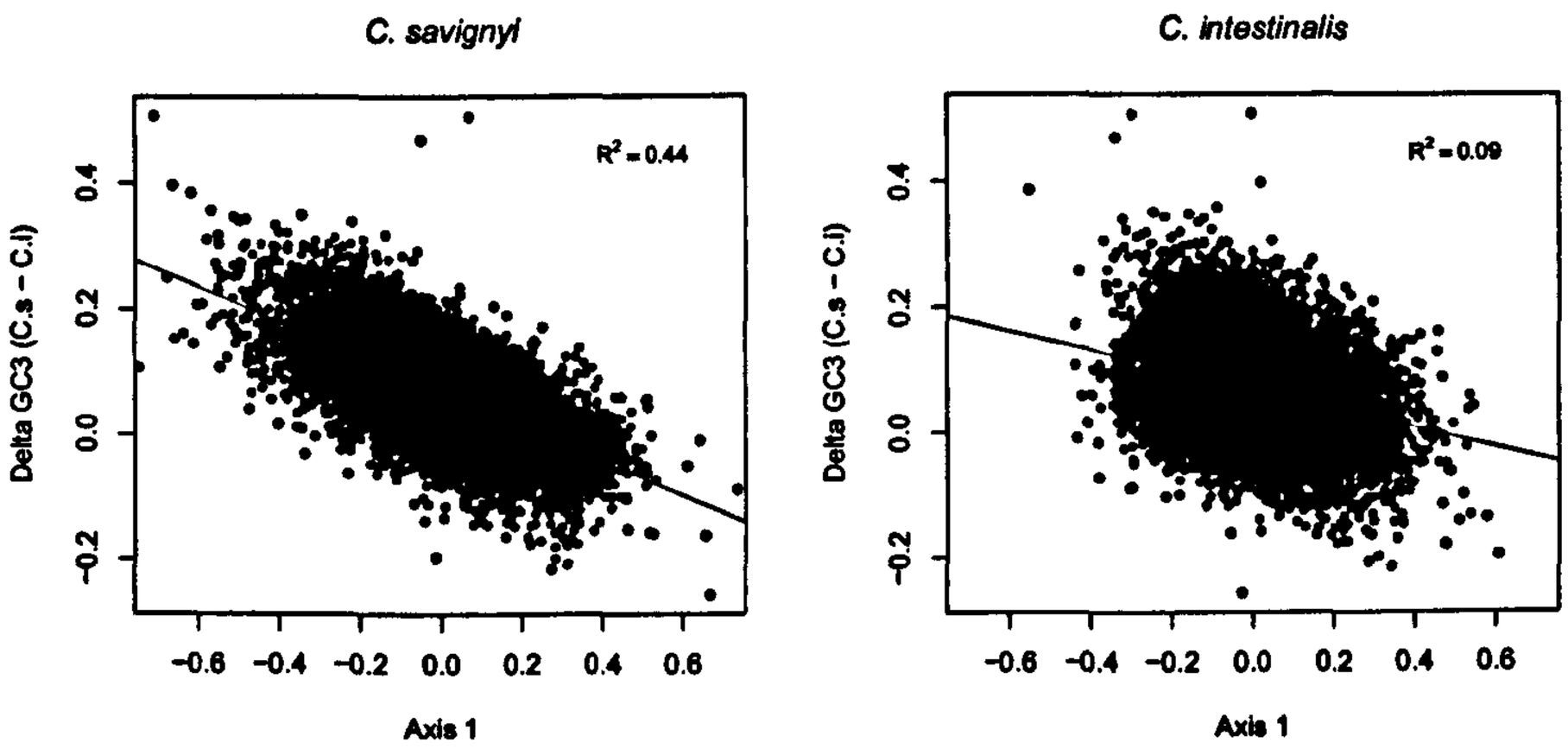


Figure 2.9: Correlation of the ΔGC3 with axis1 in *C. intestinalis* and *C. savignyi*

2.3 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 vertebrate genomes (Aota and Ikemura, 1986; Assani et al., 1991), afterwards observed in other eukaryotic genomes. The inter-genome comparison of the GC levels showed that *C. savignyi* was GC-richer, a trend found in all regions, coding and non-coding, so far analysed. Surprisingly, the highest increment of the GC level was observed at the third codon positions (about 7%, against an average of 2%). An observation confirmed by the analysis of orthologous genes. A simple explanation invoking the mutational bias, according to the intrinsic architecture of the genetic code, could not completely enlighten an increment of the GC3 levels higher than those observed in non-coding regions. Indeed, the selective pressure acting at synonymous codon positions and at intergenic/intervening sequences was generally accepted to be very similar (Graur and Li, 2000).

Intrinsically, a Δ has no directionality and could be the result of a decrement in one species, or an increment in the other, or either in both species. Regarding the decrement, CpG doublet has been reported to be a fast evolving site, because of the well described methylation/tautomerization process transforming C in T, through the 5-methylcytosine (5mC). The iteration of this process can lead to the so-called "CpG shortage", hence affecting the DNA base composition by lowering the GC level (Bird, 1980; Jabbari et al., 1997; Fryxell and Moon, 2005). The methyl-transferase has been reported to act effectively in ascidian genomes (Suzuki et al., 2007). Indeed the *C. intestinalis* genome "was found to be a stable mosaic of methylated and nonmethylated domains" being promotors, intergenic DNA, active long terminal repeat retrotransposon, long interspersed element and highly expressed genes not preferentially methylated. On the contrary, evolutionary conserved genes and other transcription units are methylated domains(Simmen et al., 1999; Suzuki et al., 2007). Therefore, the frequencies of the CG doublets, as well as that of the derivative doublets, produced by the methylation/deamination process, were checked. However, neither in coding regions where methylation occur predominantly, nor in non-coding regions any trace of CpG shortage was found. In both genomes, indeed, the derivative doublets TC and CA, showed practically the same frequencies (Tables 2.2 and 2.3). Several authors pointed out a collateral

effect (theoretically proposed, but not fully elucidated) linking CpG to TpA throughout a double 5mC, claiming a correlation with the GC content in vertebrate genomes (Bulmer, 1987; Duret and Galtier, 2000). An hypothesis not confirmed by detailed analysis of vertebrate genomes (Jabbari and Bernardi, 2004). The TpA/CpG mirror effect was also found in the ascidian genomes, but the lack of any evident CpG shortage suggested a simple AT/GC equilibrium.

Afterwards, a possible role played by the codon usage was checked. The multivariate analysis by CoA on RSCU was performed by means of orthologous sequences (>7000 cds). In *C. intestinalis* the first four axes accounted for 24.40% of the total variation, whereas in *C. savignyi* the corresponding value was slightly higher, 25.74%. A very low amount, indeed, taking into account that in Xenopus, also considered an exceptionally low case, the first axis accounted alone for 20% of the total variance (Musto et al., 2001). Plotting axis1 *versus* axis2, and highlighting ribosomal proteins, differences between the two organisms were found.

On the first axis of CoA, accounting for 11.32% and 14.48% of the total GC3 variability in *C. intestinalis* and *C. savignyi*, respectively, a strong correlation with the GC3 values was found, as expected. In *C. intestinalis* the second axis accounted for 4.96% of the total variation. After sorting the axis values and performing a partition in deciles, a strong clustering of putatively highly expressed genes and EST values were observed at one extreme (72% and 33% at the first decile, respectively). In *C. savignyi*, on the same axis, accounting for 3.88% of the total variability, such a strong clustering was not observed. However, putatively highly expressed genes and ESTs were not uniformly distributed along the second axis. Indeed at one extreme, high occurrence of putatively highly expressed genes and EST values were found (26% and 20% at the first decile, respectively). Plotting the EST values against those of the axis2, a statistically significant correlation was found, $R^2 = 0.6641$. The above results clearly showed that in *C. intestinalis* translational selection affects the synonymous codon usage. A different picture was found in *C. savignyi*, raising the question about the forces apparently slowing down selection for codon usage, as well as slowing the percent of variability accounted by the second axis (*i.e.* from 4.96% to 3.88%).

A further analysis of axis1 of the CoA revealed interesting features: i) a significant correlation with ΔGC3 in *C. savignyi* ($R^2 = 0.4355$; $p - value < 0.0001$), absent in *C. intestinalis* ($R^2 = 0.0919$); and ii) putatively highly expressed genes and ESTs values were not uniformly distributed, but significantly increasing along the axis1 in *C. savignyi*; ESTs indeed, showed a coefficient of

determination ($R^2 = 0.744$; $p - value < 0.0013$) even higher than that found on axis2 ($R^2 = 0.6641$). In turn, ΔGC3 and ESTs were significantly correlated ($R^2 = 0.741$; $p - value < 0.0014$). Incidentally, the same correlation in *C. intestinalis* was at the limit of significance ($R^2 = 0.5156$; $p - value < 0.043$).

The above result answered about the direction of ΔGC3, that indeed, was increasing in *C. savignyi*. Therefore, from a phylogenetic point of view, *C. intestinalis* is probably closer than *C. savignyi* to the common ancestor. Further comparisons of the two organisms could be of crucial interest not only to understand vertebrate evolution, but also to get deeper in the speciation processes. Several authors suggested that compositional divergences are the basic step to prevent recombination events (Forsdyke, 1996, 2007; Wei et al., 2008). Mitogenomic analysis carried out on Ciona reveled the co-existence of two cryptic species, called type A and B. Interestingly type B is significantly GC richer than type A (ΔGC= 2.1%), and the difference is indeed more evident at third positions (ΔGC3=5.1%) (Iannelli et al., 2007).

Regarding the link between GC3 and gene expression found in *C. savignyi* and only weakly present in *C. intestinalis*, a question arises about the cause/effect between the two variables. The issue will be discussed in the frame of current theories about genome evolution.

The current neutral and selective hypotheses proposed to explain the base composition variation at the genome level were presented in the section "General Introduction", namely, the bias gene conversion (BGC) (Eyre-Walker, 1993; Galtier et al., 2001; Duret and Galtier, 2009), the breakpoints distribution (Lemaitre et al., 2009); the thermal stability (Bernardi (2007) for a review), and the metabolic rate (Vinogradov, 2001, 2005). Which of the above hypotheses could better fit with the data obtained from the compositional analysis of the ascidian genomes?

Regarding BGC, it is generally recognized that recombination plays a key role in genome evolution, and within a genome a correlation holds with GC content, as shown for example in human and yeast (Eyre-Walker, 1993; Charlesworth, 1994; Baudat and Nicolas, 1997; Gerton et al., 2000). More debated is the cause/effect relationship. Indeed, in 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). In mammals, recombination hotspots sites were not conserved from a phylogenetic point of view, in spite of a very close isochore pattern, like

in the human/chimpanzees genome comparison (Ptak et al., 2005). Moreover, little correlation between the two variables was found in a pseudoautosomal region (PAR) and in X-linked regions in human/orangutan genome comparison (Huang2005).

In ascidian genomes the recombination rate were reported to be 25-49 kb/cM in *C. intestinalis* (Kano et al., 2006) and 200 kb/cM in *C. savignyi* (Hill et al., 2008). As far as we know, no comparative studies have been carried out to assess a correlation between GC content and recombination rate at inter-genomic level. However, according to the correlation between the two variables, *C. intestinalis* should expected to be GC-richer than *C. savignyi*.

Regarding DNA breakpoints, it is difficult to assess if the hypothesis could fit with the ascidian genomes, since in these species little is known: i) about the distribution of fragile points along the chromosomes; and ii) about the bias of the replication/repair machinery. However, Lemaitre and colleagues (2009) stated: "In agreement with the Intergenic Breakage Model, we observed that breakpoints are under-represented in genes". Consequently, according to this observation, a higher GC delta should have been found not at the third codon position, as it was, but more in non-coding regions.

Regarding the thermal stability, a bibliography search about the ecological distribution of both Ciona species was done. The geographical origins of *C. intestinalis* and *C. savignyi* are still matter of debate. Most probably *C. intestinalis* was originally native from northern Europe (Monniot and Monniot, 1994) and *C. savignyi* from Japan (Lambert and Lambert, 1998). While *C. savignyi* is apparently restricted to North Pacific area, *C. intestinalis* is extremely widely distributed and considered as cosmopolitan species, 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 (Zvyagintsev et al., 2007). *C. intestinalis* tolerate temperatures between 1 °C and 30 °C, representing the maximum range from different populations around the globe and it is considered a cold-water or temperate species (Therriault and Herborg, 2008). *C. savignyi* is also considered a cold-water organism, and share with *C. intestinalis* the same range of optimal developmental temperatures (Takashi et al., 1997; Bellas et al., 2003). However, they have been found only in North Pacific water, possibly due to the fact that optimal range of their temperatures is narrower compared with *C. intestinalis*. In short, both Ciona species could colonize the same habitats, but *C. intestinalis* shows to be more flexible and to tolerate

different environments, specifically those that correspond to higher temperatures (e.g. Mediterranean Sea and Atlantic sea). Consequently, according to the thermal stability *C. intestinalis* was expected to show a higher or, at least equal, GC content than *C. savignyi*.

According to metabolic rate hypothesis, DNA structure shows different degree of flexibility being more bendable at higher GC levels (Vinogradov, 2001). Moreover, along the same line, nucleosome formation potential was found to be negatively correlated with the GC content (Vinogradov, 2005). Therefore, both properties converged towards the suggestion that GC-poor and GC-rich regions should have a specific chromatin structure, dictated by the trascriptional level. By *in situ* hybridization of GC-poor and GC-rich probes, a "closed" and an "open" chromatin structure was found in GC-poor and GC-rich chromosomal regions (Saccone et al., 2002).

At present, among the hypotheses discussed above, only the metabolic rate could provide a solid background to give an adequate answer to the question left open about the cause/effect of the co-occurrence on the first axis of CoA in *C. savignyi* of a correlation with ΔGC3 and ESTs values, not found in *C. intestinalis*. In this context, in spite of the difficulties for the sampling of *C. savignyi* from Pacific ocean, the measurement of the O_2 consumption is now in progress, and would be of great support to these results.

2.4 Conclusions

A genome comparative analysis of *C. intestinalis* and *C. savignyi* was done, leading to several main conclusions: i) a compositional difference exist among the two organisms, taking place mainly at the third codon position; ii) the difference are due to an increment that took place in *C. savignyi*, consequently *C. intestinalis* most probably is closer to the ancestral species; iii) both Cionas showed a selection for codon usage; and iv) the metabolic rate is at present is one hypothesis that could give an understandable theoretical frame for the differences observed between the two species.

3 | *High rate of evolution in tunicates*

3.1 Introduction

Because of a remarkable morphological similarity, *Ciona intestinalis* and *Ciona savignyi* have been long regarded as close relatives. Nevertheless, more recently, several evidences supported the notion that, in spite of their similar morphology, *C. intestinalis* and *C. savignyi* genomes are quite divergent. It was not very clear however, whether this genomic differentiation was due to an accelerated evolutionary genomic rate, to long divergence time, or a combination of both factors. Supporting the second alternative, high rates of molecular evolution have been reported in previous studies focusing on individual genes, such as *huntingtin* gene (Gissi et al., 2006), 18S rRNA (Johnson et al., 2004) and P-transposase (Kimbacher et al., 2009). However, neither their precise speed of molecular evolution, nor the time of divergence between the two Ciona species were known. Thanks to the fact that the genomes of *C. intestinalis* and *C. savignyi* have been completely sequenced, we were able to make an accurate assessment on the whole genome divergence between the two species. Furthermore, the availability of genome data from two other deuterostomes recently published, namely sea urchin (Sodergren et al., 2006) and amphioxus (Putnam et al., 2008), was of a great importance, because it allowed us to use these species as outgroup and hence to assess the relative speed of genome evolution in ascidians in comparison to vertebrates.

The high rate of evolution in Ciona appeared to be a more general feature affecting others, if not all, tunicates. Certainly, high substitution rates have been observed in other tunicates like *Oikopleura dioica* (Delsuc et al., 2006; Winchell et al., 2002). Particularly surprising was the very long branch exhibited by this larvacean species on a Bayesian phylogenetic tree of deuterostomes (Putnam et al., 2008) built on an alignment of 1090 concatenated genes. Remarkably, the phy-

logenetic tree was not only confirmatory that tunicates are, indeed, a sister group of vertebrates, but also allowed to observe a roughly constant evolutionary rate of amino acid changes across deuterostomes but a noticeable increment in tunicates (Putnam et al., 2008). In addition, different works reported recurrent problems: that is the long branches of tunicates, an observation that again stressed their high rate of evolution (Blair and Hedges, 2005; Bourlat et al., 2006; Tsagkogeorga et al., 2009). What is more, analyses of mitochondrial genomes also show that tunicates are fast evolving species compared with other metazoan species (Singh et al., 2009). Specifically in this latter work the authors estimated the evolutionary rates of rRNA and mitochondrial protein-coding genes of 54 taxa, and demonstrated that for both groups of sequences, tunicates evolve faster.

The aforementioned apparent exacerbated acceleration in the rates of amino acid substitutions in tunicates, and specially in *Oikopleura dioica*, is a subject that in our opinion deserves further attention for many reasons. In the first place, it would be of interest to analyze to what extent such an extreme increment affects the Oikopleura genome, namely whether it is restricted to a particular group of genes or if it is a genome wide phenomenon. In the second place, *O. dioica* represents a particularly interesting species for the understanding of what seems to be a general feature of all tunicates, their fast evolutionary rates, because it belongs to the Appendicularia class (larvaceans), namely this species is derived from the most basal tunicates.

In the present work we characterize the rate of molecular divergence of Ciona species. Moreover we analyze the rates and patterns of amino acid substitutions of tunicates with special interest in the exceptionally fast evolving *O. dioica* aiming to shed some light on the biological causes that led to this peculiar rate of divergence. These analyses were conducted at the genome scale level, which means using a gene sample large enough to be fully representative of the genomes they belong to. These kinds of analyses were possible because during the last few years the genome of several chordates have been published, and vast amounts of sequence data (at the genome scale) is available for many other chordates.

Parts of the results presented here have been published in:

Bernà, L., Alvarez-Valin, F., and D'Onofrio, G. (2009). *How fast is the sessile ciona?* Comp. Funct. Genomics. doi:10.1155/2009/875901.

3.2 Results

To estimate the time of divergence between *C. intestinalis* and *C. savignyi*, the approach is to compare their genomic distances with the distances exhibited by other organisms (mostly vertebrates) for which the dates of divergence are known based on the fossil record. Then, based on this information and their relative evolutionary speed, a rough estimate can be obtained. In other words, we calibrate a vertebrate's molecular clock, compare the Ciona and vertebrate speed (by using the relative rate test) and extrapolate the time of divergence between Cionas, by comparing the two group of distances.

To this aim, orthologous gene pairs of nine couples of vertebrates were obtained. The pairs were constructed with human and another species representative of different vertebrate groups encompassing a broad range of divergence times (mouse, opossum, ornithorhynchus, cow, chicken, frog and fugu), also, gene pairs of *Xenopus laevis* and *X. tropicalis*, as well as *Mus musculus* and *Rattus norvergicus* were used. The genome-wide amino acid inter-species distances among these vertebrates were calculated, as well as between the two Ciona species (for details see Chapter 6). The average amino acid distances (and their standard errors), the number of orthologous pairs used for each comparison and the time estimations obtained from fossil records are presented in Table 3.1.

3.2.1 Distance between Ciona and vertebrates

Vertebrate's molecular clock In order to determine if the distances calculated for vertebrates and the time of divergence satisfy a linear relationship (in other words if they follow a molecular clock), a graphical representation of the relationship between genomic divergence and time of divergence was made (Figure 3.1). A clear linear relationship (*i.e.* clock like) was found. However, two aspects deserve to be pointed out. First, contrary to what would be expected, the regression line does not pass through the origin (time equal zero should correspond to zero distance). This can be attributed to the fact that some of the orthologous pairs could be, most probably, paralogous. Second, the distances between human and marsupials, and human and platypus were higher than expected (points 5 and 6 in Figure 3.1). This can be readily explained by a higher rate of molecular evolution in non-placental mammals. To better observe this effect, opossum and or-

	N° sequences		N° sequences	BRH	Distances JJT	MIN (My)	MAX (My)
Mus musculus	34.966	*Rattus norvergicus*	36.496	16.138	0.077	11	12,3
Xenopus tropicalis	20.455	*Xenopus laevis*	30.455	6.613	0.097	48	52
Homo sapiens	34.180	*Bos taurus*	24.853	12.755	0.152	95,3	113
		Mus musculus	34.966	16.848	0.174	61,5	100,5
		Monodelphis domestica	32.557	15.145	0.261	124,6	138,4
		Ornithorhynchus anatinus	26.836	11.323	0.308	162,5	191,1
		Gallus gallus	18.529	8.380	0.326	312,3	330,4
		Xenopus laevis	30.455	7.813	0.373	330,4	350,1
		Takifugu rubripes	47.841	10.975	0.471	416,0	486,0
Ciona intestinalis	19.858	*Ciona savignyi*	20.143	7.815	0.335	?	?

Table 3.1: Genome wide average amino acid distances, their standard error, and the time of divergence in Million years (My) for each pair of vertebrate species analysed

nithorhynchus were removed from the analysis (dashed line). As a result, all points fitted almost perfectly the regression line, incidentally this result gives further support to the claim that the rate of molecular evolution has been accelerated in non-placental mammals.

Divergence estimation Assuming that the genomes of ascidians have the same evolutionary pace as the remaining chordates, without considering non-placental mammals, the estimation of the divergence time between the Ciona species would be approximately 308 (±16) Million years (My). Considering the close morphology between these ascidians, this estimation is surprisingly higher than one would have expected. However, taking into account the discovery done in China some years ago (Chen et al., 2003), where an Early Cambrian fossil from namely aplousobranch tunicate group showed an overall morphological aspect very similar to that of living species, which means that for more than 500 My the overall morphological features of ascidians were hardly modified, the aforementioned estimation does not seem to be too exaggerated. Needless to say, the estimation of divergence time of 308 My was obtained under the assumption that the same

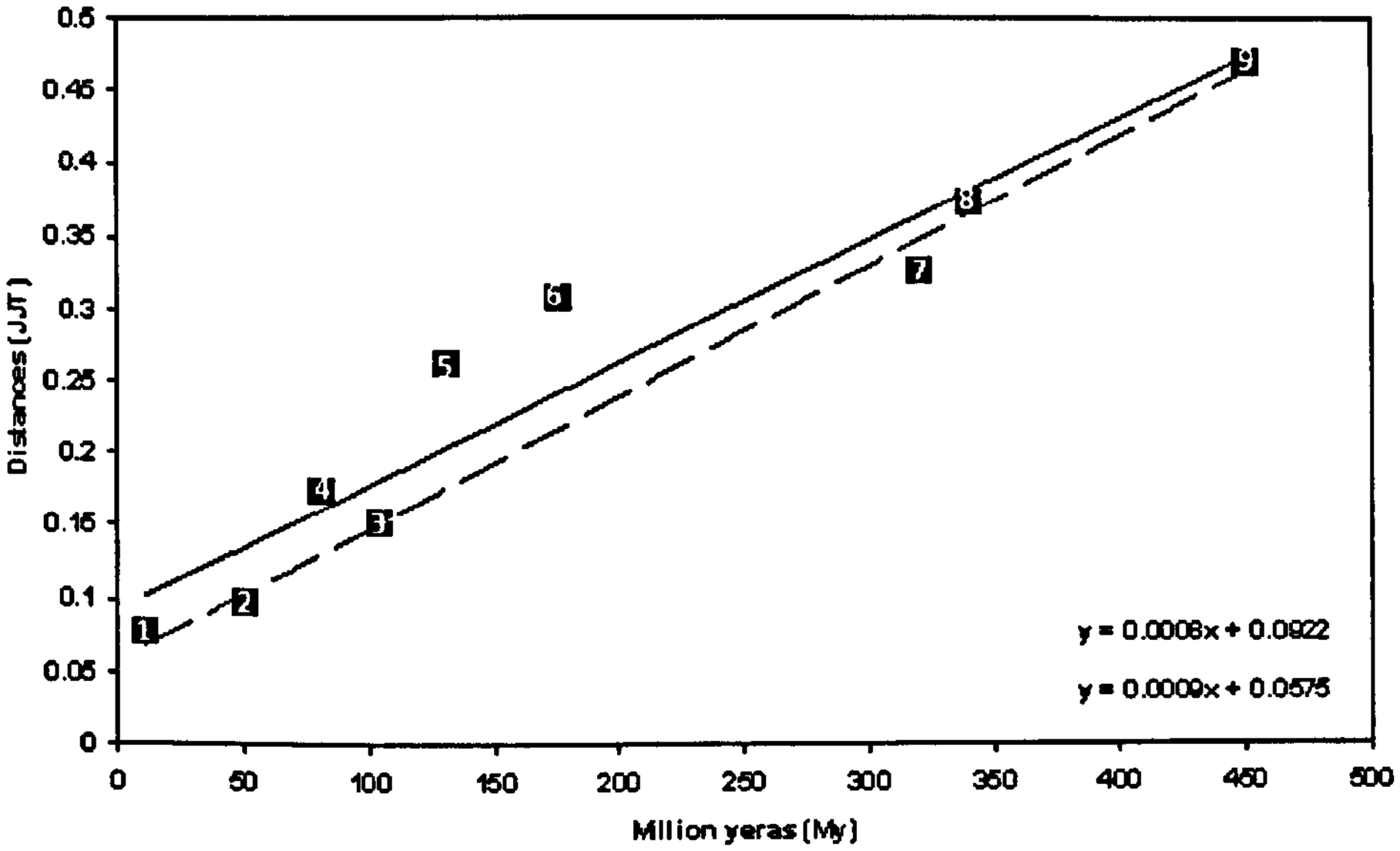


Figure 3.1: Regression line of divergence times derived from the fossil record (Table 1) and the correspondent distances (JTT method) in different vertebrate pairs. For the continuous line: 1. *M. musculus* - *R. norvergicus*, 2. *X. laevis* - *X. tropicalis*, 3. *H.sapiens* - *B. taurus*, 4. *H. sapiens* - *M. musculus*, 5. *H. sapiens* - *M. domestica*, 6. *H. sapiens* - *O. anatinus*, 7. *H. sapiens* - *G. gallus*, 8. *H. sapiens* - *X. laevis*, 9. *H. sapiens* - *T. rubripes*. For the dashed line: same pairs, comparisons involving non-placental mammals were excluded (points 5 and 6)

molecular clock governs Ciona and vertebrates. However, as mentioned phylogenetic analyses of single (Gissi et al., 2006) or large set of nuclear genes (Bourlat et al., 2006; Delsuc et al., 2008; Gissi et al., 2006; Putnam et al., 2008), as well as mitochondrial ones (Bourlat et al., 2006), have suggested that the ascidians genomes evolve faster than vertebrates. Therefore, the assumption that ascidians and remaining chordates have the same evolutionary pace should be carefully re-considered. Several very serious inaccuracies could probably affect the reported determination, specifically the time of divergence was very probably an overestimate. Therefore, to test the validity of this assumption and also to determine the acceleration extend, if any, a genome wide analysis of substitution rates was carried out.

Relative rate test In order to verify if vertebrates and ascidians have the same molecular clock, the Tajima's relative rate test was conducted. For this purpose, orthologous sequences shared by

C. intestinalis, one vertebrate and one outgroup were analysed. The cephalochordate *B. floridae*(amphioxus), whose genome has been recently published (Putnam et al., 2008), was used as the outgroup. The choice was largely justified by the new phylogeny of chordate organisms, according to which tunicates, instead of cephalochordate, are the closest relatives of vertebrates, as previously accepted (Delsuc et al., 2006, 2008; Dunn et al., 2008; Putnam et al., 2008). However, an unsolved controversy still runs about this specific phylogenetic item, therefore the same analysis was also performed using as the outgroup the *S. purpuratus* (sea urchin), whose status is not questioned (Bourlat et al., 2006; Philippe and Telford, 2006; Stach, 2008). Besides, the comparisons of ascidians with several vertebrates were performed, in order to make sure that the differences between Ciona and vertebrates were not comparison-specific, but reflecting a more general aspect of chordate evolution. Species from main vertebrate classes with the exception of reptiles, due to the lack the genome data, were used. As clearly emerged from Tajima's test results (Table 3.2), for the vast majority of the alignments, Ciona genes evolve faster than those of all vertebrate groups, including those of the non-placental mammals, which, as discussed above, were characterized by the faster evolutionary rate among vertebrates. Indeed, the proportions of genes that evolve faster in Ciona ranged from 67% to 83%, according to the different vertebrate groups. Furthermore, considering only those alignments that yielded a significant χ^2 value, the proportion of genes that are faster in Ciona increased up to 85%. The results remained basically unchanged using sea urchin as outgroup (Table 3.2). Thus, the evolutionary rate of Ciona at molecular level turned out to be faster than of all other vertebrate species analysed.

Species name	N orthologs trios	D13>D23 (%)	Total Significant	Significant at 1%	Significant at 5%	D13> D23 Significant*
2 - *Homo sapiens*	5911	4603 (77.9%)	3438	2736	702	2998 (87.2%)
2 - *Mus musculus*	5862	4519 (77.1%)	3422	2691	731	2949 (86.2%)
2 - *Xenopus laevis*	4452	3410 (76.6%)	2587	2012	575	2180 (84.3%)
2 - *Ornithorhynchus anatinus*	5392	3605 (66.9%)	2941	2263	678	2158 (73.4%)
2 - *Monodelphis domestica*	5928	4468 (75.4%)	3408	2653	755	2904 (85.2%)
2 - *Bos taurus*	4864	3906 (80.3%)	3075	2493	582	2704 (87.93)
2 - *Gallus gallus*	5221	3975 (76.1%)	3131	2466	665	2640 (84.3%)
2 - *Takifugu rubripes*	4918	4080 (83.0%)	3064	2464	600	2797 (91.3%)
Total	**42548**	**32566 (76.5%)**	**25066**	**19778**	**5288**	**21330 (85.1%)**
1 - *Ciona intestinalis*						
3 - *Branchiostoma floridae* (Amphioxus)						

Species name	N ortholog trios	D13>D23 (%)	Total Significant	Significant at 1%	Significant at 5%	D13> D23 Significant*
2 - *Homo sapiens*	4662	3720 (79.8%)	2485	1873	612	2232 (89.8%)
2 - *Mus musculus*	4659	3659 (78.5%)	2449	1880	569	2149 (87.8%)
2 - *Xenopus laevis*	3670	2847 (77.6%)	1936	1432	504	1683 (86.9%)
2 - *Ornithorhynchus anatinus*	4093	2322 (56.7%)	2063	1563	500	1363 (66.1%)
2 - *Monodelphis domestica*	4623	3570 (77.2%)	2407	1812	595	2076 (86.2%)
2 - *Bos taurus*	4431	3400 (76.3%)	2321	1714	607	1985 (85.5%)
2 - *Gallus gallus*	4166	3261 (78.3%)	2243	1703	540	1936 (86.3%)
2 - *Takifugu rubripes*	4457	3529 (79.2%)	2310	1716	594	2048 (88.7%)
Total	**34761**	**26308 (75.7%)**	**18214**	**13693**	**4521**	**15472 (84.9%)**
1 - *Ciona intestinalis*						
3 - *Strongylocentrotus purpuratus* (Sea Urchin)						

Table 3.2: The Tajima's Relative Rate Test, orthologous sequences from 1- *C. intestinalis*, 2- one vertebrate and 3- an outgroup (Amphioxus above, Sea Urchin below). (*) Related to the total of significant differences.

3.2.2 Acceleration in Ciona

In order to be able to use the calibration of vertebrate molecular clock presented in Figure 3.1, it becomes also necessary to determine the extent of acceleration in Ciona molecular clock. The simplest approach was to estimate the branch lengths separating both Ciona and vertebrates from their common antecessor (Sarich and Wilson, 1973). The results showing the acceleration of rates between *C. intestinalis*, versus human, mouse, frog, cow, chicken and fish, using amphioxus as an outgroup are reported in Table 3.3. The distance **a**, between *C. intestinalis* and the common ancestor, was always greater than the distance **b**, between each vertebrate, by a factor of 1.50 (for more details see 6.9, Figure 6.1). That is, on the average, *C. intestinalis* evolves 50% faster than all vertebrates, with the exception of *O. anatinus* and *M. domestica* (for which the acceleration is slightly lower).

	d12 JTT	d13 JTT	d23 JTT	**a**	**b**	(a-b)/b
2 - *Homo sapiens*	0.659	0.6501	0.5138	0.3976	0.2613	0.5217
2 - *Mus musculus*	0.6608	0.6472	0.5195	0.3942	0.2666	0.4789
2 - *Xenopus laevis*	0.6377	0.6316	0.4917	0.3888	0.2489	0.5621
2 - *Bos taurus*	0.6638	0.6471	0.5193	0.3958	0.268	0.4769
2 - *Gallus gallus*	0.6517	0.6439	0.5082	0.3937	0.258	0.5258
2 - *Takifugu rubripes*	0.6684	0.6497	0.5273	0.3954	0.273	0.4486
1 - *Ciona intestinalis*					Average:	**0.5023**
3 - *Branchiostoma floridae* (Amphioxus)						

Table 3.3: Average distance of orthologous sequences from 1- *C. intestinalis*, 2- one vertebrate and 3- *B. floridae*. **a** - correspond to the distance between *C. intestinalis* and the common ancestor with vertebrate, and **b** is the distance between vertebrate and the common ancestor with *C. intestinalis*.

Nevertheless, it is worth bringing to mind that the previous analysis was performed with *C. intestinalis* genome. Thus, it is necessary to determine if it is possible to expand this result for *C. savignyi*. In other words, determine if both ascidians evolve at the same speed, nearly 1.5 faster than vertebrates, or on the other hand, if this peculiar acceleration is only related to *C. intestinalis* genome. This was accomplished by comparing 5320 orthologs between *C. intestinalis*, *C. savignyi* and *B. floridae*. Each trio of orthologous sequences was aligned and the relative distances between both ascidians and the outgroup amphioxus were calculated. The average distance of *C. intestinalis* and *C. savignyi* to *B. floridae* were 1.001 and 1.004 respectively and there was no

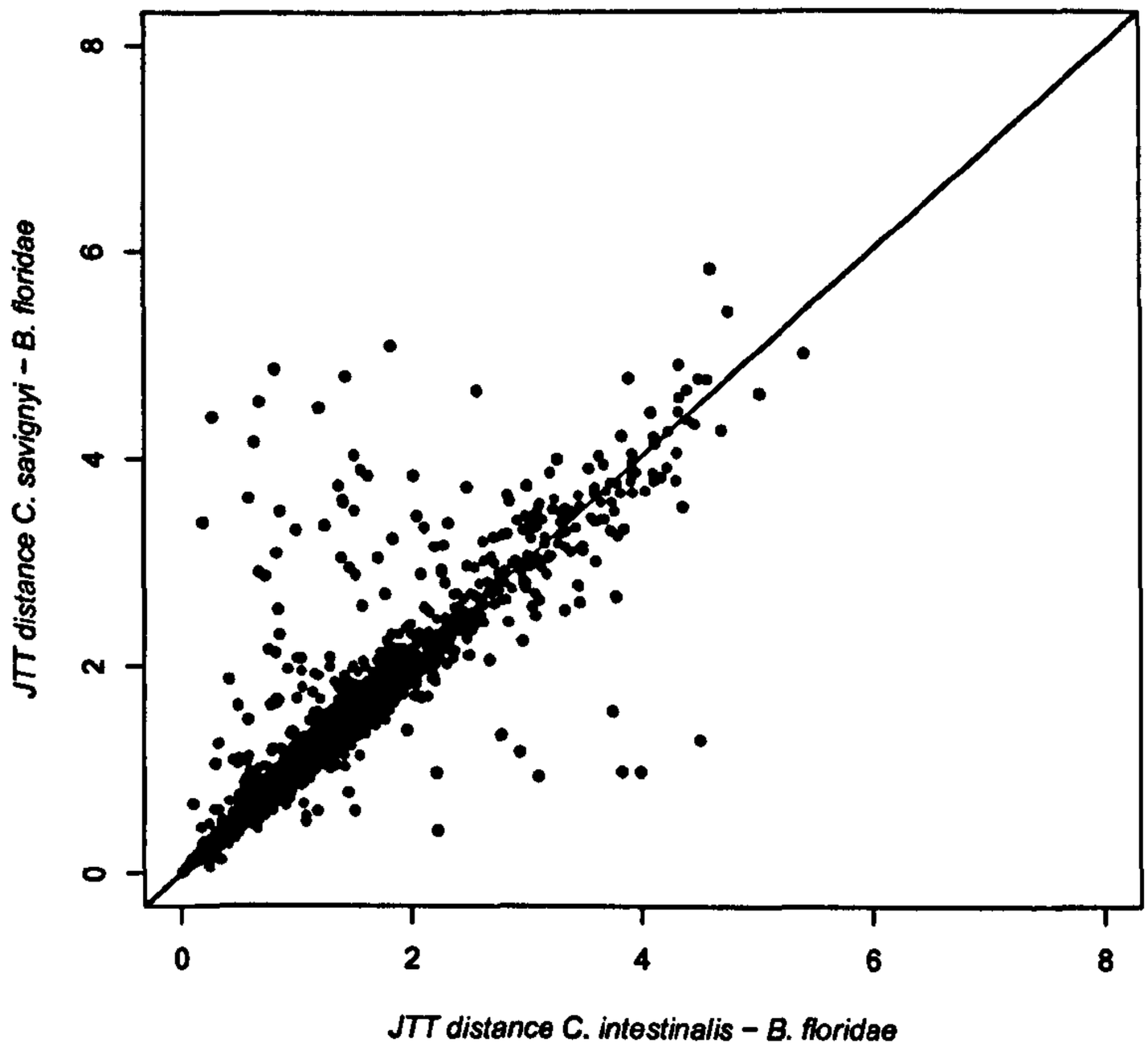


Figure 3.2: Scatter plot of JTT distances between the ortologous pairs of *B. floridae* (as an out-group) with *C. intestinalis* and *C. savignyi*

statistically significant difference between them (t-test for paired comparisons). This result shows clearly that both Ciona genomes evolve at approximately the same rate. This can also be readily observed graphically, indeed Figure 3.2 shows how the distances of *C. intestinalis* and *C. savignyi* to amphioxus are homogeneously dispersed around the diagonal line, which indicates equal molecular evolutionary rate. Moreover, the Tajima's relative rate test for each group of orthologs was also performed, and as expected, not significant differences were obtained between the rate of evolution of these two ascidians.

Divergence estimation Taking in consideration the evolutionary rate of Ciona species, and the calibration of vertebrate's molecular clock (Figure 3.1), the divergence between *C. intestinalis* and *C. savignyi* was re-estimated and found to have taken place approximately 184 ($\pm$15) My ago. This new estimation, considerably lower than the previous one obtained without adjusting

the relative molecular clocks, seems to be still large for two species almost indistinguishable one to the other. However taking in consideration the close morphology in living tunicates and specimens of an Early Cambrian fossil tunicate (Hoshino and Nishikawa, 1985; Chen et al., 2003) this divergence estimation appears to be very realistic.

3.2.3 Evolutionary rates in tunicates

In order to determine if the this acceleration present in Ciona species is, indeed, a characteristic of tunicates genomes, the evolutionary pace and substitution patterns were studied in *O. dioica*, a species that belongs to the Appendicularia class (larvaceans), namely this species is derived from the most basal tunicates (Swalla et al., 2000). These analyses were conducted using a data set comprising 3221 groups of orthologous genes. These groups consist of genes from the tunicate species *O. dioica* and *C. intestinalis*, and from two outgroup species: the vertebrate *Bos taurus* and the cephalochordate amphioxus (**Branchiostoma floridae**). Because of the very high rates that tunicates and in particular *O. dioica* seem to exhibit, the orthology assignment was difficult in some cases. Taking this fact into account, a subset of more reliable orthologous groups was analysed separately. This subset of orthologs contains alignments with less than 60% of indels and distances lower than 1 amino acid substitutions per site (for details see materials and methods). The restricted subset was used to verify if the conclusions drawn from the whole data set also hold when less divergent and more accurate alignments were used.

Table 3.4: Molecular JTT distances: 1- *O. dioica*, 2- *C. intestinalis*, 3- *B. taurus*, 4- *B. floridae*

	#	d12	d13	d14	d23	d24	d34
All	3220	1.281	1.328	1.320	0.809	0.789	0.657
Filtered	1610	0.975	0.994	0.992	0.573	0.563	0.46
Very slow	176	0.283	0.325	0.308	0.24	0.223	0.206
Slow	517	0.557	0.621	0.623	0.443	0.442	0.391
Intermediate	415	0.798	0.847	0.851	0.582	0.582	0.480
Fast	2112	1.637	1.678	1.667	0.990	0.961	0.795

Distances averaged across all groups of orthologous genes, between each pair of organisms, as well as those corresponding to the more reliable subset are present in Table 3.4. As it can be

clearly observed, the distances between *O. dioica* and either outgroup (vertebrate or cephalochordate), namely d13 and d14, are much higher, almost twice as much, than those between *C. intestinalis* and the same outgroup (d23 and d24). The same relation also holds when only the dataset containing the more reliable group or orthologs is considered (filtered). It is worth stressing that this difference in average distances is consistent along the whole dataset and cannot be attributed to a minority of genes that are very fast evolving in *O. dioica*, as it emerges from the pairwise comparisons of distances between *O. dioica/B. taurus* and *C. intestinalis/B. taurus*. Specifically out of 3220 COGs used in the analysis, *O. dioica* genes are faster in 3060 COGs, which represents 95% of the dataset, being this trend highly statistically significant (t-test for paired comparisons, $p - value << 2.2e^{-16}$).

Acceleration in *O. dioica* The acceleration that *O. dioica* genes underwent in relation to those of the *C. intestinalis* is more evidently appreciated in Figure 3.3 that shows the scatter-plot of the distances between of *O. dioica* and *C. intestinalis* to the same outgroup (*B. taurus*). Note, that almost all points fall well above the diagonal line, which represents the place where points should fall if both species of tunicates had the same molecular evolutionary rate. Almost identical results were obtained using amphioxus as the outgroup or any other vertebrates instead of *B. taurus* (SupFig. 2).

In order to quantify the differences in evolutionary rates between *O. dioica* and *C. intestinalis* the Tajima's relative rate test for each group of orthologs was carried out. This allows to determinate which genes present statically significant acceleration in any of the species. As already mentioned 95% of the genes evolve faster in *O. dioica* than *C. intestinalis*. Remarkably, 2486 genes (which represent 81.2% of the sample) have a statistically significant difference (χ^2 in Tajima's test). The results remained basically the same when different groups of orthologous sequences (i.e, using a different outgroup) were utilized.

It is worth wondering what the extent of this acceleration is, in other words how much faster is *O. dioica* in comparison to Ciona. It is important to remind that Ciona is already a fast evolving species, being 1.5 faster than vertebrates, as the previous analyses have shown. To address this point the standard approach is to estimate the branch lengths that separate *O. dioica* and Ciona

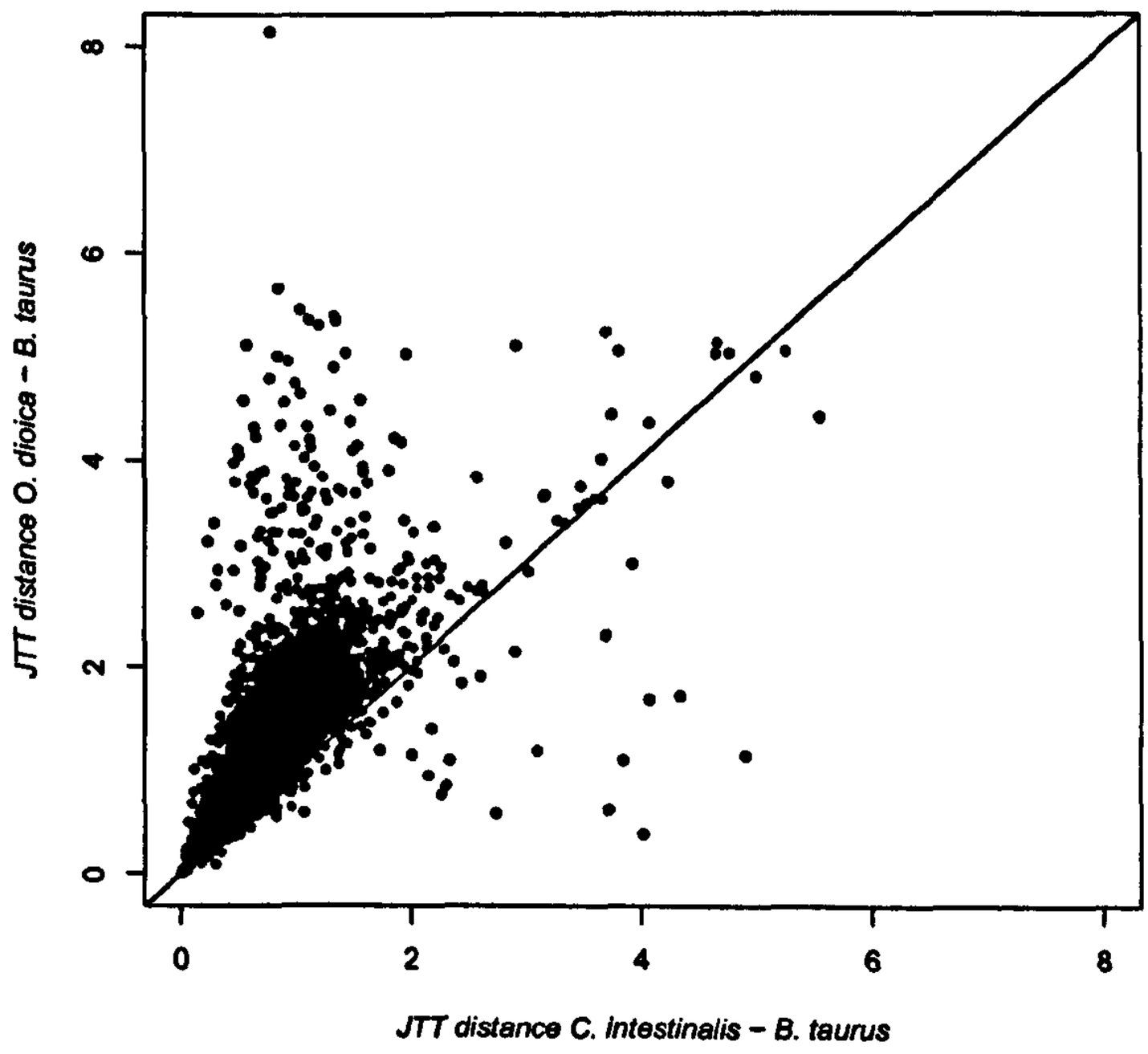


Figure 3.3: Scatter plot of JTT distances between the ortologous pairs of *B. taurus* (as an outgroup) with *C. intestinalis* and *O. dioica*

from their common ancestor (i.e the so called Sarich and Wilson test). For this purpose these branch lengths were calculated using different outgroups in order to avoid possible biases due to the sample used. As shown in Table 3.5 the inferred distance **a**, between *O. dioica* and its common ancestor with Ciona was, in all the datasets, greater than the distance **b**, namely the distance between *C. intestinalis* to the same common ancestor, by a factor of more than 2. That is, on average *O. dioica* genes evolve 2.1 times faster than *C. intestinalis* (Table 3.5). The distribution of the ratios **a/b** for all genes in one of these datasets is presented in Figure 3.4.

The results presented in this section show that the acceleration observed in both Cionas, is not a particular feature of ascidians, but seems to be a general feature of tunicates, since the rate of molecular evolution is even greater in *O. dioica*, who belong to a different class (larvacean) within tunicates.

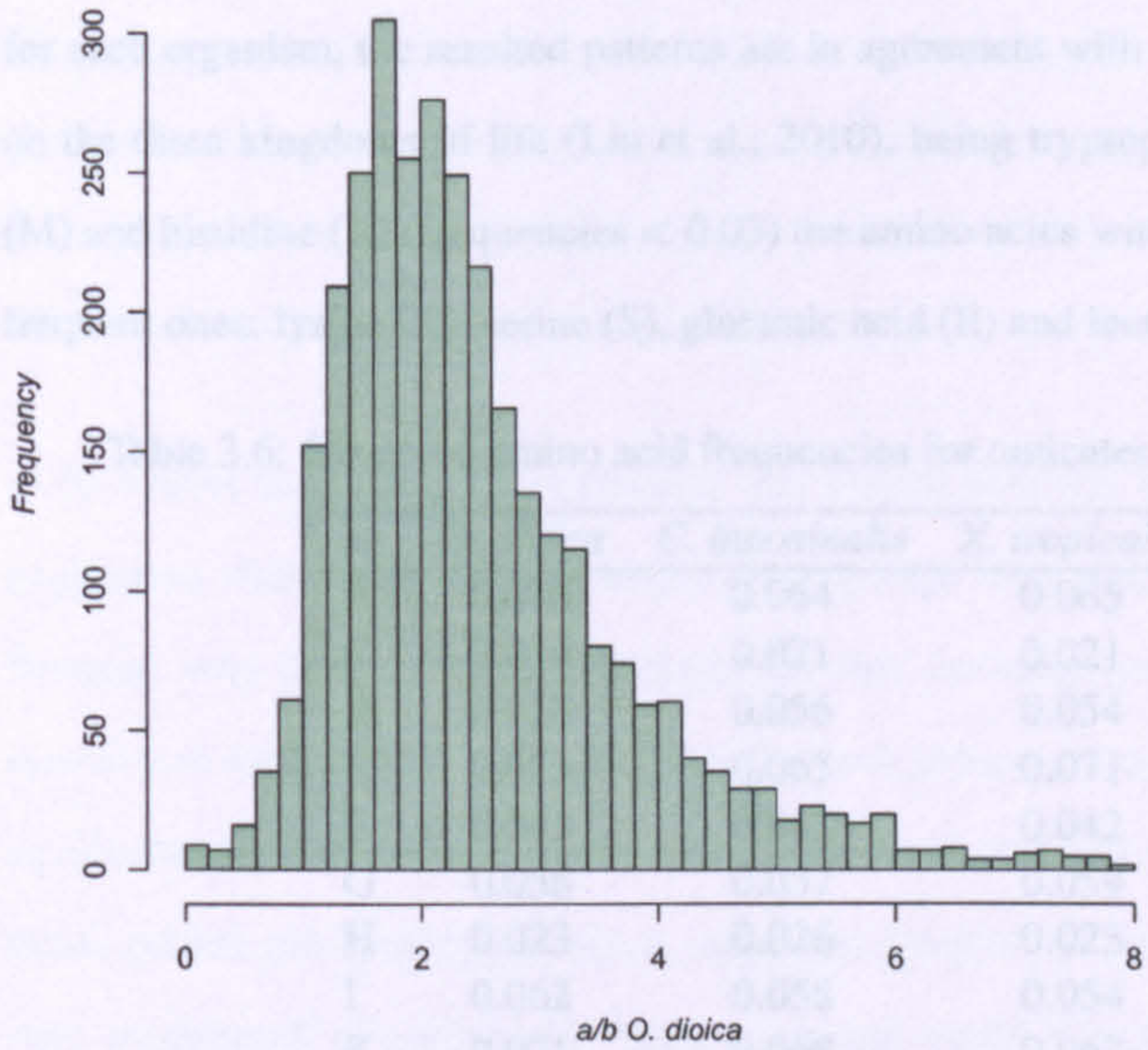


Figure 3.4: Distribution of **a/b** ratio. In the figure, **a** and **b** correspond to the distances between the species *O. dioica* and *C. intestinalis* to their common ancestor, respectively. In this case *B. taurus* was used as the outgroup

3.2.4 Amino acid composition in tunicates

Patterns of amino acid frequencies of *C. intestinalis* and *O. dioica* were compared with those of vertebrates and amphioxus. For this purpose quartets of orthologous protein were used, and the amino acid frequencies were calculated for each species. Clearly all studied species have very

Table 3.5: Average distance of orthologs sequences from 1- *O. dioica*, 2- *C. intestinalis* and 3 - one outgroup. **a** - correspond to the distance between *O. dioica* and the common ancestor with *C. intestinalis*, and **b** is the distance between *C. intestinalis* and the common ancestor with *O. dioica*. $*$ using averaged branch length

	# orthologs	d12	d13	d23	**a**	**b**	**a/b**	**a/b***
3- *H. sapiens*	3569	1.31	1.3895	0.9085	0.8955	0.4145	2.1612	2.1602
3- *B. taurus*	3546	1.308	1.409	0.9374	0.8898	0.4182	2.7053	2.1276
3- *Amphioxus*	3534	1.329	1.3958	0.8837	0.9206	0.4084	3.2762	2.254
3- *Sea Urchin*	3316	1.3321	1.488	1.0112	0.9045	0.4276	2.5788	2.1151
1- *O. dioica*								
2- *C. intestinalis*						average =	**2.6804**	**2.1642**

similar amino acid frequencies (Table 3.6), being the differences between them not statistically significant in any case. Indeed, when the amino acids were ranked according to their frequencies for each organism, the resulted patterns are in agreement with a general amino acid composition on the three kingdoms of life (Liu et al., 2010), being tryptophan (W), cysteine (C), metionine (M) and histidine (H) (frequencies < 0.03) the amino acids with lowest frequencies, and the most frequent ones: lysine (K), serine (S), glutamic acid (E) and leucine (L) (frequencies <0.065).

Table 3.6: Observed amino acid frequencies for tunicates, vertebrates and amphioxus.

aa	*O. dioica*	*C. intestinalis*	*X. tropicalis*	*H. sapiens*
A	0.070	0.064	0.065	0.069
C	0.019	0.021	0.021	0.021
D	0.058	0.056	0.054	0.052
E	0.073	0.065	0.071	0.072
F	0.045	0.043	0.042	0.041
G	0.058	0.057	0.059	0.060
H	0.023	0.026	0.025	0.026
I	0.062	0.058	0.054	0.051
K	0.071	0.068	0.067	0.064
L	0.091	0.095	0.100	0.103
M	0.025	0.027	0.025	0.024
N	0.046	0.046	0.041	0.038
P	0.042	0.043	0.047	0.050
Q	0.041	0.041	0.045	0.045
R	0.052	0.051	0.052	0.055
S	0.071	0.071	0.072	0.069
T	0.052	0.055	0.052	0.051
V	0.062	0.069	0.065	0.066
W	0.012	0.012	0.012	0.013
Y	0.030	0.033	0.032	0.031

3.2.4.1 Patterns of amino acid substitutions in tunicates

In order to further characterize the process of amino acid substitution in tunicates their substitution patterns were analysed. That is, determining which amino acids are replaced by others during the course of evolution, if there are amino acids that are more evolutionary stable than others, and if the present day amino acid composition is at evolutionary equilibrium. For doing this, two groups of matrices of counted amino acid substitutions were built, namely one accounting for those amino acid substitutions that took place in the branch leading to *O. dioica* and another in that leading to

C. intestinalis.

It should be taken in mind that for this purpose the ancestral sequence (*i.e.* the node) must be inferred, since the matrices are polarized. This means that, each entry M_{ij} in the matrix represents the number of substitutions from amino acid *i* to amino acid *j*. This in turn implies that one or more outgroup are required. The inference was done by both maximum parsimony and maximum likelihood (see Materials and methods 6.10).

To control the behaviour of proteins that exhibit different evolutionary speed, four groups of alignments were analysed separately. These groups were defined according to their divergence. Namely: very slow, slow, intermediate, and fast, evolving proteins (Table 3.3). The mentioned matrices of substitutions were generated for each group of proteins for both tunicate species. The equilibrium amino acid frequencies, which are those that will be eventually reached if the substitution pattern observed in nowadays sequences remains for a long period of time (*i.e.* the steady state equilibrium of amino acid frequencies that would be achieved by the inferred substitution pattern) were also estimated (see Material and Methods 6.10).

These analyses yield some interesting results that deserve to be mentioned. In the first place, the four groups of Oikopleura alignments (namely those ones that were separated according to their divergence) have very similar amino acid equilibrium frequencies. As it can be observed in Figure 3.5, in *O. dioica* the expected equilibrium frequencies for each one of the four groups, is similar to that obtained with the whole dataset. Note that due to the fact that the majority of genes belong to fast evolving group (2112 sequences , 66% of the sample), the excellent agreement between the expected equilibrium frequencies in the comparison involving fast genes and the whole dataset is not surprising. This in part can be attributed to the fact that this is an almost self comparison. However, the agreement in expected amino acid equilibrium frequencies is quite good for the three other groups as well. The all to all comparisons (slow vs. intermediate, fast and so on) gave again the same results: the expected amino acid equilibrium frequencies are basically the same for all groups of genes. Needless to say, these latter comparisons involved completely independent sets of genes. Two straightforward conclusions can be drawn. The first one, technical, yet important conclusion, is that the estimations could be considered as "reliable", since four independent assessments gave basically the same results. The second, and more biological mean-

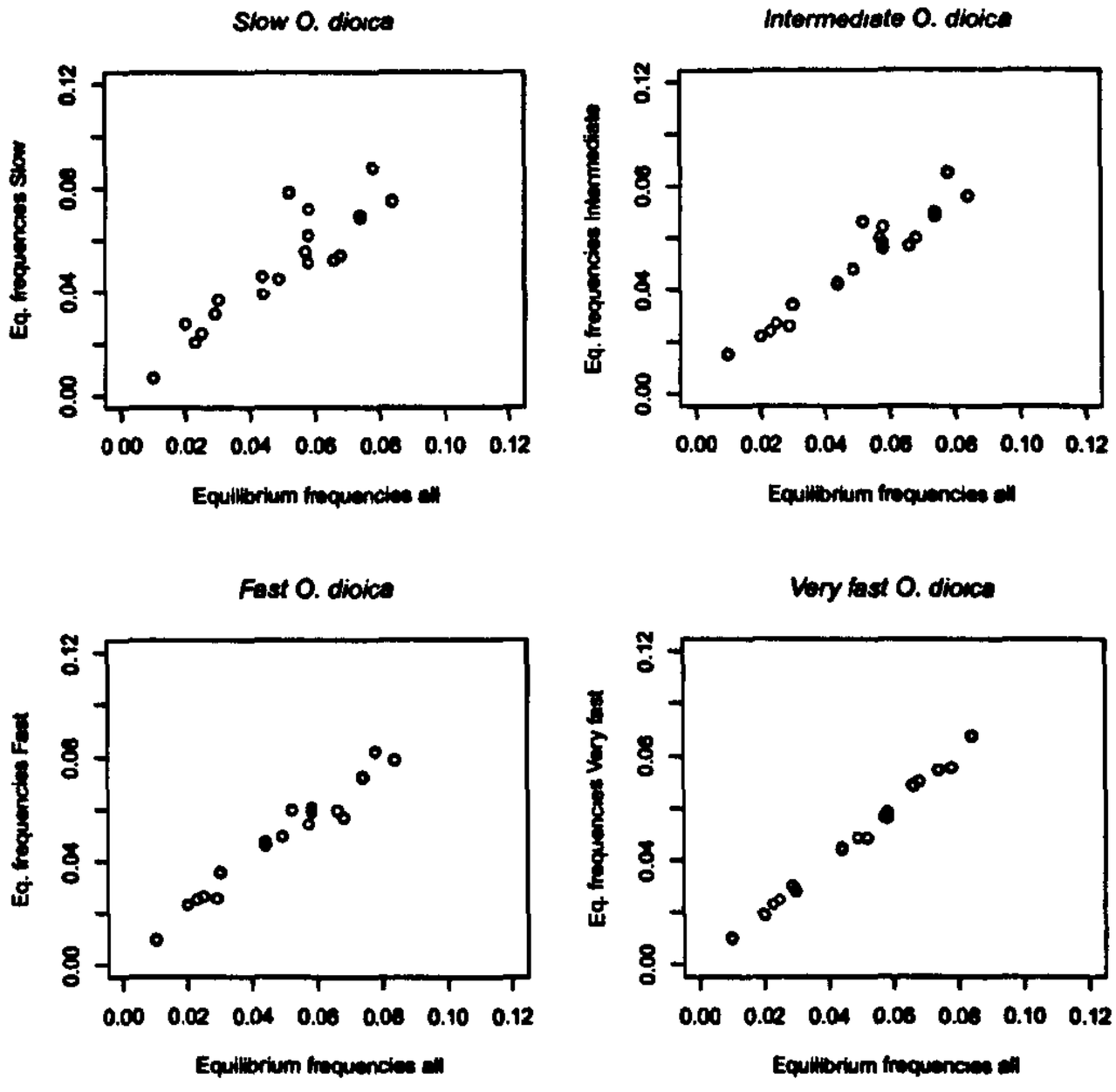


Figure 3.5: Equilibrium frequencies of the overall dataset of alignments versus each group of sequences analysed of *O. dioica*

ingful conclusion is that the equilibrium state is independent from the rate of evolution. In other words, this means that fast and slow evolving genes exhibit virtually the same pattern of amino acid substitutions. Similar results were obtained for *C. intestinalis*, *i.e.* the four groups of genes yield basically the same amino acid equilibrium frequencies, indicating that in both tunicates the expected equilibrium state is independent from the rate of evolution (SupFig. 5).

Additional aspects of the amino acid substitutions patterns in tunicates were also analysed. For this purpose the observed amino acid frequencies in nowadays sequences versus the expected frequencies at equilibrium were compared. Interestingly enough, *O. dioica* genome exhibited amino acid frequencies not far away from that expected at equilibrium, as it can clearly be observed in Figure 3.6 a (The same result was observed for each group of alignments so far analysed, and was reported in SupFig. 3). In contrast, the same analysis in *C. intestinalis* gives results that are essentially different (Figure 3.6 b). In effect, for this latter species, the observed amino acid frequencies were visibly far away from the inferred equilibrium state (SupFig. 4). On the other

hand, the amino acid frequencies are quite similar between the two species (Table 3.6).

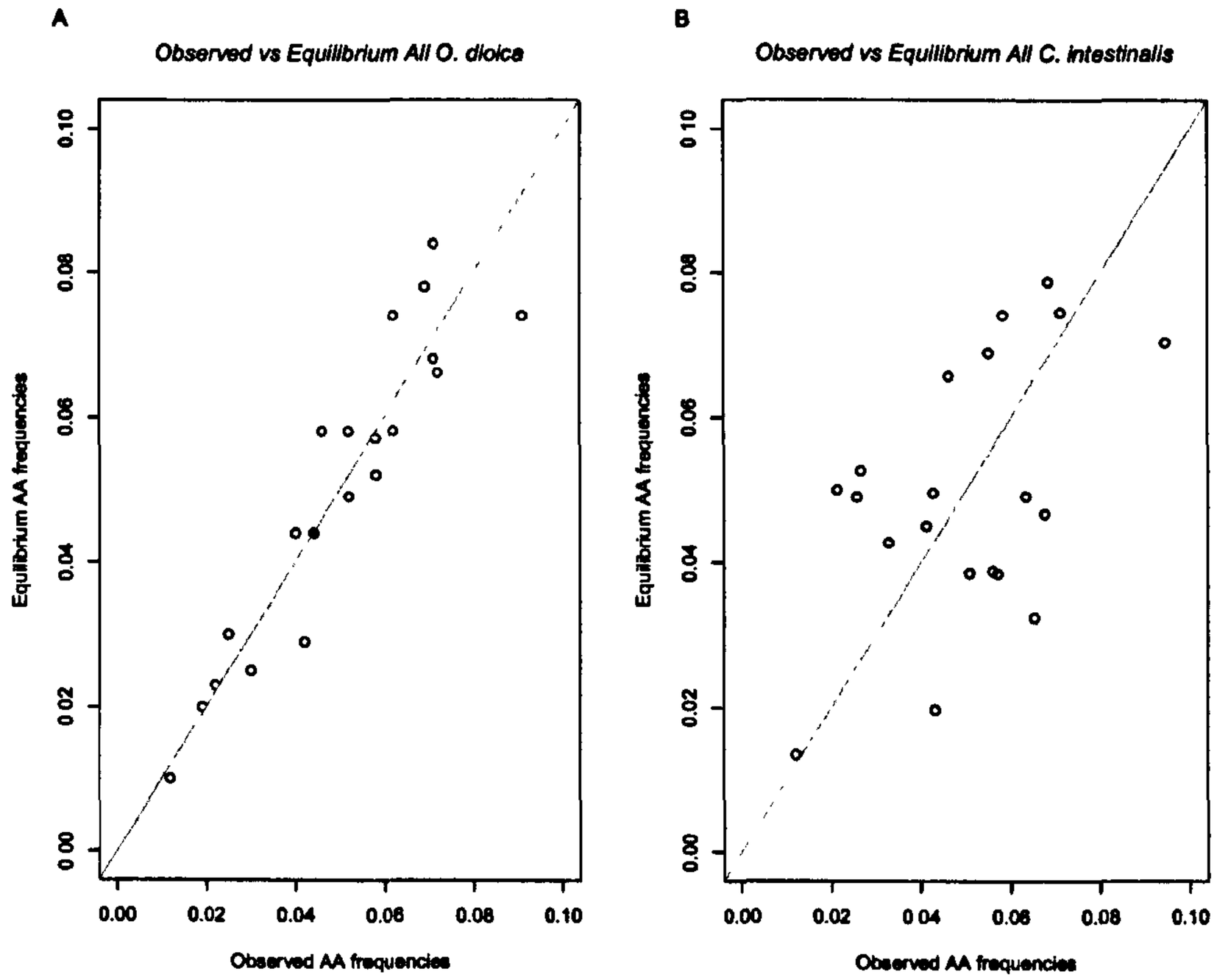


Figure 3.6: Equilibrium frequencies vs Observed frequencies for all orthologs analysed of *O. dioica* and *C. intestinalis*

3.2.4.2 Pattern of amino acid gain and loss

A recent investigation shows that there may exist a universal trend of amino acid gain and loss in protein evolution (Jordan et al., 2005). However, the possible causes that contribute to this phenomena are still in debate (Goldstein and Pollock, 2006; Hurst et al., 2006). With the purpose of establishing if this trend is also observable in these tunicates, the gain and loss phenomenon was investigated. The total gain and loss for each amino acid (normalized fraction of all substitutions that created (C) and substitutions that removed (R) the amino acid) was calculated for both organisms. Figure 3.7 shows in alphabetical order the ratio of create/remove substitutions. As expected from the previous results, the amino acid variation [(C - R)/(C + R)] are smaller and closer to zero in *O. dioica*, indicating that its amino acid frequencies are near to the equilibrium state. Conversely, *C. intestinalis* present patterns of gain and loss similar to that described for the

other species, in agreement with the fact that it exhibits greater distance to equilibrium.

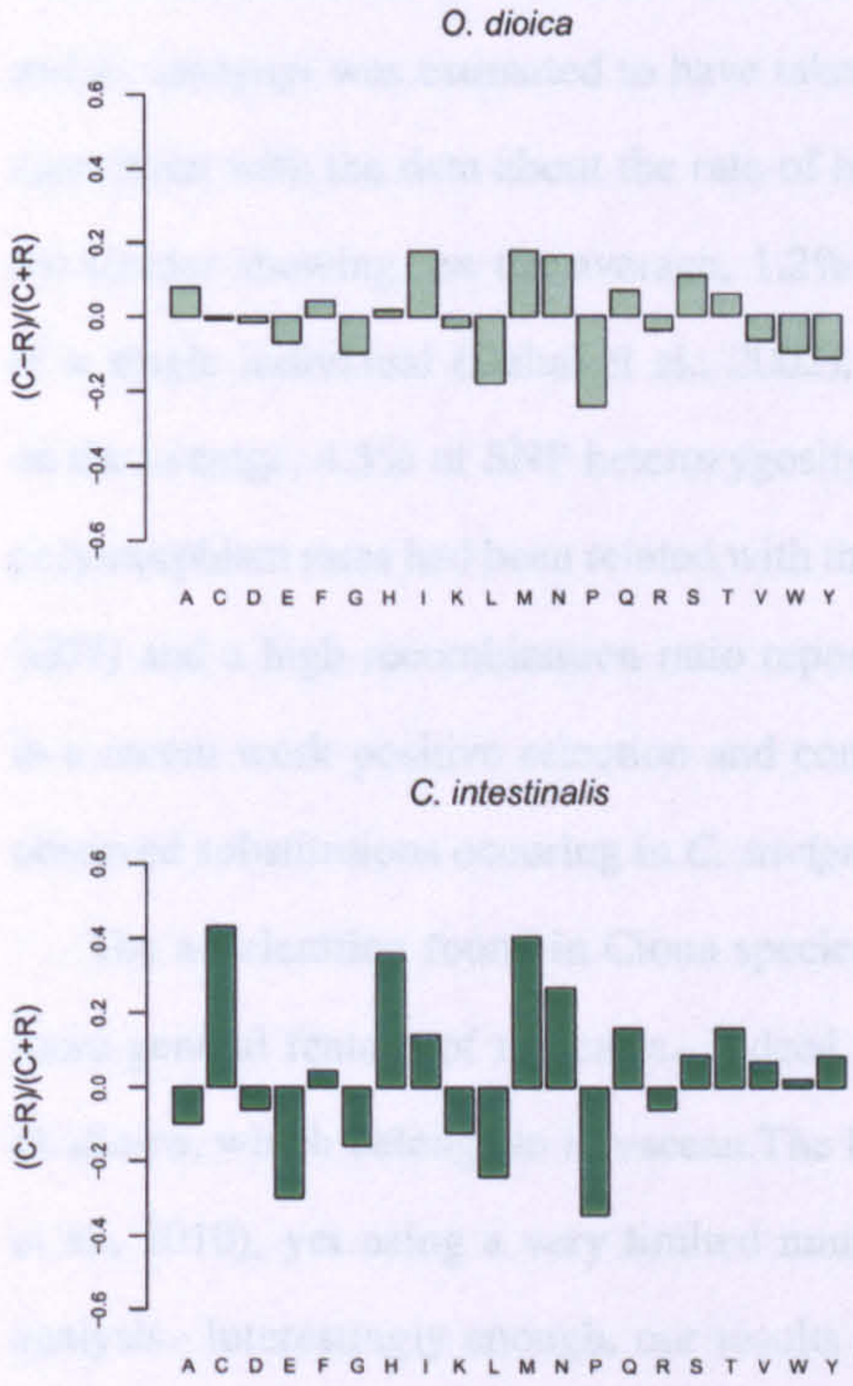


Figure 3.7: Amino acid gain and loss in *O. dioica* (light green) and *C. intestinalis* (green)

3.3 Discussion

The relationships between distances and divergence time (obtained from the fossil records), were established for vertebrates. Indeed, the relationship fitted quite well a clocklike behaviour for all vertebrates. Two exceptions, the non-placental mammals opossum and ornithorhynchus, that presented a slightly accelerated molecular evolutionary rate, were removed from the analysis. In order to use the vertebrate molecular clock to estimate the divergence time between Cionas, we first tested if they evolve at the same molecular rate by applying the Tajima's relative rate test. The results showed that Ciona, in fact, does not evolve at the same rate of vertebrates. On the contrary, the acceleration was estimated in a factor of 1.5, that means, *C. intestinalis* evolves 50%

faster than vertebrates (the same result is also valid for *C. savignyi*). With this information on one hand and the calibration of molecular clock on the other, the divergence between *C. intestinalis* and *C. savignyi* was estimated to have taken place about 184 (±15) My ago. This result was in agreement with the data about the rate of heterozygosity in both *C. intestinalis* and *C. savignyi*, the former showing, on the average, 1.2% of nucleotides differing between chromosome pairs of a single individual (Dehal et al., 2002), and the latter, even much more polymorphic, with, on the average, 4.5% of SNP heterozygosity (Vinson et al., 2005; Small et al., 2007). These high polymorphism rates had been related with the effect of large effective population size (Small et al., 2007) and a high recombination ratio reported in *C. intestinalis* (Kano et al., 2006). Moreover, in a recent work positive selection and compensatory substitutions were invoked to explain the observed substitutions occuring in *C. savignyi* genome (Donmez et al., 2009).

The acceleration found in Ciona species is not only a peculiarity of ascidian genomes, but a more general feature of tunicates. Indeed, an even higher degree of acceleration was found in *O. dioica*, which belongs to larvacean.The latter result was recently confirmed by (Tsagkogeorga et al., 2010), yet using a very limited number of genes, instead of conducting a genome wide analysis. Interestingly enough, our results show that *O. dioica* evolve 2 time faster than Ciona. This observation implies that *O. dioica* is at least 3 time faster than vertebrates. Certainly, this high molecular evolutionary rate is represented as a long branch leading to *O. dioica*, when a maximum-likelihood phylogenetic tree was obtained using concatenated alignments comprising 3220 genes (Figure 3.8).

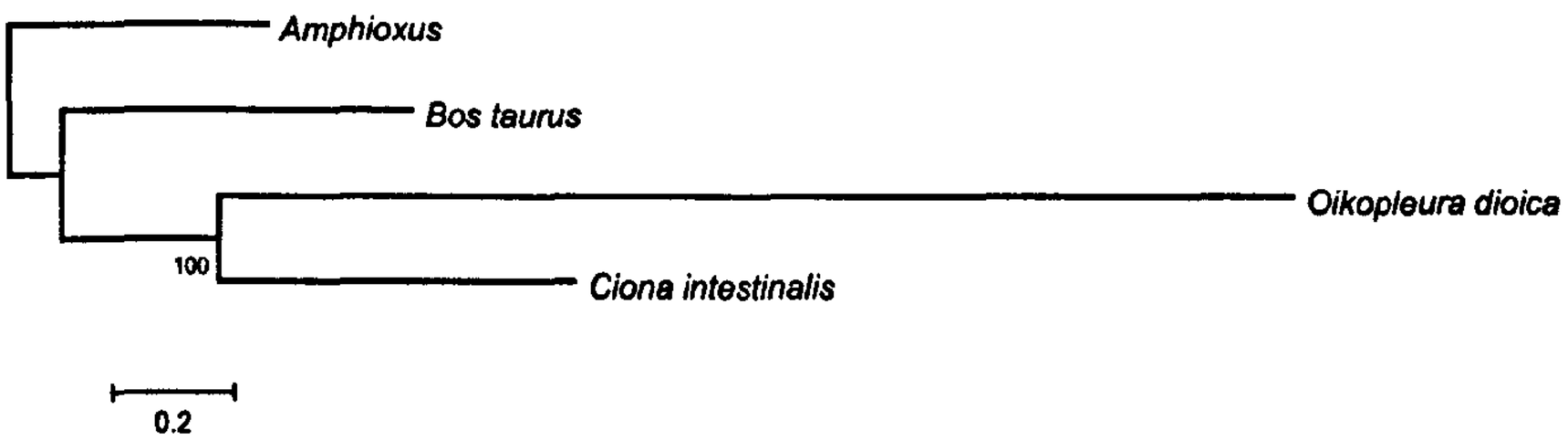


Figure 3.8: Maximum likelihood phylogeny of 3220 concatenated COGs

The disparity in rates between tunicates and vertebrates can either be the result of an acceleration in the former or a slowing down in the latter. The second possibility has been proposed as

the most probable by Peterson who compared vertebrate with the remaining metazoan (Peterson et al., 2004). One may wonder, what are the evolutionary forces responsible of the changes in molecular evolutionary pace described above. Changes in mutation rate, population size, selection coefficients, metabolic rate and generation time, have been claimed to affect the variation in substitution rates among species (Martin and Palumbi, 1993; Bromham and Penny, 2003). However, more recent reassessments of these factors did not fully support many of the initial claims. For instance, in spite of the fact that the metabolic rate is approximately one order of magnitude higher in endothermic amniotes than in ectothermic ones of similar body mass (Montes et al., 2007), the substitution rate was found to be of the same order of magnitude (Peterson et al., 2004). Moreover, the so called "generation time effect" was not supported in many cases, since, for instance rodents and carnivores, that have disparate generation times, exhibit close substitution rates (Huttley et al., 2007). Other analysis and interpretations have been put forward to explain why Ciona evolves faster than vertebrates. In this regard, it is worth mentioning that according to Peterson et al. (2004) the main factor that underlies the different evolutionary rate in chordates is related to the issue of genome duplication. Indeed, two rounds of whole genome duplications took place in vertebrates, the first one before the branching of agnates from the remaining vertebrates, but after the separation of ascidians, and the second one just before the split of cartilaginous fish and bony vertebrates (Putnam et al., 2008). According to this point of view, vertebrate genomes would evolve more slowly as a result of an increased number of protein-protein interactions which implies that a higher proportion of surface amino acids would be involved in these interactions. Participating in specific protein-protein interactions limits the possibility to vary, namely the participant amino acids would be under more stringent selective constrains (Fraser et al., 2002). This double duplication hypothesis nevertheless is at odds with the classical hypothesis proposed by Ohno according to which genome duplication results in higher evolutionary rate because one of the gene copies would be more free to accept mutations (Ohno, 1970). We can conclude that further data and analyses are required in order to shed light on this controversial point.

Even if it seems very reasonable that vertebrates have lowered down their rate of amino acid substitution, it also seems almost a fact the rate of evolution has also been increased in tunicates, and particularly in *O. dioica*. In fact when one compares the rates of evolution with those of representative invertebrates like echinoderms, insects, hemichordates, mollusc, etc, it appears that

O. dioica is the fastest evolving metazoan so far recorded. Although the biological basis for this extreme and unusual acceleration remains obscure in many aspects, some points can be made: The acceleration is not due to a small number of genes that are very fast evolving, but an effect affecting almost all genes in the genome. This fact rules out adaptive evolution (i.e positive selection) as the possible cause of the acceleration. Note that if adaptive evolution was the cause of this unusual acceleration, then it is to be expected that it would affect the group of genes that help the organism to get adapted to new environments or to new challenges. Even if this group in some situations may represent substantial fraction of the whole genome, it seems very unlikely that almost all genes were under adaptive pressure. By contrast, it is more plausible that the rate increase is due to relaxation of selective (functional) constraints or increase in the mutation rates. These two alternatives can be tested if DNA sequence data were available. The former situation would result in a decrease in the Ks/Ka ratio, while in the latter situation this ratio is expected to remains unchanged. Regretfully no DNA sequence is publically available to test these two hypotheses.

Certainly, determining the rate of protein evolution, as well as understanding the causes of variation in protein evolutionary rates is fundamental to understand protein evolution. Recently diverse works applied statistical approach to analyse amino acid composition and try to increase knowledge in evolution of genetic codes, evolution of protein families, prediction of secondary structures and the origin of amino acid among other (Liu et al., 2010). Regarding to the latter point, Jordan and collaborators determined an universal trend of amino acid gain an loss in protein evolution (Jordan et al., 2005). If proteins were at compositional equilibrium and their evolution was stationary and governed by a reversible process, symmetrical matrices of amino acid substitution would be generally observed. However using 15 sets of organisms (two ingroups and two outgroups) and analysing all the possible amino acid changes, they establish that proteins are not in equilibrium and their evolution is not reversible (Jordan et al., 2005). Thus, they claim that asymmetric amino acid substitution matrices ought to be used in phylogeny studies based on multiple alignments sequences. Moreover, taking into consideration all the amino acid substitutions regardless of the number of nucleotide replacements required, they described five amino acid as strong "gainers" (cysteine, metionine, histidine, serine and phenylalanine) and four as strong "losers" (proline, alanine, glutamic acid, and glycine) and considered this trend as universal and not driven at DNA level. Although Goldstein and Pollock claimed that this pattern could be an

artifact explained by statistical bias (Goldstein and Pollock, 2006) the patterns described by Jordan and collaborators, were confirmed by McDonald (2006) and Hurst et al. (2006) and by those reported here for *C. intestinalis*. Indeed, when the total gain and loss for each amino acid was calculated for *O. dioica* small differences between gained and lost amino acids were found, indicating that its amino acid frequencies are near to the equilibrium state. Conversely, *C. intestinalis* present patterns of gain and loss similar to those described in other organisms. Specifically, cysteine metionine and histidine were found as "gainers" and proline glutamic acid and glicine as "losers". Jordan and collaborators claimed that this trend is due to the fact that amino acid that were incorporated in the last term to the genetic code (with a low frequency) are still increasing their frequencies, and those that are thought to be among the first incorporated are nowadays decreasing their frequencies. This conclusion was immediately questioned, for instance McDonald in a detailed work argued that the trend found by Jordan and collaborators far from being universal could be explained by nearly neutral theory (McDonald, 2006). In the same line, Hurst, Feil and Rocha added more evidence to explain this pattern as an effect of nearly neutral theory, "mutation is biased towards the newer or costlier amino acids, but time-lagged fixation is biased toward older or cheaper amino acid replacement" (Hurst et al., 2006).

3.4 Conclusions

Genome wide analyses of molecular rates in tunicates and careful use of vertebrate fossil records permit to estimate a reliable divergence between *C. intestinalis* and *C. savignyi*, that took place around 180 My ago. What is more, tunicates show to have a striking high molecular evolutionary rate, being at least 50% faster than vertebrates. Particularly, the genome of *O. dioica* appears to be one of the fastest evolving metazoans. The acceleration that tunicates present is at genomic level and cannot be attributed to selective pressures, on the contrary, it is possible due to a relaxation of selective constraints or an increase in the mutation rates.

4 | *Comparative analyses of base compositions in vertebrate genomes*

4.1 Introduction

To understand the mechanisms affecting the nucleotide substitution is fundamental to comprehend evolutionary biology, population genetics and mutation research (Martin and Palumbi, 1993). Mutation in turn is the ultimate source of genetic variation. In other words, the mutational process is one of the fundamental forces of evolution enabling evolutionary changes (Barton, 2010; Loewe and Hill, 2010). Already in 1968, Kimura, taking in consideration that each nucleus present similar DNA content among different mammals, and that the average GC content ranged from 40% to 44%, established that "nucleotide substitution played a principal part in mammalian evolution" (Kimura, 1968). Needless to say, he also considered that most mutations are produced by nucleotide replacement that are almost neutral. To a certain extent this is true because mutations do not originate in a certain way or time that is related to whether their effects are beneficial, but, on the other hand, mutations are the result of complex biochemical reactions resulting in a not random distribution of mutation frequencies, favouring some changes over others (Loewe and Hill, 2010).

Mutations can occur by different process and in different ways, for instance there are: point mutations, insertion and deletions, transpositions, inversions and chromosome mutations, among other. They are in turn affected by biochemical factors that affect the frequency in which these changes tend to occur. For example, transitions occur more often than transversions (Graur and

Li, 2000) and methylation of CpG sites leads to a higher mutation rates at these sites (Rosenberg et al., 2003).

Mutations are also classified as non-synonymous or synonymous, according if they produce or not an amino acid change. As a general rule, synonymous site are believed to be under weak selection or to evolve neutrally, on the other hand, non-synonymous site are considered to be under strong purifying selection (Loewe and Hill, 2010).

Different hypotheses have been proposed to explain differences in rate of DNA evolution, such as DNA repair efficiency, rate of cell division, generation time, and metabolic rate. Moreover many of these physiological variables are correlated with each other, and traditionally are correlated with body size, which "probably does not control the rate of DNA substitution directly but serves as a convenient guidepost for understanding the biological correlates of molecular rate heterogeneity" (Martin and Palumbi, 1993). While it is true that no single factor is likely to explain completely the variation in molecular rates, some process are more appealing to have an important role in DNA evolution. Particularly Martin and Palumbi considered that generation time and metabolic rate effects explain it equally well. Generation time effect considered that nucleotide substitutions are mainly the results of errors during replication, thus, theoretically if most species have similar numbers of cell division, species with shorter generation time will accumulate higher number of DNA changes per year. This effect was considered by Otha to be cancelled by the effect of population-size in mammals, supporting the nearly neutral theory of molecular evolution (Ohta, 1993). Metabolic rate is related to the molecular process of mutation, because metabolism produce oxygen radicals (highly reactive molecules with free electrons) and mediated it the damage to DNA sequences occur (Fraga et al., 1990). Indeed, oxidative damage is extensive in nuclear DNA. In this way, species with higher metabolic rates should have higher DNA substitution rates, due to the fact that DNA damage and mutation rate are positive correlated. Naturally, the DNA damaged by oxidation is continuously repaired and mutations may occur by incorrect reparation (Martin and Palumbi, 1993). The important effect of temperature and body size on metabolic rate was also studied. Metabolic rate scales with the $\frac{3}{4}$ power of body mass and increases exponentially with temperature, thus, mass specific metabolic rate is governed by two factors, the Boltzmann's factor describing the temperature dependence of biochemical processes ($e^{-E/kT}$) and the quarter-power allometric relation describing how biological rate scales with body size (Gillooly et al., 2001).

According to those authors, to normalize both, body mass and temperature, in order to compare the metabolic rate of fish living in different environments, the Boltzmann's factor was used.

It was shown in chapter 2 that the metabolic rate hypothesis could provide an appropriate theoretical frame to understand the observed compositional differences between tunicate species. In this chapter other chordate organisms were utilized in order to understand the possible effect of the metabolic rate hypothesis on driving the base composition evolution at the genome level. In this regard, two different strategies were used to investigate fish and mammalian genomes.

Regarding the first, the rationale of the choice was grounded on the consideration that, differently from terrestrial species, where the O_2 is free and is not a limiting factor, in aquatic animals the O_2 available in the environment is regulated by the Henry's law. Using the Boltzmann's factor, let us to disentangle temperature from the metabolic rate, in order to test if the two variables were playing and independent the role on the GC content. The result showed that both variables were decreasing according to the temperature of the habitats, and a significant correlation between the metabolic rates and genomic GC levels in teleostean fish was found.

Regarding the second, the use of KOG classification of human genes let us to perform comparative analysis of different functional classes in several mammalian and non-mammalian genomes. The results clearly showed that, mammalian genes involved in metabolic processes were characterized by higher GC3 level than those involved in "information store and processing" or those involved in "cellular processing and signaling". This clearly implies that a correlation between GC3 and genes involved in metabolic processes holds. The overall result presented in this chapter reinforces the metabolic rate hypothesis as one of the factors shaping genome composition.

Parts of the results presented here have been published in: Uliano, E., Chaurasia, A., Bern, L., Agnisola, C., and DOnofrio, G. (2010). Metabolic rate and genomic gc. what we can learn from teleost fish. Marine Genomics, 3(1):29 34.

4.2 Results

4.2.1 Learning from teleostean fish

In order to understand which forces are driving the evolution of vertebrate genomes and controlling the genome base composition, two datasets of fish were analysed. Data about taxonomic classification, geographical distribution and metabolism were retrieved (www.fishbase.org). The mass specific metabolic rate measured at different temperature was corrected using the Boltzmann's factor, more suitable than the Q_{10} factor (Clarke and Johnston, 1999; Gillooly et al., 2001). Data about genomic GC levels were obtained from available literature (Bucciarelli et al., 2002; Varriale and Bernardi, 2006). Thereby, all the data about fish were divided in five groups according to the habitats, namely, polar, temperate, subtropical, tropical and deep-water.

Body mass and metabolic rate according habitats The logarithmic values of body mass and metabolic rate were plotted according to the five habitat groups (Figure 4.1). The same decreasing trend, from polar to deep-water, was found (4.1, panels A and B), confirming that the body mass effect on metabolic rate was removed after the correction of the metabolic rate according to the Boltzmann's factor.

Table 4.1: p-values of Mann-Whitney test for metabolic rate of fish among different habitats.

		Polar	Temperate	Subtropical	Tropical
Teleostean	Polar	-			
	Temperate	$< 2.8x10^{-2}$	-		
	Subtropical	$< 7.1x10^{-3}$	ns	-	
	Tropical	$< 1.0x10^{-4}$	$< 1.0x10^{-4}$	$< 1.1x10^{-3}$	-
	Deep-water	$< 5.0x10^{-4}$	$< 2.2x10^{-3}$	$< 5.0x10^{-2}$	ns
Perciformes	Polar	-			
	Temperate	$< 5.7x10^{-2}$	-		
	Subtropical	$< 3.9x10^{-2}$	ns	-	
	Tropical	$< 1.0x10^{-4}$	$< 2.5x10^{-2}$	$< 8.2x10^{-3}$	-
	Deep-water	$< 1.0x10^{-2}$	ns	ns	ns

Regarding metabolic rate, polar fish showed the highest values, followed by temperate and subtropical fish, significantly higher than tropical and deep-water fish. However, the average metabolic rate of deep-water fish was not significantly different from that of tropical fish. The results were of great interest, considering that polar and deep-water fish live in environments

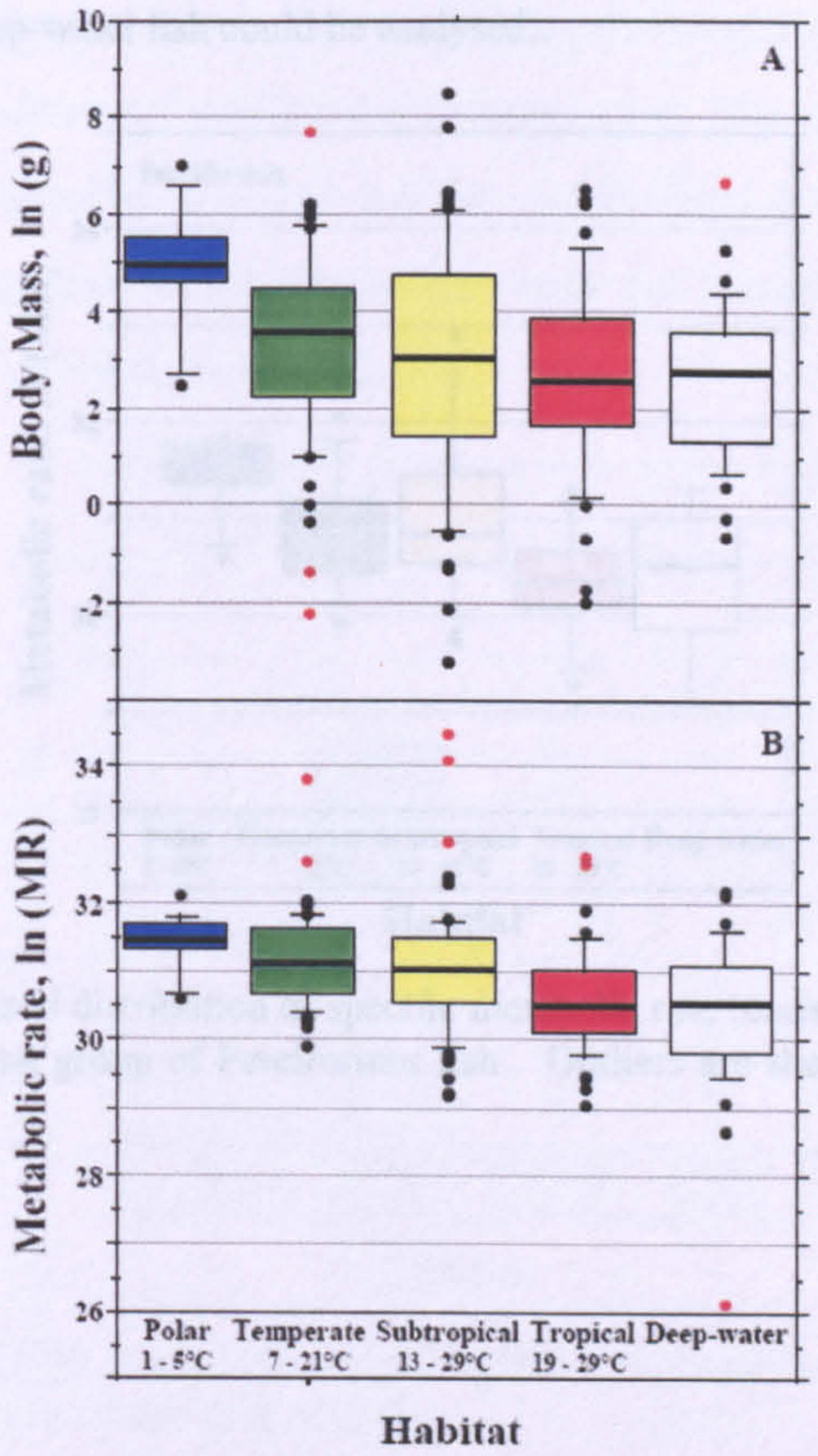


Figure 4.1: Log-normalized distribution within each habitat group of fish A) Box plot of body mass, B) specific metabolic rate (corrected for the Boltzmann's factor). Outliers are shown as a red points (from Uliano et al. (2010)).

characterized by very low temperature. The results of statistical analysis were presented in Table 4.1.

Effects of phylogenetic relationship The same decreasing trend of metabolic rate according to different habitats was found when only Perciformes fish were analysed (Figure 4.2), indicating that metabolic rate is not affected by phylogenetic relationship among species, an observation already reported by Clarke and Johnston (1999). Also in this case the metabolic rate of polar fish was significantly the highest (Table 4.1). However the metabolic rate of deep-water fish turned out to be not significantly different in any pairwise comparison, probably due to the very low number

of species (only four deep-water fish could be analysed).

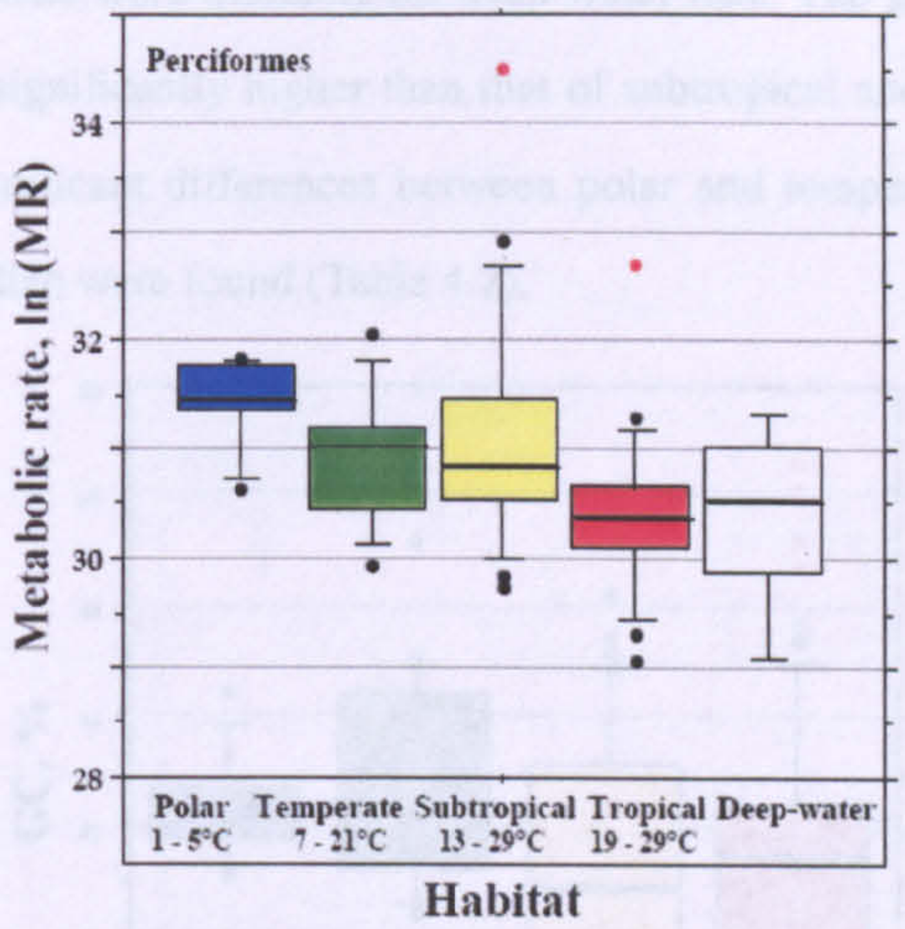


Figure 4.2: Log-normalized distribution of specific metabolic rate (corrected for the Boltzmann's factor) within each habitat group of Perciformes fish . Outliers are shown as a red points (from Uliano et al. (2010)).

GC among habitats The distribution of genomic GC content among different habitats was analysed. Unfortunately no data were available for deep-water fish. The genomic GC level of polar and temperate fish was significantly higher than that of subtropical and tropical habitats (Figure 4.3; Table 4.2). No significant differences between polar and temperate fish, neither between subtropical and tropical fish were found (Table 4.2).

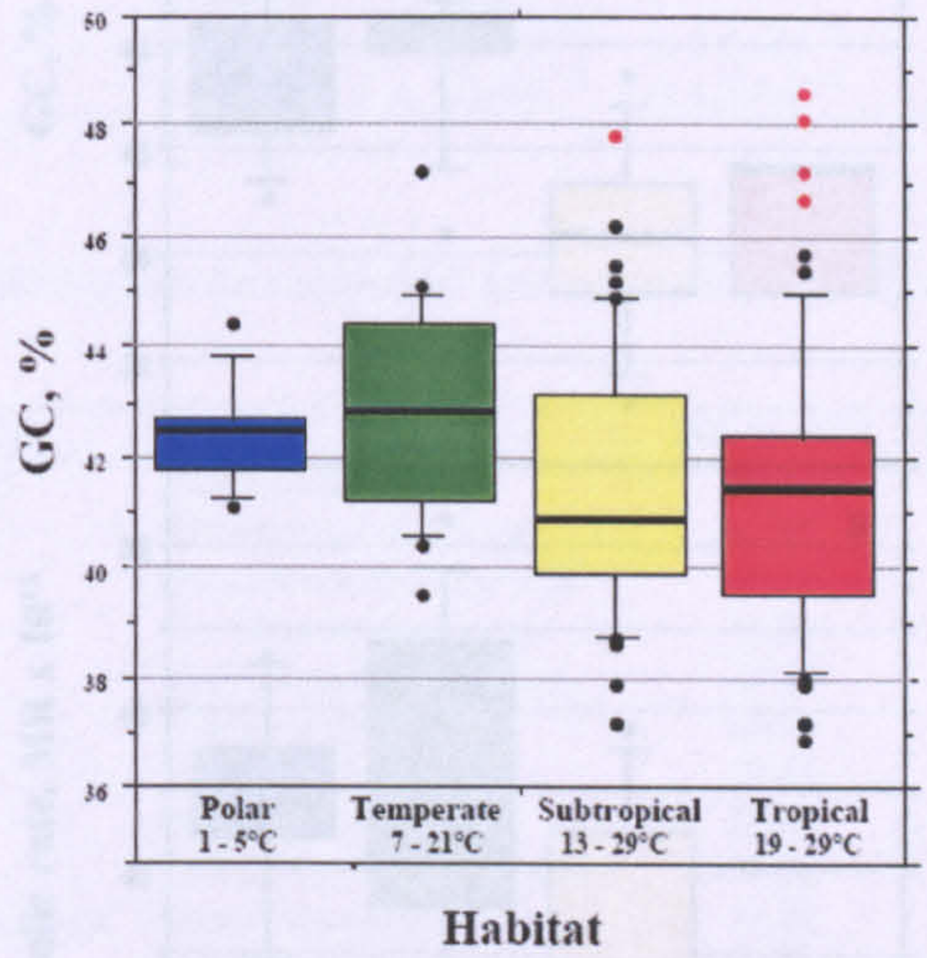


Figure 4.3: Box plot of GC% genomic levels distribution within each habitat groups. Outliers are shown as a red points (from Uliano et al. (2010)).

Table 4.2: p-values of Mann-Whitney test for GC levels among different habitats.

		Polar	Temperate	Subtropical
%GC	Polar	-		
	Temperate	ns	-	
	Subtropical	$< 5.5x10^{-2}$	$< 5.0x10^{-3}$	-
	Tropical	$< 1.9x10^{-2}$	$< 1.0x10^{-3}$	ns

Metabolic rate and GC Crossing data for metabolic rate and genomic GC level, data about 34 fish were obtained (Table 4.4). The distribution according to the four different habitats was analysed. The subset of 34 fish showed the same properties of the whole dataset (Figure 4.4,B). Certainly, regarding the metabolic rate no significant difference was found between polar and temperate fish. However, both presented a metabolic rate significantly higher than those of subtropical

and tropical fish, that in turn were not significantly different (Table 4.3). Regarding the genomic GC the same picture was found (Figure 4.4,A and Table 4.3).

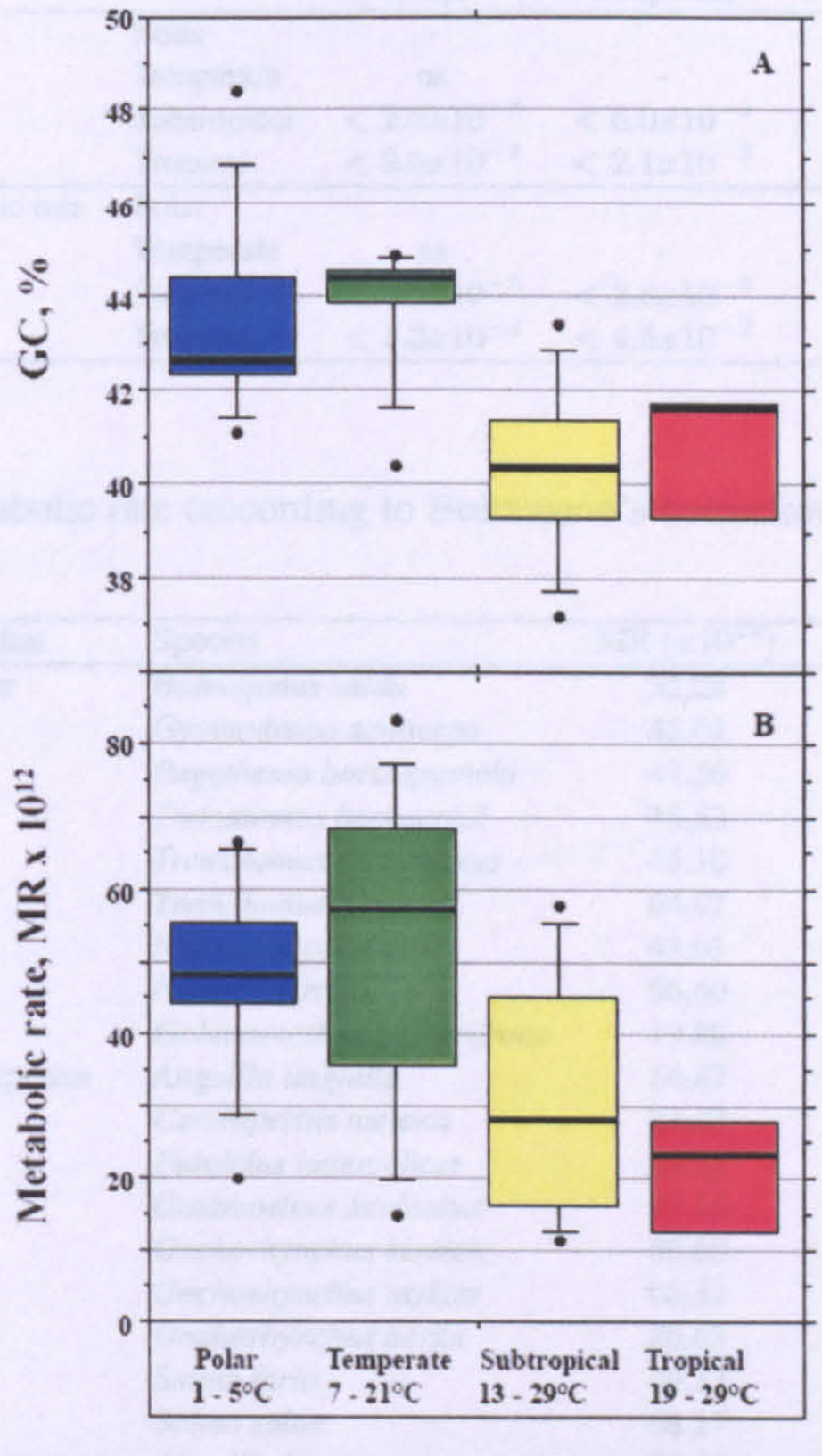


Figure 4.4: Distribution within each habitat group of fish A) GC level, B) specific metabolic rate, corrected for the Boltzmann's factor (from Uliano et al. (2010)).

Plotting the metabolic rate of the 34 species against the corresponding genomic GC, a positive and significant correlation was found (Figure 4.5 $R^2 = 0.252$, $p - value < 2.5x10^{-03}$). On the contrary, no significant correlation between GC levels and body mass was found ($R^2 = 0.004$).

Table 4.3: p-values of Mann-Whitney test for GC levels and metabolic rates of fish among different habitats.

		Polar	Temperate	Subtropical
%GC	Polar	-		
	Temperate	ns	-	
	Subtropical	$< 2.0x10^{-3}$	$< 6.0x10^{-4}$	-
	Tropical	$< 2.5x10^{-2}$	$< 2.1x10^{-2}$	ns
Metabolic rate	Polar	-		
	Temperate	ns	-	
	Subtropical	$< 3.3x10^{-2}$	$< 2.8x10^{-2}$	-
	Tropical	$< 1.3x10^{-2}$	$< 4.5x10^{-2}$	ns

Table 4.4: Average metabolic rate (according to Boltzmann's correction) and genome base composition.

Habitat	Species	MR ($x10^{12}$)	GC%
Polar	*Boreogadus saida*	52.28	48.4
	Gymnodraco acuticeps	45.62	42.6
	Pagothenia borchgrevinki	47.56	41.8
	Trematomus bernacchii	48.53	43.6
	Trematomus centronotus	45.10	42.5
	Trematomus hansoni	64.07	41.1
	Nototenia coriiceps	42.01	44.4
	Nototenia rossii	66.60	44.5
	Gobionotothen gibberifrons	19.88	42.5
Temperate	*Anguilla anguilla*	14.87	44.0
	Centropristis melana	83.43	44.9
	Fundulus heteroclitus	38.56	40.4
	Gasterosteus aculeatus	57.15	44.0
	Onchorhynchus kisutch	60.60	44.5
	Onchorhynchus mykiss	68.32	43.5
	Onchorhynchus nerka	26.65	44.4
	Salmo fario	49.33	44.8
	Salmo salar	68.27	44.4
Subtropical	*Anguilla rostrata*	37.51	42.6
	Brevoortia tyrannus	41.69	43.4
	Carassius auratus	20.65	37.9
	Chromis chromis	33.19	40.1
	Clinocottus analis	16.55	41.2
	Cyprinodon variegatus	48.40	40.6
	Cyprinus carpio	11.57	37.2
	Embiotoca lateralis	22.65	40.1
	Gillichthys mirabilis	16.11	38.4
	Ophiodon elongatus	13.23	41.5
	Opsanus tau	54.27	40.9
	Orizya latipes	58.02	40.1
Tropical	*Oreochromis aureus*	6.41	41.6
	Oreochromis mossambicus	27.90	41.8
	Oreochromis niloticus	18.93	41.7
	Danio rerio	27.50	36.9

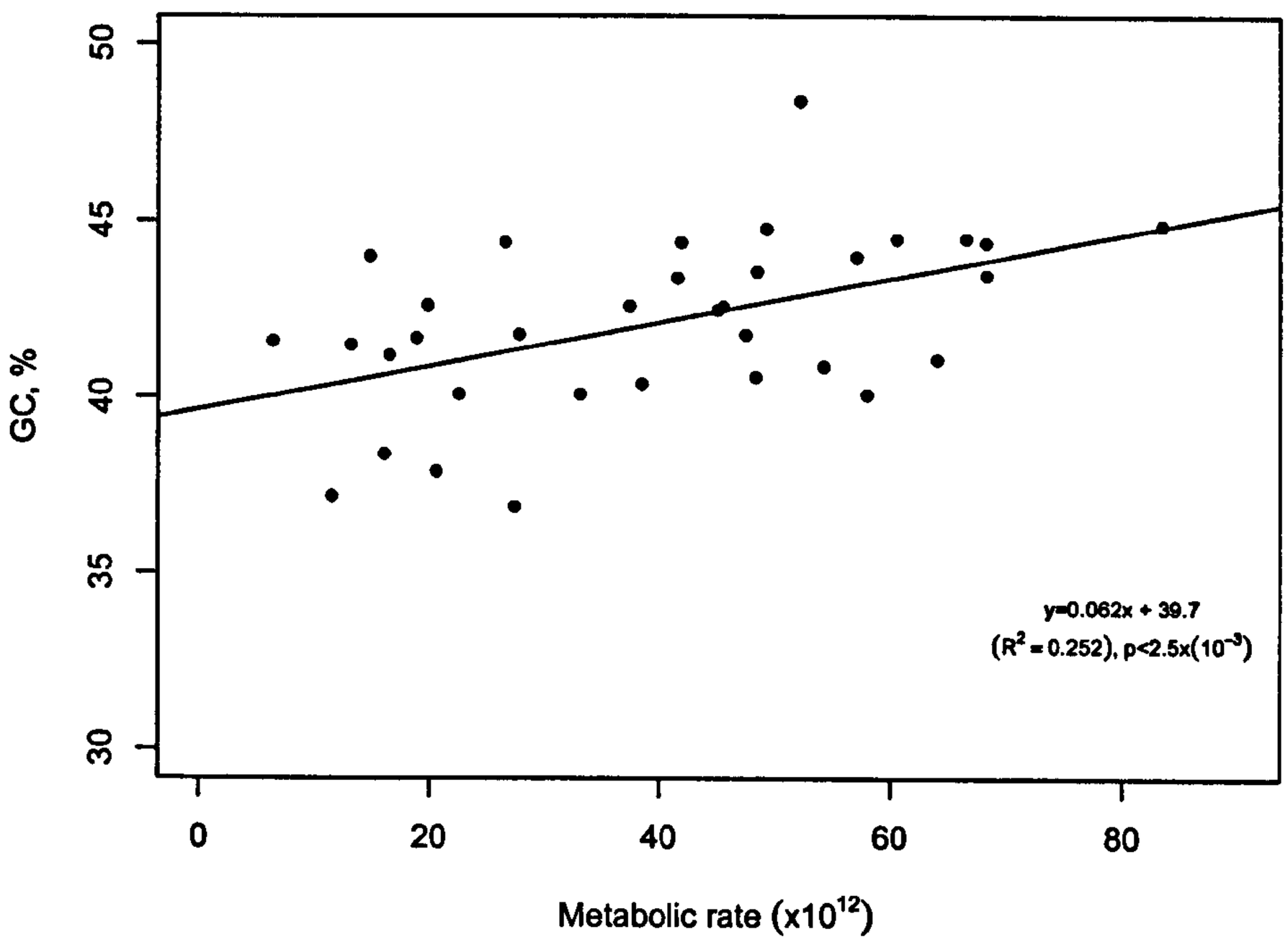


Figure 4.5: Specific metabolic rate (corrected for the Boltzmann's factor) versus the average GC content for 34 fish.

4.2.2 Human KOG genes

The human sequences retrieved from the Clusters of Orthologous Groups of proteins for eukaryotic complete genomes (KOG database), were classified in 25 functional classes, denoted by capital letter in square brackets (Tatusov et al., 2001, 2003). The corresponding number of genes and average GC3 level were reported in Table 4.5. Three main categories were defined: i) information storage and processing (in Blue), grouping five functional classes; ii) cellular processes and signalling (in Black), grouping ten functional classes; and iii) metabolism (in Red), grouping eight functional classes. Each category respectively represented, about 22%, 42% and 16% of the whole dataset. Genes whose function was predicted only [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.

4.2.3 Classification of vertebrate KOG genes

Using the best reciprocal hits approach, orthologs to human genes were searched in thirteen mammalian genomes (for details see Chapter 6), representing the following orders: primates (*G. gorilla* and *P. pygmaeus*), rodents (*M. musculus*, *O. cuniculus* and *S. tridecemmilineatus*), laurasiatheria (*B. taurus*, *E. caballus*, *P. vampyrus* and *T. truncatus*), afrotheria (*L. africana*), xenarthra (*D. novemcintctus*), marsupials (*M. domestica*) and monotremes (*O. anatinus*). The same approach was used for the amphibian *X. tropicalis* and for the reptile *A. carolinensis*. A gene of each species found to be ortholog of a specific human gene acquired the same KOG classification. In other words, through out orthology the KOG classification was extended to the genes of the 15 species so far analysed. For each species the whole number of orthologous genes, the amount of genes belonging to the Blue, Black and Red categories, as well as the corresponding GC3 levels and the standard deviation, were reported in Table 4.6.

4.2.4 Base composition of KOG genes

Preliminary analyses of human functional classes highlighted that the average GC3 level was significantly different among the three categories. Indeed, genes involved in the metabolic processes were those showing the highest GC3 levels. More precisely, the GC3 of the Blue category was

Table 4.5: 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	1273	0.545
[K]	Transcription	1137	0.619
[L]	Replication, recombination and repair	300	0.546
	CELLULAR PROCESSES AND SIGNALLING		
[D]	Cell cycle control, cell division, chromosome partitioning	267	0.552
[M]	Cell wall/membrane/envelope biogenesis	63	0.576
[N]	Cell motility	26	0.586
[O]	Post translational modification, protein turnover, chaperones	1471	0.557
[T]	Signal transduction mechanisms	2214	0.616
[U]	Intracellular trafficking, secretion, and vesicular transport	685	0.571
[V]	Defense mechanisms	1023	0.527
[W]	Extracellular structures	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 number of genes	16118	0.581

(#) Number of genes.

Table 4.6: Average GC3 levels, standard deviation and gene number of KOG's functional categories

		KOG[a]			BLUE[b]			BLACK[b]			RED[b]		
	Organism	GC3	s.d.	#	GC3	s.d.	#	GC3	s.d.	#	GC3	s.d.	#
Mammals (placental)	*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
(non placental)	*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
Reptiles	*A. carolinensis*	0.539	0.159	5959	0.535	0.158	3063	0.546	0.1655	1498	0.539	0.153	1398
Amphibians	*X. tropicalis*	0.500	0.112	3584	0.499	0.112	1753	0.499	0.1174	961	0.501	0.108	870

(a) Blue, Black and Red refers to the gene classification of KOG database (see Materials and Methods).
(b) Genes orthologous to KOG human genes.
(#) Number of genes.

significantly lower than that of the Black one, in turn significantly lower than Red one (D'Onofrio et al., 2007). This observation was confirmed by the present analyses taking into consideration a bigger set of genes (Table 4.6). More important, this trend was also studied in other vertebrate genomes, in order to determine to which extent this observation is applicable to other organisms. Interestingly, the same trend (Blue<Black<Red) was observed in several genomes, but not in the amphibian and reptile genomes, suggesting that base compositional differences among functional categories took place after the mammalian radiation (see Table 4.7). For almost all mammals the average GC3 was significantly different among the three categories (Red, Blue and Black), with the exception of *B. taurus*, showing not highly significant difference between Red and Black categories (p-value 0.1007), and for non-placental mammals (*O. anatinus* and *M. domestica*) where the differences between Blue and Black categories were not significant at all (p-value 0.7639 and 0.3252 respectively). This clearly demonstrates that the differences in base composition are not due to a random process, but there is a relation between higher GC content and genes related to metabolism.

4.2.5 De Finetti's diagram

In order to better assess within each mammalian genome the compositional/spatial distribution of the three functional categories, an analysis of the distribution of the functional classes over a GC3 range was performed, and represented by the de Finetti's diagram (Figure 4.6, see also Materials and Methods 6.13). This diagram was used to assess the compositional/spatial distribution of the three categories, i.e. Blue, Black and Red, in different genomes. Shortly, for each organism the whole GC3 range was split in three equal size intervals, corresponding to the levels denoted as Low, Medium and High, respectively. The number of functional classes belonging to the Blue, Black and Red categories were counted in each intervals and normalized to 1 for plotting.

In this figure the color of the dots corresponded to the three functional categories, and numbers close to each dot represent the occurrence of overlapping genomes. The de Finetti's diagram clearly showed that in the large majority of mammalian genomes: i) the Red category was rarely present in the lowest GC3 range, therefore confined to a restricted part of the space on the diagram; and ii) no overlap was observed with the spatial distribution of the Red category and Blue or

Black categories, being the former clearly grouped separately to one side of the triangle. On the contrary, Blue and Black categories are characterized by a partial spatial overlap. The significance of the first observation was tested performing 1000 class permutations and observing the diagram distribution of the Red class. The probability to find the spatial distribution of the Red category as far from the low interval was estimate to be $p - value < 1.76x10^{-2}$. This means that the observed configuration is due to a higher GC3 content of red classes and not by a random effect. Analysing reptile and amphibian genomes no different spatial distribution of the three categories was observed (data not shown).

4.2.6 The *Butterfly* plot

In order to better understand the genome organization of the species so far analysed, within each genome an analysis of the average GC3 of each class defined according to KOG was performed. More precisely, the delta of the average GC3 levels of each functional class against that of whole

Table 4.7: p-value of Mann-Whitney U-test among categories

	Organism	Red *vs.* Blue[a]	Red *vs.* Black[a]	Black *vs.* Blue[a]
Mammals (placental)	*H. sapiens*	0.0000	0.0000	0.0000
	G. Gorilla	0.0000	0.0027	0.0012
	P. pygmaeus	0.0000	0.0001	0.0300
	M. musculus	0.0000	0.0004	0.0006
	O. cuniculus	0.0000	0.0001	0.0005
	S. tridecemilineatus	0.0000	0.0022	0.0000
	B. taurus	0.0000	0.1070	0.0002
	E. caballus	0.0000	0.0002	0.0023
	P. vampyrus	0.0000	0.0276	0.0005
	T. truncatus	0.0000	0.0012	0.0028
	L. africana	0.0002	0.0101	0.0916
	D. novemcinctus	0.0000	0.0001	0.0002
(non placental)	*O. anatinus*	0.0958	0.0224	0.7639
	M. domestica	0.0015	0.0050	0.3252
Reptiles	*A. carolinensis*	0.6258	0.1748	0.0625
Amphibians	*X. tropicalis*	0.4994	0.8865	0.5715

(a) The null hypothesis is that the two variables have equal distributions.

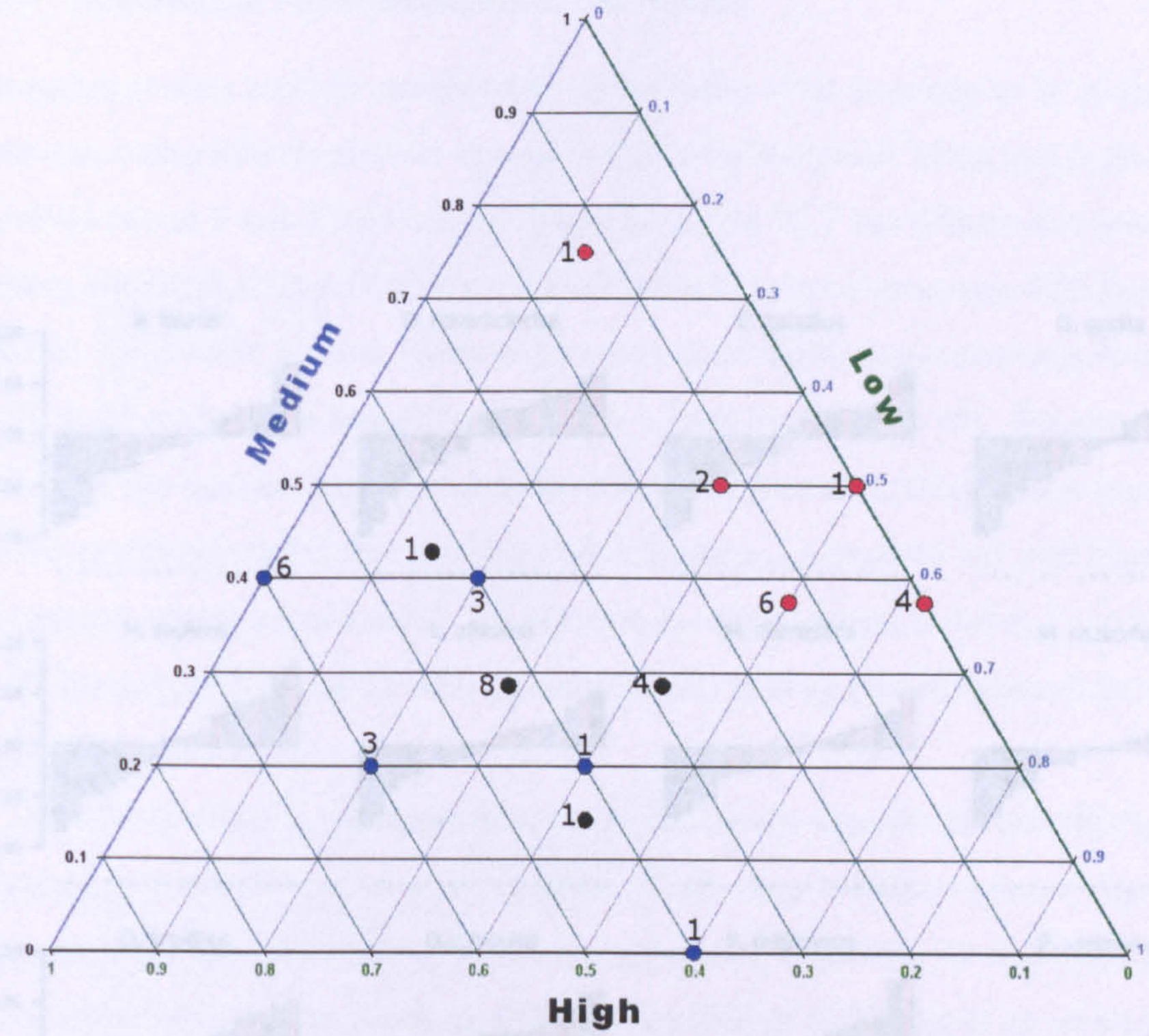


Figure 4.6: De Finetti's diagram showing the spatial distribution of the three functional categories: (i) information storage and processing (Blue dots); (ii) cellular processes and signalling (Black dots) ; (iii) metabolism (Red dots). Numbers close to dots refer to the occurrence of overlapping genomes.

corresponding genome was calculated. An overview of histograms obtained for each genome are presented in Figure 4.7, where the color code of the functional classes correspond to that of the three main categories. For sake of simplicity, this kind of plot will be defined as "butterfly plot". A more detailed representation for the butterfly plot of each genome so far analysed is reported as supplementary figures (SupFigs. 6, 7, 8, 9 and 10). At first glance it was evident that all mammalian genomes showed an unbalanced distribution of the bars according to the color code. Indeed, the blue bars where mainly in the negative side of the butterfly plot (i.e. showing an average GC3 level lower than the corresponding genomic one). On the contrary the red bars where mainly in the positive side, showing an average GC3 level higher than the corresponding genomic one (Figure 4.7).

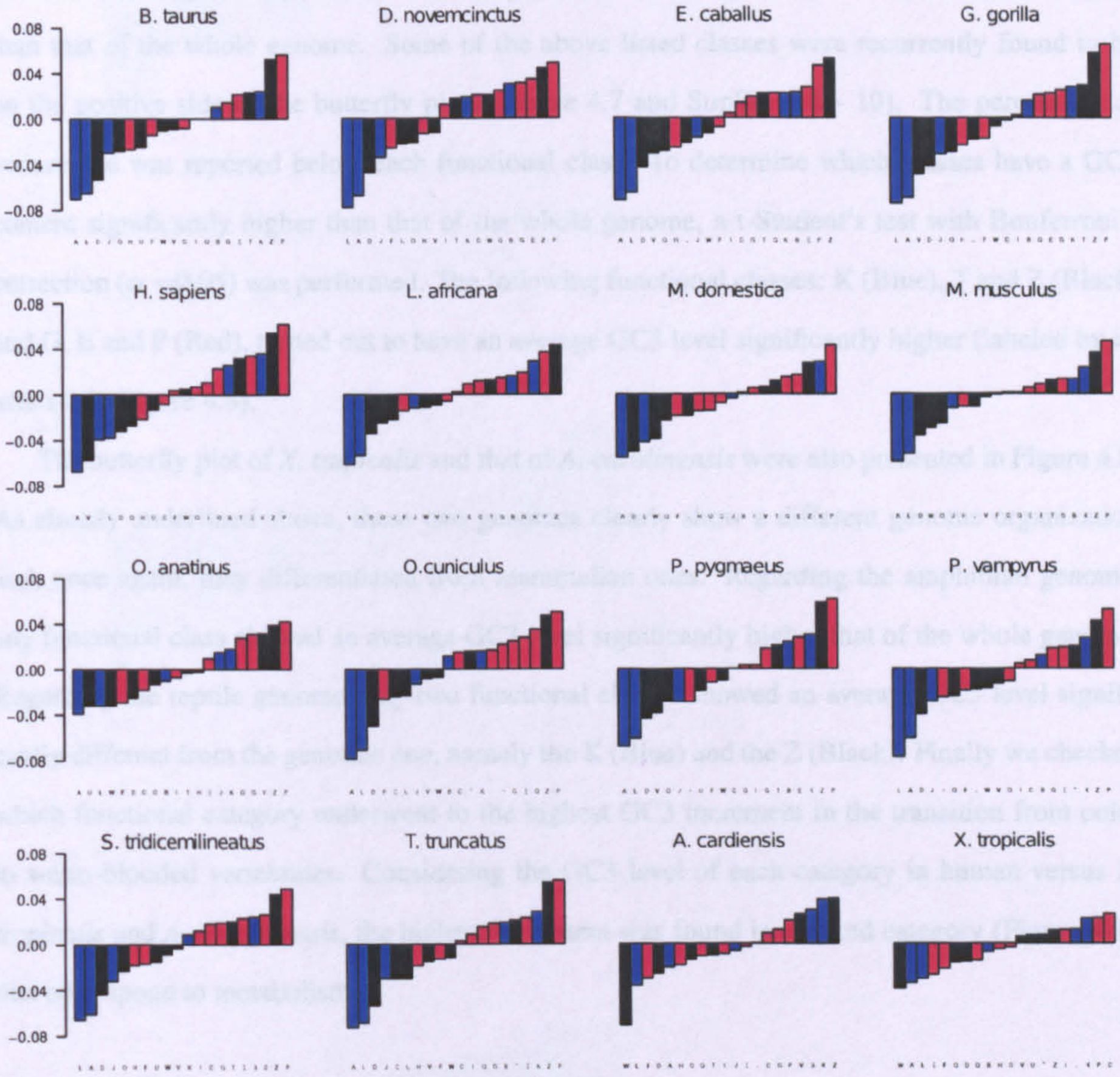


Figure 4.7: Histograms of the delta between average genomic GC3 level against that of each functional class within each genome. Color code as in Figure 4.6

4.2.7 Mammalian versus amphibians and reptiles

The human genome could be considered as representative of all mammals so far analysed and the corresponding butterfly plot was reported in Figure 4.8 (top panel). In the human genome the functional classes B and K (two over five Blue classes), the W, T and Z (three over seven Black classes), and F, Q, I, G, E and P (five over eight Red classes) showed an average GC3 level higher than that of the whole genome. Some of the above listed classes were recurrently found to be on the positive side of the butterfly plots (Figure 4.7 and SupFigs. 6 - 10). The percentage of occurrence was reported below each functional class. To determine which classes have a GC3 content significantly higher than that 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 level significantly higher (labeled by an asterisk in Figure 4.8).

The butterfly plot of *X. tropicalis* and that of *A. carolinensis* were also presented in Figure 4.8. As already underlined above, these two genomes clearly show a different genome organization and, once again, they differentiated from mammalian ones. Regarding the amphibian genome, any functional class showed an average GC3 level significantly higher that of the whole genome. Regarding the reptile genome only two functional classes showed an average GC3 level significantly different from the genomic one, namely the K (Blue) and the Z (Black). Finally we checked which functional category underwent to the highest GC3 increment in the transition from cold- to warm-blooded vertebrates. Considering the GC3 level of each category in human versus *X. tropicalis* and *A. carolinensis*, the highest increment was found in the Red category (Figure 4.9), that correspond to metabolism.

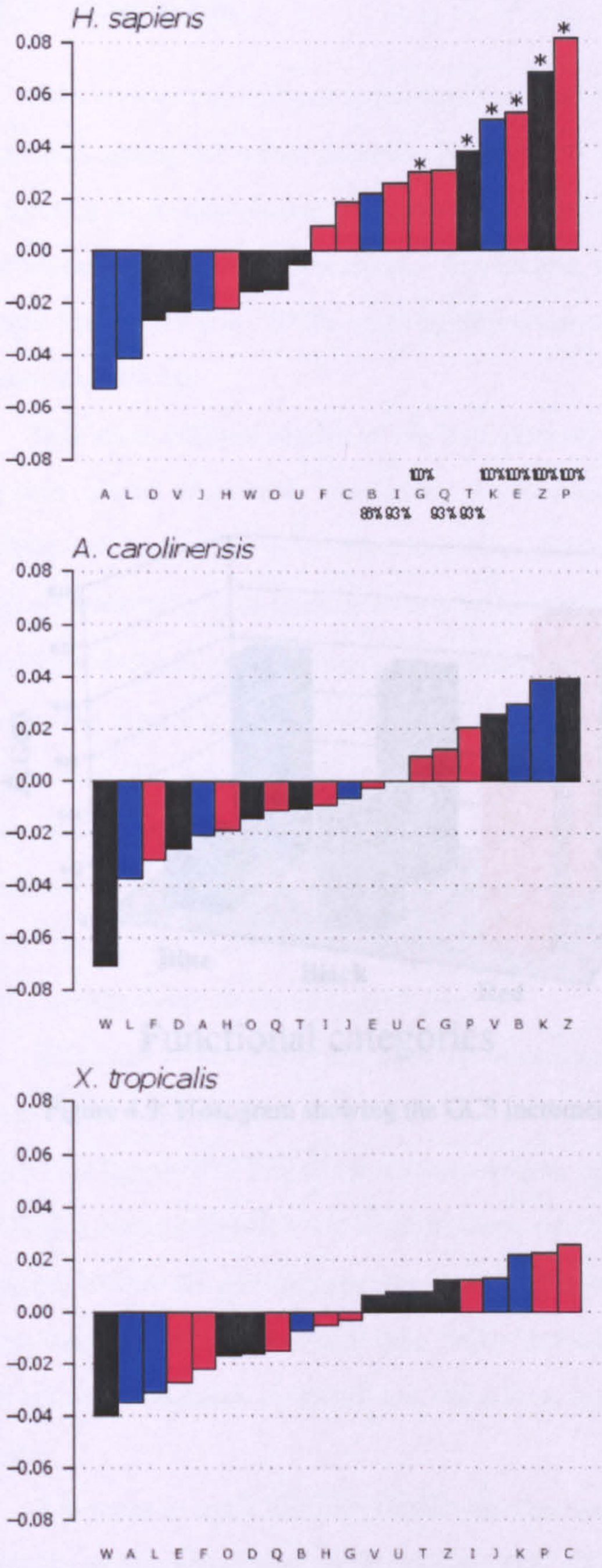


Figure 4.8: Distribution of functional classes within mammalian and non-mammalian genomes. (*) represent those categories with GC3 level significantly higher than that of the whole genome

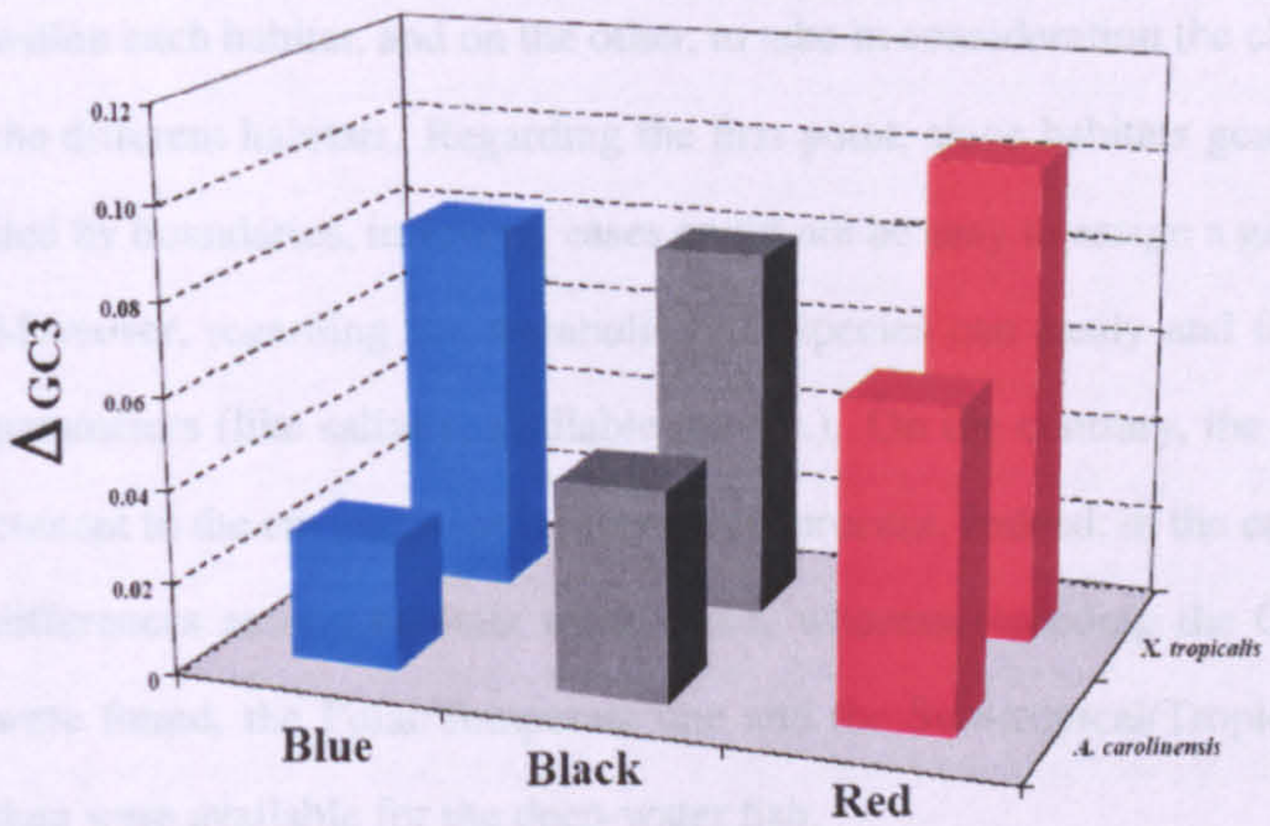


Figure 4.9: Histogram showing the GC3 increment for each human functional category

4.3 Discussion

In this chapter two different approaches were shown to tackle the problem of the GC content variation, among and within genomes. In the case of fish two databases were crossed. The first, collecting all the data produced by several laboratories about the oxygen consumption in several teleostean species (www.fishbase.org). The second, in reality not an interactive database but a paper (Bucciarelli et al., 2002), reporting the measurement of the average genomic GC content in more than 200 fish.

Data about different species of fish were grouped according to their habitats. Using this approach, instead of a simple correlations, it was possible, on one side, to detect some outliers within each habitat, and on the other, to take in consideration the chemical-physical parameters of the different habitats. Regarding the first point, since habitats generally speaking, are not delimited by boundaries, in several cases could not be easy to assign a given species to one habitat only. Moreover, regarding the metabolic rate, species can easily and fastly adapt to different aquatic parameters (like salinity, available 0_2, etc.). On the contrary, the adaptation of the genomic GC content to the environment is a very slow process. Indeed, in the case of metabolic rate significant differences among habitats were found, whereas regarding the GC content two "main blocks" were found, the Polar/Temperate one and the Sub-tropical/Tropical one. Unfortunately no GC data were available for the deep-water fish.

In Chapter 2 was already discussed the evolutionary hypotheses proposed to explain the GC variation among genomes. Regarding fish, unfortunately no data about recombination rate or about the hypothetical bias of the replication/repair machinery are available. However, as matter of fact, a strong correlation was found between metabolic rate and genomic GC content, in good agreement with the genomic analyses carried out in tunicates. The observation that GC content was higher in fish living in "cold" than "warm" habitats suggest that the temperature is not acting in this case, and hence, exclude it as a major force explaining higher GC contents at the genome level.

It is worth to recall here that vertebrates show two different kind of genome evolution: the shifting and the transitional mode (Figure 4.10). The former, mainly found among fish, is, as indicated by the term, a complete compositional shift, like the one observed in bacteria. Com-

positionally far genomes, such as those of zebrafish and fugu (Costantini et al., 2007), showed no overlap at all. The transition mode, was found comparing "cold- and warm blooded- vertebrates" (a terminology used by Bernardi and Bernardi (1986) to stress the fact that, considering temperature, both environmental and body temperature should be taken in to account), mainly amphibians versus mammals. Thus, in mammals an increment of the genome heterogeneity due to an increment of the GC content was shown (see Figure 4.10). The analysis of the functional classes performed according to the KOG classification (Tatusov et al., 2001, 2003), clearly showed that in spite of the different mode of genome evolution, the metabolic rate seems to play the same role in fish and mammals. The result was interesting for the following reasons: i) the temperature effect on the GC content, invoked as factor stabilizing DNA, RNA and proteins, can be excluded, also in warm-blooded vertebrates; and ii) pointed out some critical questions regarding the BGC hypothesis, as well as other hypotheses grounded on random processes. Regarding the first point, in the frame of the thermal stability hypothesis, is very difficult to explain why genes coding for proteins involved in metabolic processes should be stabilized by a GC3 increment, more that those involved, for example, in information storage and processing, since both kind of proteins are affected by the same temperature. Regarding the BGC hypothesis, phylogenetic relationship could explain the very similar compositional pattern found in all mammalian orders. Indeed, the fact that a very similar "butterfly plots" were found in mammals could be simply explained by the existence of a common ancestor. However, two considerations should be done. The first regards the recombination hot spots. Indeed, the analyses of very close genomes, such as those of primates, failed to find conserved hot spot for recombination, reaching the conclusion that they are "highly mobile" and therefore not phylogenetically related (Huang et al., 2005) The second regards the fact that random processes could not easily explain the differences appearing in a specific subset

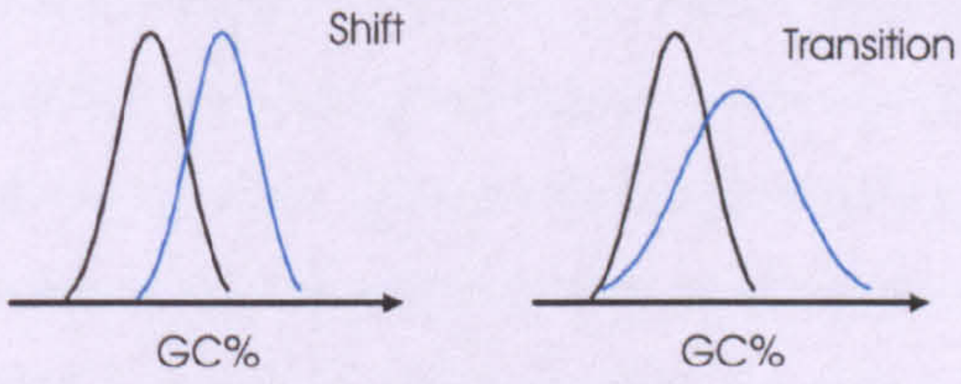


Figure 4.10: Representation of compositional transitions or shift

of functional genes (Berná et al., paper in preparation). Finally, it is worth to stress that, the GC of coding sequences, especially the GC3, was always found to be higher than the GC content of the surrounding non-coding regions (Aota and Ikemura, 1986; Assani et al., 1991). This feature, lastly found in ascidian genomes, most probably is shared by all eukaryotic species. The different GC content between coding and non-coding regions was tentatively explained by the occurrence in the latter, GC-poor transposable elements (Duret et al., 2002). The hypothesis was rejected in yeast (Birdsell, 2002) and was not holding also in human (D'Onofrio, unpublished results). Regarding this point further investigation should be done in ascidian genomes.

4.4 Conclusions

Here two different approaches were carried out to shed some light on the compositional genome evolution, trying to understand which are the forces driving it. The results presented here reinforce the metabolic rate hypothesis as an important factor shaping genome composition. Indeed in fish, a significant correlation was found between metabolic rates and genomic GC levels. Furthermore, in all mammalian genomes so far analyzed turned out that: genes involved in metabolic processes showed the highest GC level at the third codon positions, and underwent to the higher GC3 increment in the genome transition from cold- to warm-blooded vertebrates.

5 | *General Conclusions*

After more than thirty years the problem of the variation of base composition among genomes is an open question and still matter of debates in the neutralist/selectionist frame. Several forces have been proposed to drive the changes of base composition at coding and non-coding regions, the bias gene conversion (BGC) (Eyre-Walker, 1993; Galtier et al., 2001; Duret and Galtier, 2009), the breakpoints distribution (Lemaitre et al., 2009); the thermal stability (Bernardi (2007) for a review), and the metabolic rate (Vinogradov, 2001, 2005). The aim of the present thesis was to determine if the metabolic rate hypothesis could affect the base composition of genomes. In this regards, the GC content was tested along the phylogenetic line of deuterostomes performing different strategies and approaches dictated by the available data.

The first one was applied to tunicate organisms. A phylogenetically important group since they were found to be sister branch of vertebrates. Several genomes of this group are available nowadays. Our work undoubtedly showed that an exuberant molecular evolutionary rate distinguish tunicates from vertebrates. More precisely *C. intestinalis* and *C. savignyi* turned out to be 1.5 times faster than vertebrates, whereas the recently sequenced *O. dioica* was even faster, being at least three times faster than vertebrates, and most probably one of the fastest metazoans. The reliable estimation of speed lied on the fact that: i) the analyses were performed at genome level, that is, using a great amount of sequences to reduce as much as possible erratic errors inherent to algorithms; and ii) different vertebrates and out-groups were used in order to avoid species-specific effect in multiple comparisons. Therefore the ending results described a more general aspect of chordate evolution. Calibrating the vertebrate molecular clock by fossil records, the time of divergence between *C. intestinalis* and *C. savignyi*, was estimated to be nearly 180 My, a huge amount of time for species belonging to the same genus and almost morphologically indistinguishable. However, taking into account the high mutation rate found in both species, the

estimated time of divergence was quite realistic. Moreover, studying the base composition of these ascidians significant differences were found, especially at the third codon position. The analysis of the possible causes of this disparity showed that: i) compositional differences were not due to a CpG methylation/deamination process, producing a CpG shortage; ii) the common ancestor, most probably, was closer to *C. intestinalis* than *C. savignyi*, since the compositional differences were due to a GC increment that took place in the latter; iii) the fact that GC increments were higher in coding than in non-coding regions questions about the evolutionary forces driving the genome base composition, since a simple mutational bias could not account for higher increment at the third codon positions than in intergenic regions, since both are close to neutrality; and iv) the different GC3 level found between the two species, most probably can be ascribed to a different transcriptional level, suggesting that a different metabolic rate could be one of the factors driving base compositional evolution.

The second was applied to fish genomes. Thanks to the fact that metabolic rate and genomic GC content were available for more than 200 species, a detailed study comparing fish living in different habitats was carried out. The results clearly showed that significant differences subsisted among habitats. More precisely, polar fish were characterized by the highest metabolic rate and by a high GC content. The two variables turned out to be significantly correlated, being GC content increasing as increasing metabolic rates.

Finally, analyzing 13 mammalian genomes, the link between metabolism and GC content was further supported. In this analysis the CDSs of each genome were classified in different functional classes according to the KOG database (Tatusov et al., 2001, 2003). In each mammalian genome the average GC3 level of the genes involved in "metabolic processes" was the highest.

As mentioned in Chapter 2, several hypotheses have been proposed to explain the forces driving the base composition evolution at genomic level: the bias gene conversion (BGC) (Eyre-Walker, 1993; Galtier et al., 2001; Duret and Galtier, 2009), the breakpoints distribution (Lemaitre et al., 2009); the thermal stability (Bernardi (2007) for a review), and the metabolic rate (Vinogradov, 2001, 2005). The former three hypotheses failed to explain adequately the results obtained using the different approaches described above.

Specifically, the BGC hypothesis could hardly explain: i) the opposite relationship between recombination rate and GC content found comparing *C. intestinalis* and *C. savignyi*, since the

latter showed a lower recombination rate and a higher GC content. It is worth to say that this negative relationship is itself an interesting result, that could be further analyzed. However in the frame of present thesis this is representing a weak point of BGC hypothesis towards explaining the observed compositional differences in tunicates; as well ii) the observation that in all mammalian genomes so far analyzed genes involved in "metabolic processes" showed the highest GC3 level, since no conservation of recombination hot spots was found in phylogenetically closer species (Huang et al., 2005).

Regarding the breakpoints distribution, a discrepancy was found between the fact the GC increment in ascidians took place mainly in coding than non-coding regions, whereas breakpoints have been reported to be under-represented in genes.

Finally, regarding the thermal stability hypothesis: i) no evidences of different environmental temperature among the habitats colonized by *C. intestinalis* and *C. savignyi* were found; ii) fish living in polar/temperate habitats showed high genomic GC content and high metabolic rate as well; and iii) why proteins related to metabolic process have to be more stabilized by the increment of GC than those involved, for example, in the pathway of "information storage and processing". On the other hand, the observations: i) that a strong correlation was found between metabolic rate and genomic GC content in fish; ii) that mammalian genes involved in metabolic process showed higher GC3 levels; and iii) that on the first axis of CoA in *C. savignyi* a correlation with ΔGC3 and ESTs values co-occurred, all converged towards the conclusion that metabolic rate most probably play a crucial role among the forces driving the genome evolution.

The measurement of the O_2 consumption in tunicates, as well as in more teleostean fish, would be of great support for the results discussed in the present thesis.

6 | Appendix 1

Materials and Methods

6.1 Sequences

6.1.1 Coding sequences and ESTs

Coding sequences of *C. intestinalis* were retrieved from the databases JGI -version 1.0 2002- (http//genome.jgi-psf.org), and from ENSEMBL -version:48.2h.- (http//ftp.ensembl.org).

From the Ensembl database coding sequences of the following species (in alphabetical order) were retrieved : *B. taurus*, *C. savignyi*, *D. novemcinctus*, *E. caballus*, *G. gorilla*, *H. sapiens*, *L. africana*, *M. domestica*, *M. musculus*, *O. anatinus*, *O. cuniculus*, *P. vampyrus*, *P. pygmaeus*, *R. norvegicus*, *S. tridecemilineatus* and *T. truncates*, (www.ensembl.org). Those of *S. purpuratus*, *T. rubripes* and *G. gallus* were retrieved from NCBI (www.ncbi.nml.hih.gov). Coding sequences of *X. laevis* and *X. tropicalis* were retrieved from www.xenbase.org, those of *B. floridae* from JGI, while those of *O. dioica* from Genoscope (www.genoscope.cns.fr).

Expression sequence tags (ESTs) of *C. intestinalis* and *C. savignyi* were downloaded from NCBI (ftp.ncbi.nih.gov/repository/UniGene).

Redundant sequences were identified and removed by Cleanup (version 1.8.1) program (Grillo et al., 1996), using a standard threshold of 95% of identity and 80% of overlapping with other sequences. Repeatmasker was used to identify interspersed repeats and low complexity DNA sequences. (Version: open-3.2.6) (Smit and Hubley, 1996).

6.1.2 Fossil record data

The time of divergence among vertebrates estimated from the fossil records, was collected from the literature (Benton and Donoghue, 2007).

6.1.3 Human KOG sequences

Proteins belonging to clusters of orthologous groups were retrieved from the KOG database (www.ncbi.nlm.nih.gov/KOG/), and the corresponding coding sequences from NCBI using a batch Entrez function. As described in KOG database (Tatusov et al., 2001, 2003) functional classes, denoted by capital letters in square brackets, were grouped in three large categories, namely: (i) information storage and processing; (ii) cellular processes and signaling; and (iii) metabolism. For sake of simplicity, the three categories were denoted as Blue, Black and Red, respectively. Proteins with more than one classification (KOG id) and functional classes with less than 100 proteins in human, as well as poorly characterized classes were removed from the dataset.

6.2 Primary annotation of Ciona

Homologs for the gene models of *C. intestinalis* and *C. savignyi* were searched using Blastp software (Altschul et al., 1997) against sequences annotated in databases publicly available. The databases used were: NCBI-nr database (5/28/08), Swiss-Prot (manually annotated and reviewed), and TrEMBL (automatically annotated but not reviewed) (UniProtKB/TrEMBL release 14.0, www.uniprot.org). A filter with a threshold at an e-value $< 1x10^{-10}$ was used. For each sequence the best hit (higher score and higher identity) was selected and their information was acquired by the sequence.

GO annotation was performed by Blast2GO software (Conesa et al., 2005) with the following parameters: 1) Blastp vs NCBI nr database. E-value lower than $1x10^{-03}$, HSP (High Scoring Segment Pairs) length cut-off 33% and less than 20 hits retrieved; 2) Four different mapping process to obtain gene ontology information; 3) Annotation process. E-value lower than $1x10^{-06}$. HSP must cover at least 20% of the longitude of its hit. Annotation cut-off 55 (Maximum similarity weighted by a factor corresponding the best belonging GOs evidence code).

6.3 Orthologs

Orthologs gene pairs were identified by a Perl script, performing reciprocal Blastp (Altschul et al., 1997) and selecting the Reciprocal Best hit (RBH). Applying this strategy it is possible to avoid to use proteins over represented in the data set, in other words, truly orthologs are identified, not considering alignments that correspond to paralogous sequences. Using this procedure orthologs between two species (pairs orthologs) were identified. In order to obtain orthologs between three species (triplets), the orthologous sequences for the first species were used to determine (with the same procedure) the orthologs to a third one. It should be considered that the above procedure do not imply that each member is RBH-connected with every other member in the triplets, but at least with one. The same process was used to obtain orthologs between four species (quartets). These groups of orthologs were used to carry out a variety of analyses:

i) To the ascidian comparative analysis, 7747 orthologous protein pairs between *C. intestinalis* and *C. savignyi* were identified.

ii) To establish a vertebrate's molecular clock, different couples of vertebrates were utilized. Orthologous pairs were constructed between *H. sapiens* and another vertebrate species, namely: *M. musculus*, *M. domestica*, *O. anatinus*, *B. taurus*, *G. gallus*, *X. laevis* and *T. rubripes*. Moreover, orthologs of *X. laevis* and *X. tropicalis*, as well as *M. musculus* and *R. norvegicus* were performed.

iii) With the purpose of comparing substitution rates of ascidians and vertebrates, orthologous genes shared by *C. intestinalis*, one vertebrate and an outgroup (amphioxus or sea urchin) were obtained.

iv) To study in depth this phenomeno in Urochordata, quartets of orthologous sequences were constructed, containing two urochordates (*O. dioica*, *C. intestinalis* or *C. savignyi* and two outgroup species (Amphioxus and one vertebrates (i.e. *B. taurus*, *M. musculus*, *H. sapiens* and *X. tropicalis*).

v) Comparisons of orthologous sequences between KOG-classified human genes and thirteen representative mammals (belonging to primates, rodents, laurasiatheria, afrotheria, xenarthra,

marsupials and monotremes) were performed. Each mammalian genome acquired the same KOG classification of the corresponding human protein.

6.4 Membrane proteins identification

Membrane proteins were identified through the gravy score calculated for each sequence with CodonW (1.4.4). Sequences with a gravy score equal or higher than 0.45 were considered integral membrane protein according to Lobry (1997).

6.5 Base Composition

6.5.1 Composition in *C. intestinalis* and *C. savignyi*

CodonW (1.4.4) was used to calculate the molar ratio of guanine plus cytosine (GC) of the whole genome (GCg), as well as that of both non-coding and coding regions. More precisely, the GC content of introns (GCi), flanking regions (5'GC and 3'GC), coding sequences (GCcds) and that of each coding positions (GC1, GC2 and GC3) were calculated. Several compositional analyses were restricted to the orthologous sequences between *C. intestinalis* and *C. savignyi* calculating the average values of: i) the molar ratios of guanine and cytosine at the third codon position, G3 and C3 respectively; ii) the frequencies of CpG in the three reading frames (C1pG2 C2pG3 C3pG4); and iii) di-nucleotides frequencies produced after the deamination of 5-methyl-cytosine (5mC), namely, CG, TG, CA, TA. In order to determine the statistical significance of the compositional differences between ascidians genomes, a Shapiro's test was conducted to verify the normality of the data, and thereafter a Welch's t-Test was performed, using the R software.

6.5.2 GC3 of KOG classes

In order to determine the statistical significance of the differences in GC3 content between the three main KOG categories of genes (Blue, Black and Red), a two-sided Mann-Whitney test was performed. For each species 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 (α =0.05) for multiple-comparisons.

6.6 Codon usage, Correspondence Analysis (CoA) and Amino acid frequencies

The codon usage pattern in *C. intestinalis* and *C. savignyi* and the Relative Synonymous Codon Usage (RSCU Sharp and Li (1986)) were computed by CodonW (1.4.4) for all available genes in both genomes. The same analysis was performed for the orthologous sequences between both species. The RSCU value for a codon *i* is defined as follows:

$$RSCU_i = \frac{Obs_i}{Exp_i} \tag{6.1}$$

where Obs_i is the observed number of occurrences of codon *i*, and Exp_i is the expected number of occurrences of the same codon for a uniform synonymous codon usage. In the absence of any codon usage bias, the RSCU values would be 1.00. Codons used less or more frequently than expected will have an RSCU value lower or higher than 1.00. A χ^2 contingency test was performed to detect statistical significance between codon frequencies among species or different group of genes.

In order to determine the major source of variation among genes, CoA on RSCU values, as well as, on codon usage were carried out for orthologous sequences between *C. intestinalis* and *C. savignyi*. In CoA on RSCU, each sequence is described by a vector of 59 variables, which is the number of codons for which there are synonyms. CoA plots these genes in a multidimensional space of 59 dimensions. Therefore, a certain number of new axes that represent the most prominent factors of variation are identified.

CodonW (1.4.4) was also used to calculate the total amino acid frequencies of the genome of *C. intestinalis*, *C. savignyi* and *O. dioica* as well as for their correspondent orthologous sequences.

6.7 Sequence alignment and distance calculation

Protein orthologous pairs, trios and quartets were aligned by ClustalW (Larkin et al., 2007). To calculate the amino acid distance the program Fprotdist (Felsenstein, 2004) with the Jones-Taylor-Thornton model (JTT) was used (Jones et al., 1992). In order to avoid bias distance estimations, due to a substitution saturation effect, pairwise distances > 1 (except those related to *O. dioica*) were disregarded for further analysis.

For the specific case of quartets of orthologous sequences (with *O. dioica* and *C. intestinalis* and two outgroup species) the contents of GAPs were calculated and removed through a perl script. Filtered groups of alignment were defined as those alignments with less than 60 percent of GAP and distances < 1 except those that involve *O. dioica* in order to select those with higher homology and, to avoid bias due to a substitution saturation effect.

6.8 Relative Rate Test

A relative rate test compares the relative substitution rates between two species (since their common ancestor) using as a reference a third species (outgroup) branching off earlier than the species to be tested. In particular, the Tajima's test compares the number of substitutions between species 1 and outgroup (N13) with the number of substitutions between species 2 and outgroup (N23). The comparison is restricted to those sites in which both sequences 1 and 2 are different. The null hypothesis is that N13 = N23. The statistical significance of the difference between these two sequences is assessed by the chi-square (χ^2) test with one degree of freedom (Tajima, 1993). A Perl script was developed to conduct the Tajima's test (Lamolle http://lib.bioinfo.pl/auid:4696352) to determine:

i) whether the same molecular clock applies to all chordates (namely Ciona, amphioxus and vertebrates), that is, if chordates evolve at the same rate at the molecular level. This procedure was performed for each one of the most representative vertebrates for which their genomes were available (human, mouse, frog, opossum, platypus, cow, chicken and fish).

ii) if the two Ciona species evolve at the same velocit, in this case *C.intestinalis* and *C. savignyi* were compared using amphioxus as an outgroup.

iii) to test if ascidians and appendicualarians evolve at the same rate at the molecular level (*C.intestinalis* and *O. dioica*), using different outgroups: amphioxus, sea urchin, cow and human.

6.9 Estimation of acceleration rate

In order to calculate the rate of acceleration between two organisms (1 and 2), their respective distances to a common ancestor (a and b) were determined (Sarich and Wilson, 1973; Li et al., 1981). Let be: i) d12 = a + b, the distance between the two organisms to be tested; ii) d13 = a + D1 + D2, the distance between one organism (1) and the outgroup; and iii) d23 = b + D1+ D2, the distance between the other organism (2) to be tested and the outgroup (Figure 6.1). The rate of acceleration, using different outgroups is estimated by a/b. Branch lengths were determined using the program Fprotdist, with the Jones-Taylor-Thornton model (JTT) (Felsenstein, 2004).

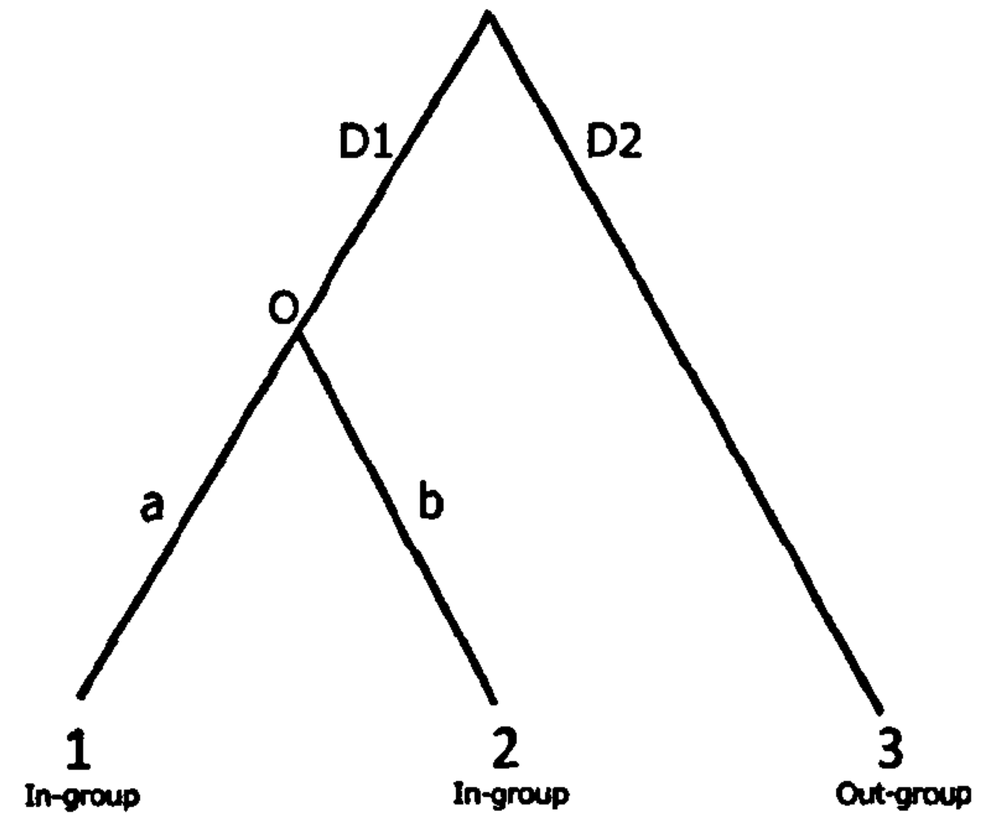


Figure 6.1: Schematic representation of branch distances to calculate molecular rate of evolution

The rate of acceleration was calculated to different species. Using amphioxus or sea urchin as outgroup:

i) *C. intestinalis* - one vertebrates (from Table 3.3)

Using amphioxus or one vertebrate (mentioned before) as outgroup:

i) *C. intestinalis - C. savignyi.*

ii) *C. intestinalis - O. dioica.*

6.10 Amino acid transition matrices

The aligned groups of orthologus sequences (*O. dioica*, *C. intestinalis*, amphioxus and *B. taurus*) were divided in four groups according to their divergence, namely slow evolving, intermediate evolving, fast evolving, and very fast evolving. For each group of four-aligned sequences the amino acid substitution matrices were computed for *O. dioica* and *C. intestinalis*.

To this aim first two groups of matrices of counted amino acid substitutions were built, namely one accounting for those amino acid substitutions that took place in the branch leading to *O. dioica* and another in that leading to *C. intestinalis*. For this purpose the ancestral sequence (i.e. the node) must be inferred, since the matrices are polarized. This means that each entry M_{ij}, in the matrix represents the number of substitutions from amino acid i to amino acid j. This in turn implies that one or more outgroups are required. The inference was done by both maximum parsimony and maximum likelihood. The amino acid substitution matrices were constructed taking into account only the informative sites and storing them in a 20 x 20 counter matrix. Sites where the amino acid of both outgroup differed from the amino acid of both in-groups were ignored, because it is not possible to infer the ancestral state, as well as, those sites where the pair of in-group have the same configuration of the corresponding pair of outgroups. The amino acid equilibrium state were performed raising the matrix to the power 1000, using Matlab (2007a, The MathWorks).

6.11 Phylogenetic analyses

3220 aligned orthologs of *O. dioica*, *C. intestinalis*, *B. taurus* and *B. floridae*, and a "filtered" subset of 588 were concatenated and analyzed using maximum-likelihood approaches of Phyml program (Guindon and Gascuel, 2003).

6.12 Metabolic rate and genomic GC in fish

Information of taxonomic classification, geographical distribution and metabolism of teleostean fish were downloaded principally from www.fishbase.org and from the available literature. Data of metabolic rate that was obtained applying any kind of stress was discarded (*e.g.* in hypoxia, feeding or starvation).

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 et al., 2001) according to the following equation: $MR = MR_O e^{E/kT}$, 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$\sim$ 0.65 eV, and k is the Boltzmann's constant equal to $8.62\text{x}10^{-5} eV K^{-1}$.

Taking in 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, namely 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.

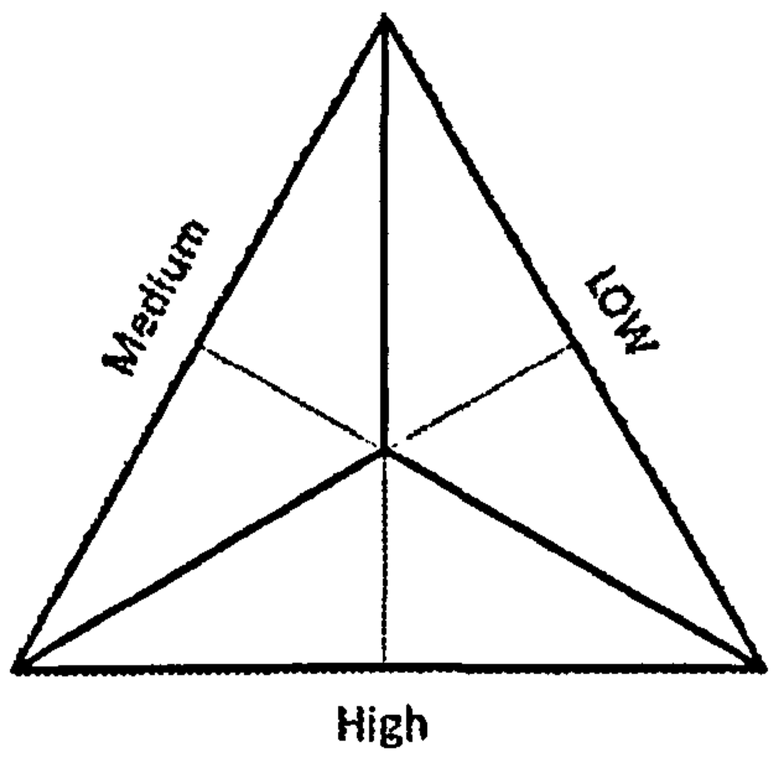
The genomic GC levels of 149 teleostean fish were retrieved from current literature (Bucciarelli et al., 2002; Varriale and Bernardi, 2006), represented by 9 polar, 22 temperate, 48 subtropical and 70 tropical teleostean fish. No data were available about the genomic GC levels of fish living in the deep-water habitat. The outlier 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, 1984). Statistical significance of pairwise comparisons was assessed by the MannWhitney U test.

6.13 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 compare such behavior across different organisms $g = 1, ..., G$, 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 the vector (C_g^L, C_g^M, C_g^H) containing the normalized frequency for the corresponding functional classes in the three levels was defined. Clearly $c_g^i \geq 0$ and $\sum_{i \in (Low, Medium, High)} c_g^i = 1$ for all organisms g. Since each vector (C_g^L, C_g^M, C_g^H) can be represented as a point (whose color corresponds to the category) in a de Finetti diagram, each organism can be coded in the diagram using three colored points drawn in correspondence of its $(C_g^L, C_g^M, C_g^H)_{c \in (blue, black, red)}$ values. The de Finetti 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, it is used for comparing the GC3 compositional/spatial distribution between the categories in different organisms. Hence, to understand its meaning in this context the Viviani's theorem is recalled, 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 this context is set to 1). According to such results each (C_g^L, C_g^M, C_g^H) 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, the three equivalent triangular sectors can be defined (each of one identified by the Low, Medium and High line, respectively) see Figure 6.2 for details.

Note that points (C_g^L, C_g^M, C_g^H) in each sectors have different relational ordering and hence show different CG3 abundance. In particular, categories belonging to a given triangular sector show minimal presence of that CG3 level with respect to the other levels. While, if six triangular sectors are defined by the altitudes then an ordering relation can be defined on the points (C_g^L, C_g^M, C_g^H). Observing that, in absence of association between the GC3 distribution and the functional categories, each configuration of the vector (C_g^L, C_g^M, C_g^H) is equally likely and, as a consequence, the different sectors are expected to be equally represented. Discrepancy from

Figure 6.2: diagram



such uniform distribution denotes a specific association. To measure such effect, first, due to the finite number of classes in each category only a finite number of configurations result potentially attainable (at most 21 for the Blue category, 45 for the Red and 36 for the Black). Such configurations are invariants with respect to $2/3\pi$ rotations of the triangle. However, due to the fact that in this construction the whole GC3 range $[a_g, b_g]$ is organism's specific the observed $(C_g^L, C_g^M, C_g^H)_{c \in (blue,black,red)}$ are not independent (by construction at least one class should be present either in the low and in the high levels) hence the number of admissible configurations results less than the potentially attainable. The $(C_g^L, C_g^M, C_g^H)_{c \in (blue,black,red)}$ for each organism g are shown in Figure 4.6 where the value of each points represents the number of times the vector (C_g^L, C_g^M, C_g^H) has been observed in the different organisms $g = 1, ..., G,$. 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., in particular to the triangular sector with the Low level as base 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, B class permutations among the categories and each time counted k_i as the occurrence of the Red category in the Low triangular sector was performed, then the p-value was estimated considering the sector as $\frac{\sum_{i=1}^{B} I(K_I \geq k^*)}{B}$ where B denotes the number of permutations and k^* was the observed occurrence of the Red class in the Low triangular sector on true configuration.

7 | Appendix 2

Data bases and annotation

This section describes in details the characteristics of the *Ciona intestinalis* genome, as well as the primary annotation of both *C. intestinalis* and *C. savignyi* genomes. This process was important because allowed to perform different analyses and obtaining some of the interesting results showed in the present thesis, particularly those related to understand biological function of particular sets of genes.

7.1 Data base selection

Before to begin with any type of analysis, the most important step was to establish a reliable data base of sequences on which ground all subsequently steps. For *C. intestinalis* two genome databases were available, both belonging to the same sequence project, namely the one carried out by Joint Genome Institute (JGI) some years ago. In order to select an accurate dataset of unique sequences, several studies to determine the advantages and disadvantages of each database were performed. Apart JGI, another database of *C. intestinalis* sequences was available: the ENSEMBL database (www.jgi.doe.gov and www.ensembl.org). Both are based on the whole genome shotgun method released by the JGI, and have gene model sequences identified by a pipeline annotation, predicting gene models by both, statistical method and experimental approach, such as EST sequences. Only the first version of JGI (version 1.0 2002) had some identification of the gene models products. ENSEMBL built additional gene models using the assembly of JGI (v. 2.0) and data supported by cDNA and EST sequences, as well as data from other species. Gene models, nucleotide and amino acid sequences were downloaded from JGI (v.1.0 2002) and

Ensembl (release 57), with 15.852 and 19.857 sequences respectively. EST from NCBI-EST (ftp.ncbi.nih.gov 04/2008) database were also retrieved, in order to perform reliable comparison supported by experimental data. Indeed, only a gene model supported by an EST sequence can be considered a reliable expressed gene.

With the purpose to assess the possible errors or missing data, and to establish the accuracy of each database, the following analyses were carried out. A summary of the results are presented in Table 7.1.

i) **NCBI hits.** The amount of protein sequences from the two databases occurring in NCBI was searched by running Blastp. The amount of protein sequences showing at least one match in NCBI was larger for ENSEMBL than JGI, respectively 15.687 and 12.345 sequences.

ii) **EST support.** The number of ENSEMBL and JGI sequences supported by at least one sequence in the EST database was searched by running Blastn. About 96% of gene models from ENSEMBL have a support in the EST database (19011 sequences). Instead, JGI gene models had only 88% sequences supported by at least one sequence in the EST database (13897 sequences).

iii) **Repeatmasker.** To asses to the amount of sequences carrying interspersed repeats, low complexity or transposable elements, the software RepeatMasker (Smit and Hubley, 1996) was performed. The results obtained showed that sequences carrying repeats were less frequently occurring in ENSEMBL than in the JGI database.

iv) **Blast comparison.** Finally, the comparison of the two databases running reciprocal blastp was done. The result establish that 13404 sequences from JGI have an homolog in ENSEMBL database, meanwhile 14908 from ENSEMBL have it in JGI. The difference indicate that some genes that were identified as entirely coding sequences for JGI, are considered two different models, that codify for two different proteins for ENSEMBL.

Considering that: i) ENSEMBL generate gene models from JGI assembly but using a different pipe-line annotation; ii) 79 % of their gene models have an homolog in NCBI database, a number

	ENSEMBL	JGI
Totalsequences	19.857	15.851
Blastp vs NCBI	15.687 (79%)	12.345 (78%)
EST support	19.011 (96%)	13.897 (88%)
Repeats	1.322 (7%)	1.888 (12%)
ENSEMBL vs JGI	14908 (75%)	13404 (84%)

Table 7.1: Data base analysis of *C. intestinalis* genome.

that correspond to 15687 sequences, similar to the total amount of JGI database; iii) 96% of the ENSEMBL gene models present a EST support, versus 88% for JGI; and iv) The amount of interspresed repeats is lower in number and percent for ENSEMBL sequences (1322, 7%) compared with JGI sequences (1888, 12%); ENSEMBL database apper to be more reliable and complete. As a consequence ENSEMBL database was selected, and their sequences were used to perform all the analysis here presented.

7.2 Primary annotation

The previous analyses established that far from representing a complete and accurate annotation, the database contained only fragmentary data, and in many cases the sequences were not at all annotated. Such information is fundamental to perform a comparative analysis and important to understand and identify which genes are absent in *C. intestinalis* genome. Hence, it was necessary to establish with a reasonably degree of confidence and accuracy, "what is what" in the genome. To this aim, a primary annotation was performed using the standard principle of "transfer of information by homology". Consecutively, a gene ontology classification was made, in order to give major support and description to each sequence.

Transfer of information by homology Gene models were searched for homology using Blast software against sequences annotated in databases publicly available (NCBI-nr, Swissprot and TEMBL). The gene models inherit the annotation of the sequence to which it had the best hit. Each protein from a total of 19697 proteins of *C. intestinalis* were compared with the available protein databases using the Blastp software (Altschul et al., 1997). Through this process 15687 (80%) sequences were identified. Each protein sequence was classified as "similar to", i.e. with

a given percent of identity with the sequence found in other organism. At this stage mistakes or misidentification could easily occur. The limitation of this annotation is discussed below. The distribution of the sequence length (bp, base pair) and the identity recovered on the primary annotation are presented in Figure 7.1.

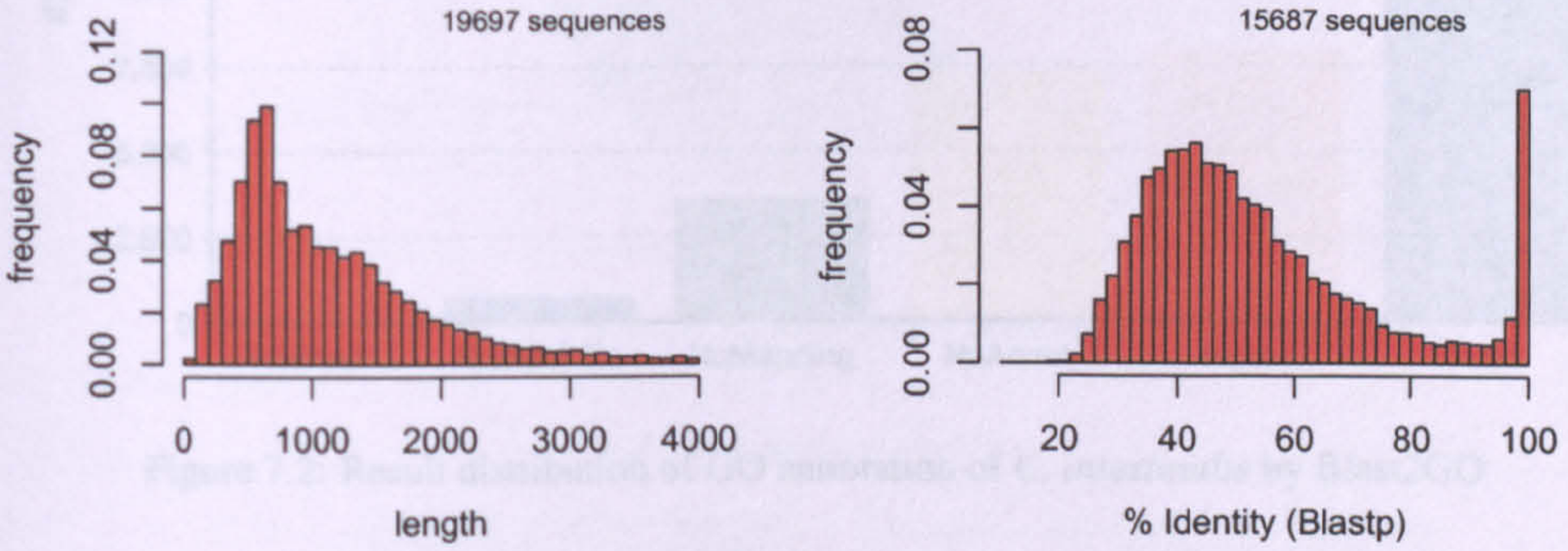


Figure 7.1: Analyses of the sequences of *C. intestinalis*

Gene Ontology Ontologies provide a common and controlled vocabulary for description of the molecular functions, biological process and cellular components of gene products. The hierarchical, and multi parent relations characteristic of the GO structure enable powerful grouping, searching and analyzing genes. It is also useful to compare datasets to have a wide panorama about the different aspect of the whole genome. Particularly, it allow to study certain groups of genes (e.g. metabolic routes). The gene ontology annotation was performed for *C. intestinalis* genome through Blast2GO software and consisted in three basic steps: i) a blast search of the fasta sequences versus an annotated databases; ii) a mapping process to obtain gene ontology information; and iii) the annotation step, where it is possible to introduce personal threshold in order to filter the information obtained and giving each sequence the adequate information. A total 7987 sequences could be annotated. A quantitative summary of these steps were presented in Figure 7.2.

The same procedures were preformed for 20143 sequences of *C. savignyi* retrieved from ENSEMBL. Almost 80% of this sequences could be identified searching for homology in different

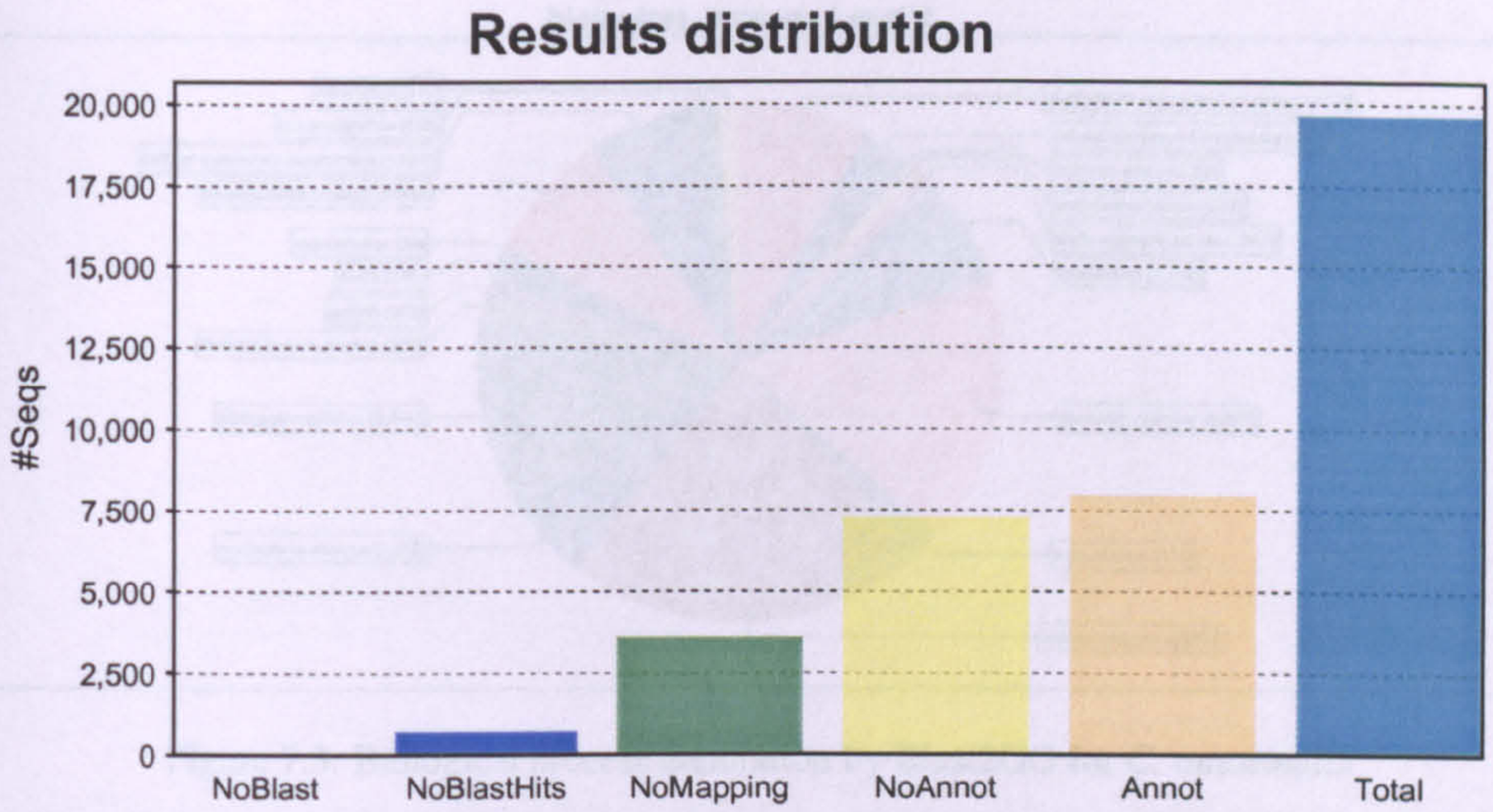


Figure 7.2: Result distribution of GO annotation of *C. intestinalis* by Blast2GO

databases (a total of 16.079 sequences). The gene ontology annotation was also perfomred, and 8100 sequences were annotated (Suplementary Figure (SupFig.) 11).

Regarding to the three descriptive categories of Gene Ontology, similar annotations were found. As an example Figure 7.3 and Figure 7.4 show a pie-plot of GO related to Biological process of level-2. Greater amount of sequences could be annotated for *C. savignyi*, indeed each process is represented by more sequences than those found in *C. intestinalis*. At this level, the same biological process were found, except for 47 sequences related to "viral reproduction" absent in *C. intestinalis*. This similarities are expected since GO annotation is performed on the base of sequence homology. This process, that could be performed with different personalized parameters and databases, has a limitation: the necessity of have an homolog gene/protein annotated, in order to identify the sequence in question.

The pie-plot (level 2) for Cellular component (SupFig. 12, SupFig. 13), and Molecular function categories (SupFig. 14, SupFig. 15) for *C. intestinalis* and *C. savignyi* respectively are presented as supplementary figures at the end of the thesis.

Membrane proteins The major difference between globular proteins and those that are present in membranes (integral membrane proteins (IMP)) is that the latter are enriched in hydrophobic

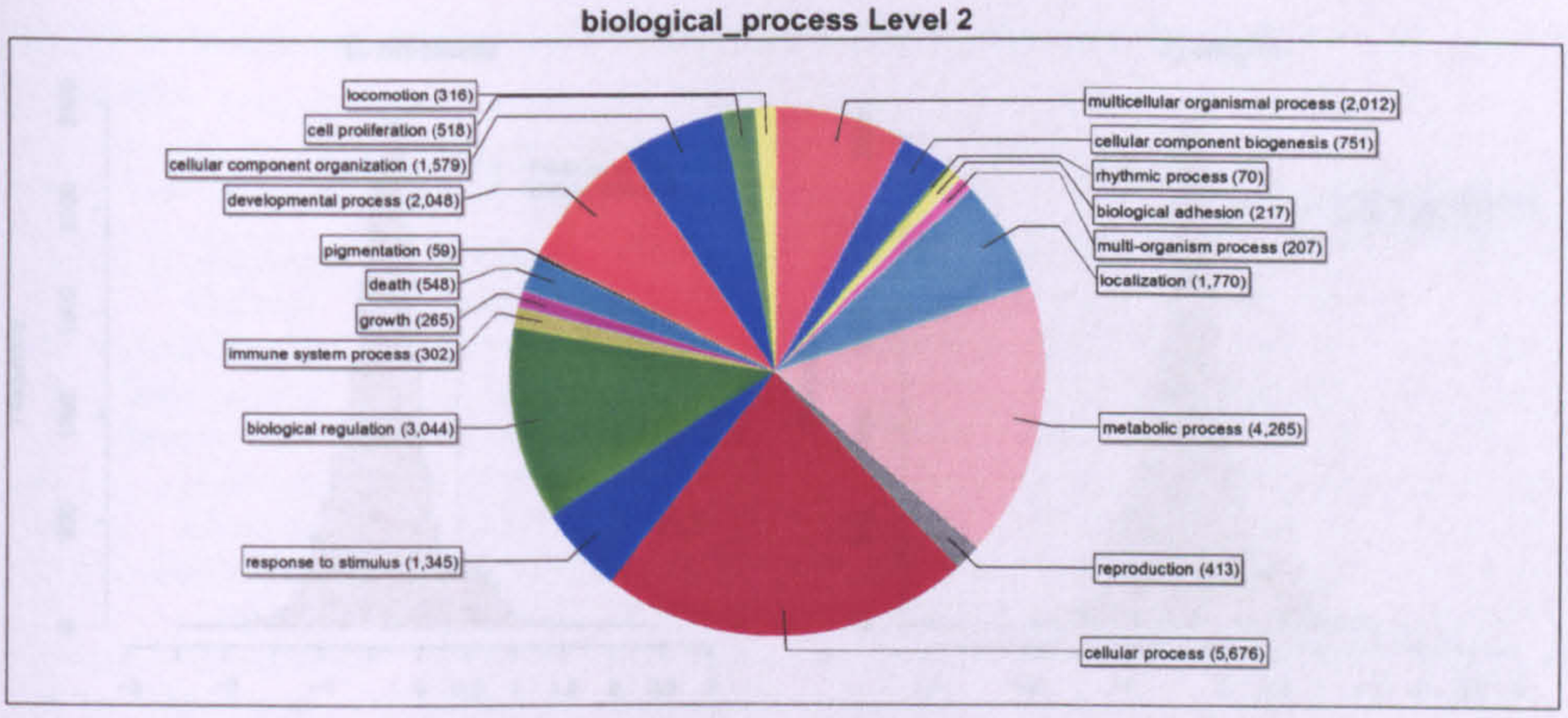


Figure 7.3: Biological process annotation by Blast2GO for *C. intestinalis*

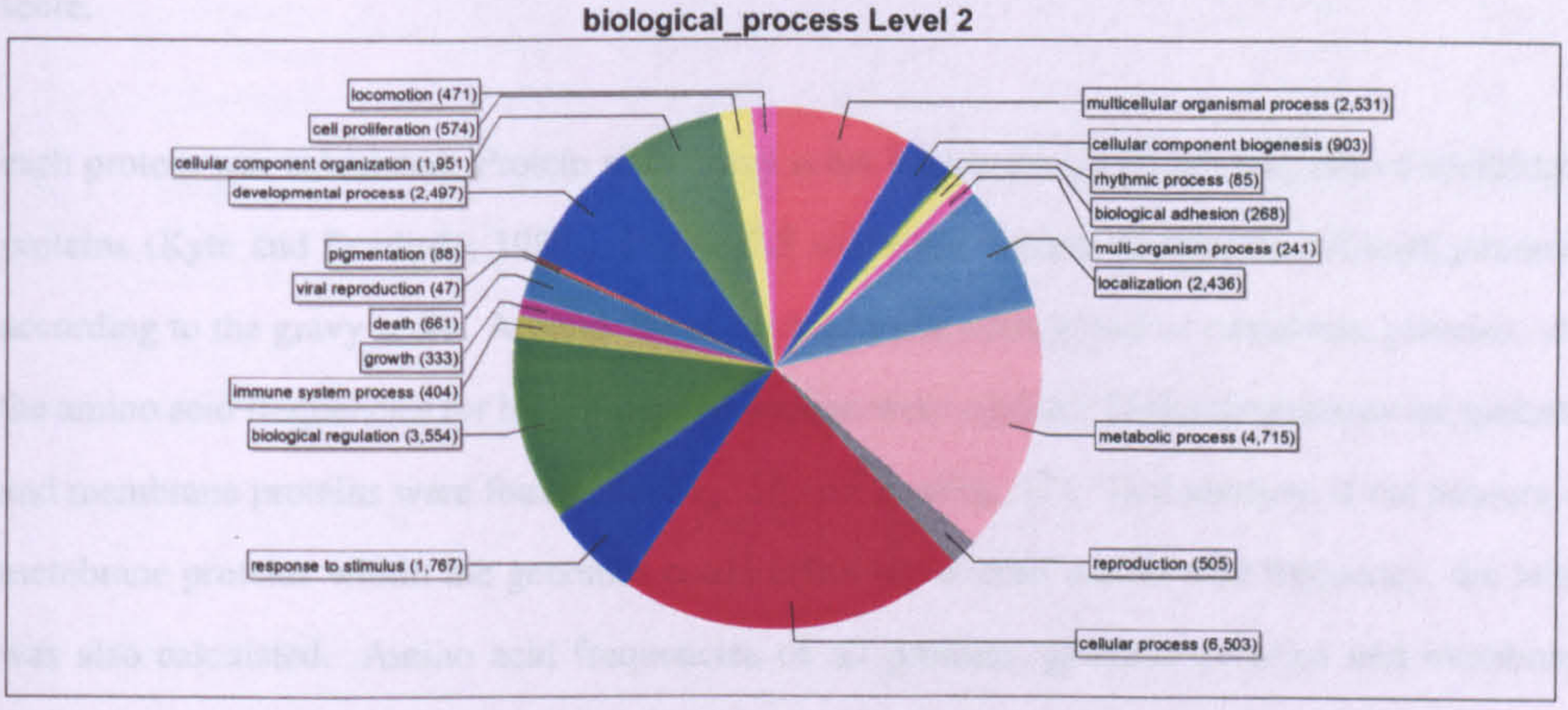


Figure 7.4: Biological process annotation by Blast2GO for *C. savignyi*

amino acids. This factor lead to variations in the global amino acid composition, at least in *E. coli* (Lobry, 1997; Lobry and Gautier, 1994). The Gravy score is an estimate of the overall hydrophobicity of proteins, higher gravy score indicate the hydrophobic character (Kyte and Doolittle, 1982). It is interesting to determine the amount of membrane proteins within the genome, in order to study possible alterations in amino acid frequencies and/or codon usage, or to be able to use all the protein sequences to perform analysis that consider as general the features of globular proteins (e.g. substitution matrices to divergence estimations).

To determine the total amount of membrane protein within the datasets, the gravy score of

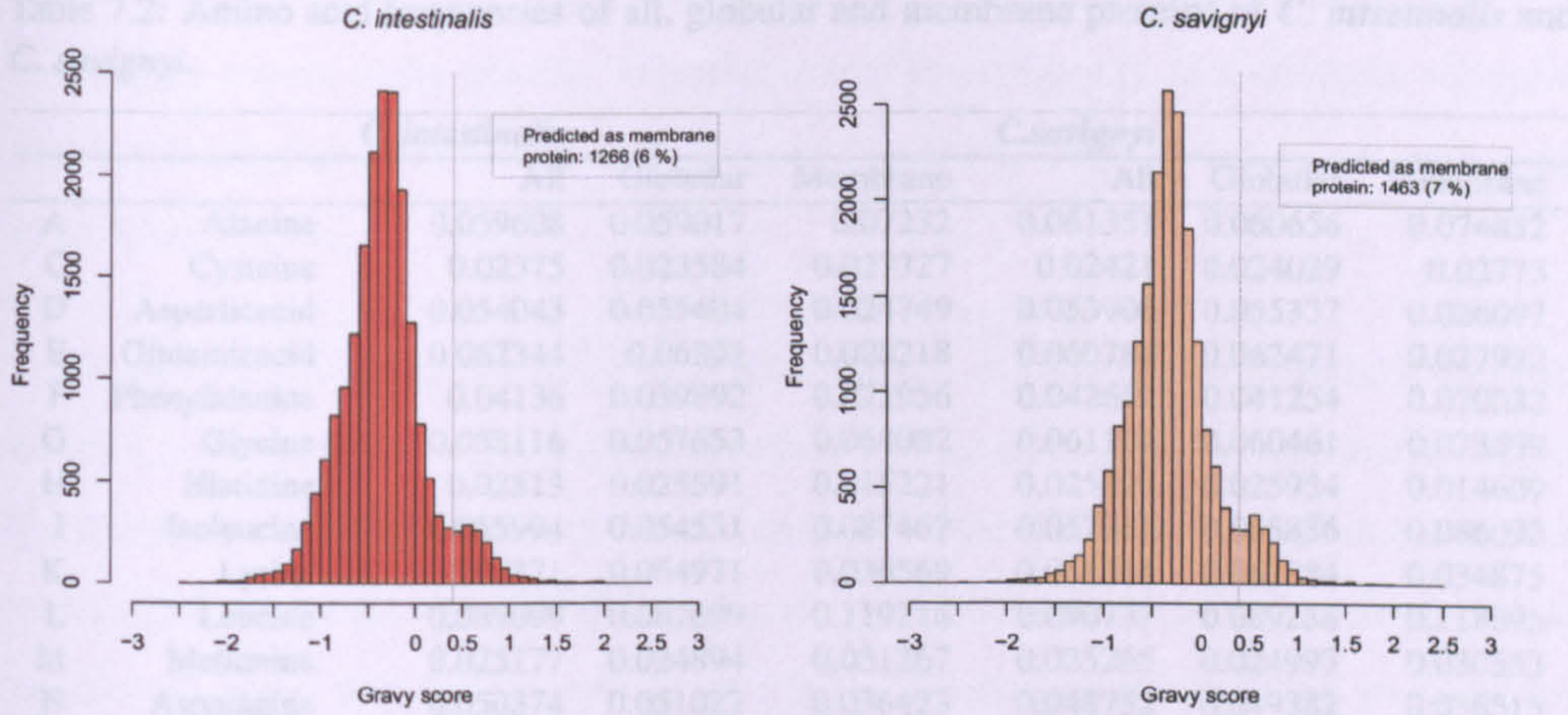


Figure 7.5: Identification of membrane proteins in of *C. intestinalis* and *C. savignyi* by gravy score.

each protein was calculated. Protein with gravy score higher than 0.45 are considered membrane proteins (Kyte and Doolittle, 1982). Figure 7.5 show the protein distribution of both genomes according to the gravy score. Around 7% of each genome correspond to membrane proteins, and the amino acid frequencies for both groups of protien were studied. Different patterns for globular and membrane proteins were found (SupFig. 16 and SupFig. 17). To determine if the amount of membrane proteins within the genomes could affect the overall amino acid frequency, the latter was also calculated. Amino acid frequencies of all proteins, globular proteins and membrane proteins are presented in Table 7.2.

Membrane proteins present higher frequencies for all hydrophobic amino acids (Alanine, isoleucine, leucine, proline, valine, and with the highest difference, phenylalanine), and lower frequencies for hydrophilic amino acid, with the exception of serine that presents similar frequencies. This pattern, as was expected, show different amino acid compositon for globular and membrane proteins. However, due to the low representation that membrane proteins have in both datasets, globular proteins present almost the same frequencies compared with those of the overall group (globular + membrane proteins). In other words, regarding to amino acid frequencies, globular proteins present very similar amino acid frequencies to those of all proteins. This observation allow to use the overall group of protein of both genomes analysed as representative of globular

Table 7.2: Amino acid frequencies of all, globular and membrane proteins of *C. intestinalis* and *C. savignyi*.

		C. intestinalis			*C.savignyi*		
		All	Globular	Membrane	All	Globular	Membrane
A	Alanine	0.059608	0.059017	0.07232	0.061351	0.060656	0.074852
C	Cysteine	0.02375	0.023584	0.027327	0.02421	0.024029	0.02773
D	Asparticacid	0.054043	0.055404	0.024749	0.053906	0.055337	0.026097
E	Glutamicacid	0.062344	0.06393	0.028218	0.060784	0.062471	0.027992
F	Phenylalanine	0.04136	0.039892	0.072956	0.042662	0.041254	0.070032
G	Glycine	0.058116	0.057653	0.068082	0.061104	0.060461	0.073599
H	Histidine	0.02513	0.025591	0.015221	0.025399	0.025954	0.014609
I	Isoleucine	0.055994	0.054531	0.087467	0.057317	0.055836	0.086092
K	Lysine	0.06371	0.064971	0.036569	0.061514	0.062884	0.034875
L	Leucine	0.089099	0.087699	0.119216	0.090737	0.089288	0.118895
M	Metionine	0.025177	0.024894	0.031267	0.025265	0.024993	0.030553
N	Asparagine	0.050374	0.051022	0.036423	0.048752	0.049382	0.036515
P	Proline	0.046998	0.047501	0.036154	0.047487	0.047895	0.039568
Q	Glutamine	0.041308	0.042054	0.025253	0.040474	0.041316	0.024094
R	Arginine	0.050881	0.051852	0.029981	0.050947	0.051992	0.030635
S	Serine	0.079386	0.079537	0.076127	0.077172	0.07721	0.076424
T	Threonine	0.061157	0.061184	0.060579	0.058986	0.058952	0.059643
V	Valine	0.06683	0.065814	0.088702	0.066283	0.065251	0.086351
W	Tryptophan	0.012087	0.011761	0.019104	0.012563	0.01219	0.019823
Y	Tyrosine	0.032291	0.031748	0.043969	0.032714	0.03227	0.041336

proteins in further analyses. Specifically, divergence estimation can be performed with the same amino acid substitution matrix (see Chapter 3).

7.3 Discussion

Gene loss and sequences annotation One of the interesting features of the Ciona genome is the absence of several genes which are present in other invertebrate as well as in vertebrate genomes. These findings allow to assume that Ciona's genome has lost these genes (Dehal et al., 2002). In effect, it has been estimated that Ciona lost 35% and 45% more ancestral gene families than pufferfish and humans respectively (Hughes and Friedman, 2005). Thus, it is interesting to assess which genes were lost in Ciona, and moreover, to identify the biological reasons related to it. Taking into consideration that they are sessile animals in the adult stage, one possible hypothesis is that the gene losses might be related to motility, and/or may reflect adaptations to the marine environment where they are bounded to live. Several specific Ciona's genes it has been claimed as lost (Holland and Gibson-Brown, 2003; Ferrier and Holland, 2002), although, in order to assess

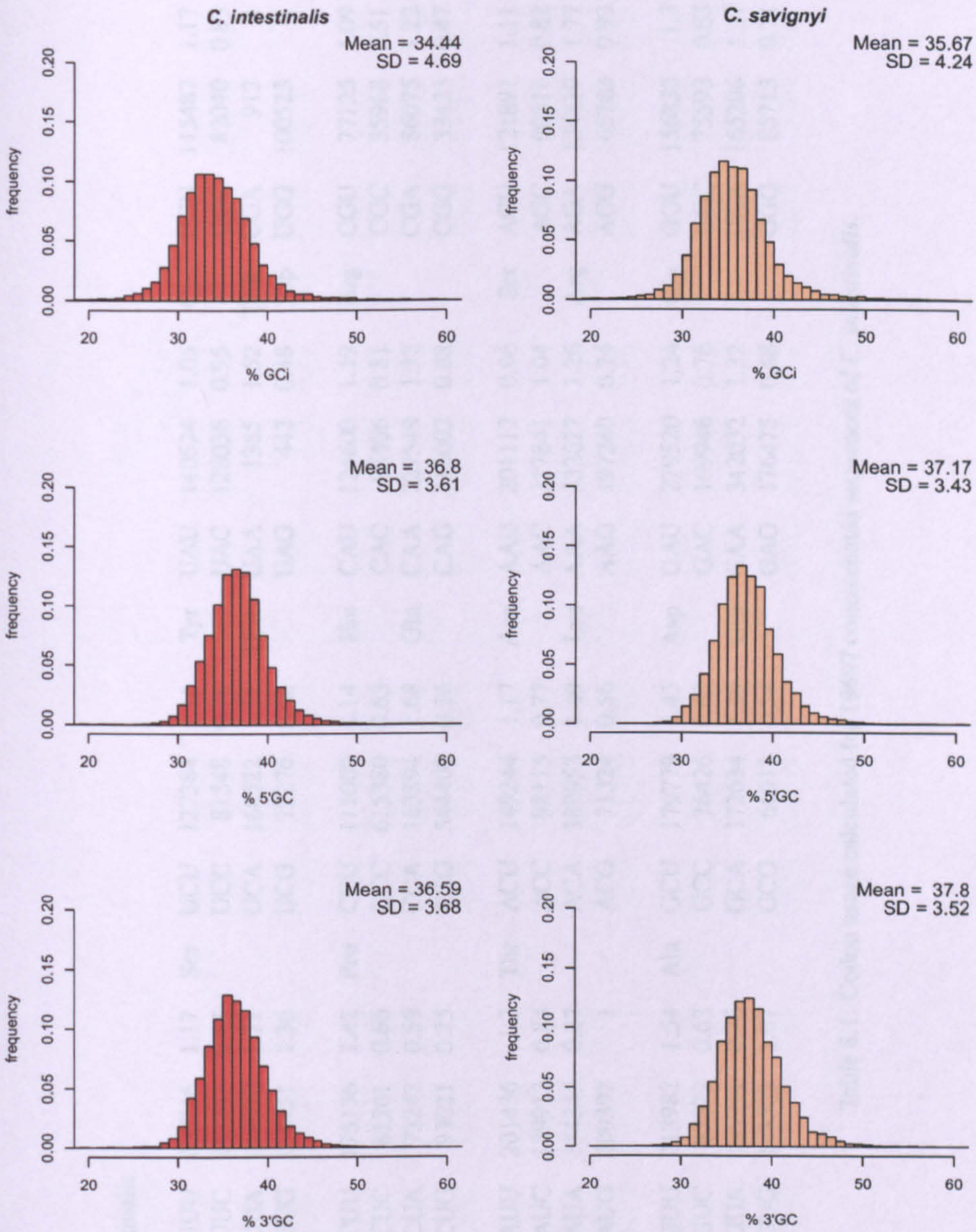
the real amount of gene losses, it is absolutely essential to have an accurate identification of the overall gene content of the genome. Moreover, it is also necessary to understand the biological function of the genes, because a missing gene (not founded by homology) could not correspond to a missing function.

The primary annotation by transfer of information and GO was exreamly useful to perform different analyses specially regarding to protein function, but was not sufficient to identify neither the overall genome of *C. intestinalis* nor *C. savignyi*. After this procedure still about 20% of both genomes remained with out identification at all. Furthermore, erroneous annotation could be inferred. Given that, first, Ciona's genome is not complete as already Dehal warns on the publication of the draft. Second, from the overall identified gene models, an important number, near to 20% of the genes, do not have a clear identification. In other words, 20% of the sequences do not have a clear homolog in a large amount of genomic data available, even having support of EST. And third, the high rate of evolution that this organism present (see 3); The idea of throw ligth upon these missing genes, specially tring to identified the biological funciotn behind it, remain unattainable.

7.4 Conclusions

This primary annotation not only permits to identify almost 80% of the sequences of both ascidian genomes, but also catalog the sequences according to the biological funcion, process and space where they "work" . This step could be crucial to understand the biological reasons of peculiarities that these genomes may have. In other words, to determine and understand the functions in which sequences are involved, is useful in genome comparative analysis. Unfortunately, not all the sequences could be identified, as was necessary to determine the gene losses that these genomes have been claimed to have. ENSEMBL databases were used to analyse Ciona's genomes, to be considered more complete and reliable. Almost 80% of the sequences of both genomes were identified, and almost 40% were annotataed with a Gene Ontology description. However, 20% still remain with out identification and no new knowledge about missing genes could be discovered.

8 | Supplementary figures



Sup.Fig 1: Different GC content distribution for *C. intestinalis* and *C. savignyi*. GCi - intronic, 5'GC - Upstream flanking region (2000pb), 3'GC - Downstream flanking region (200pb)

C.intestinalis

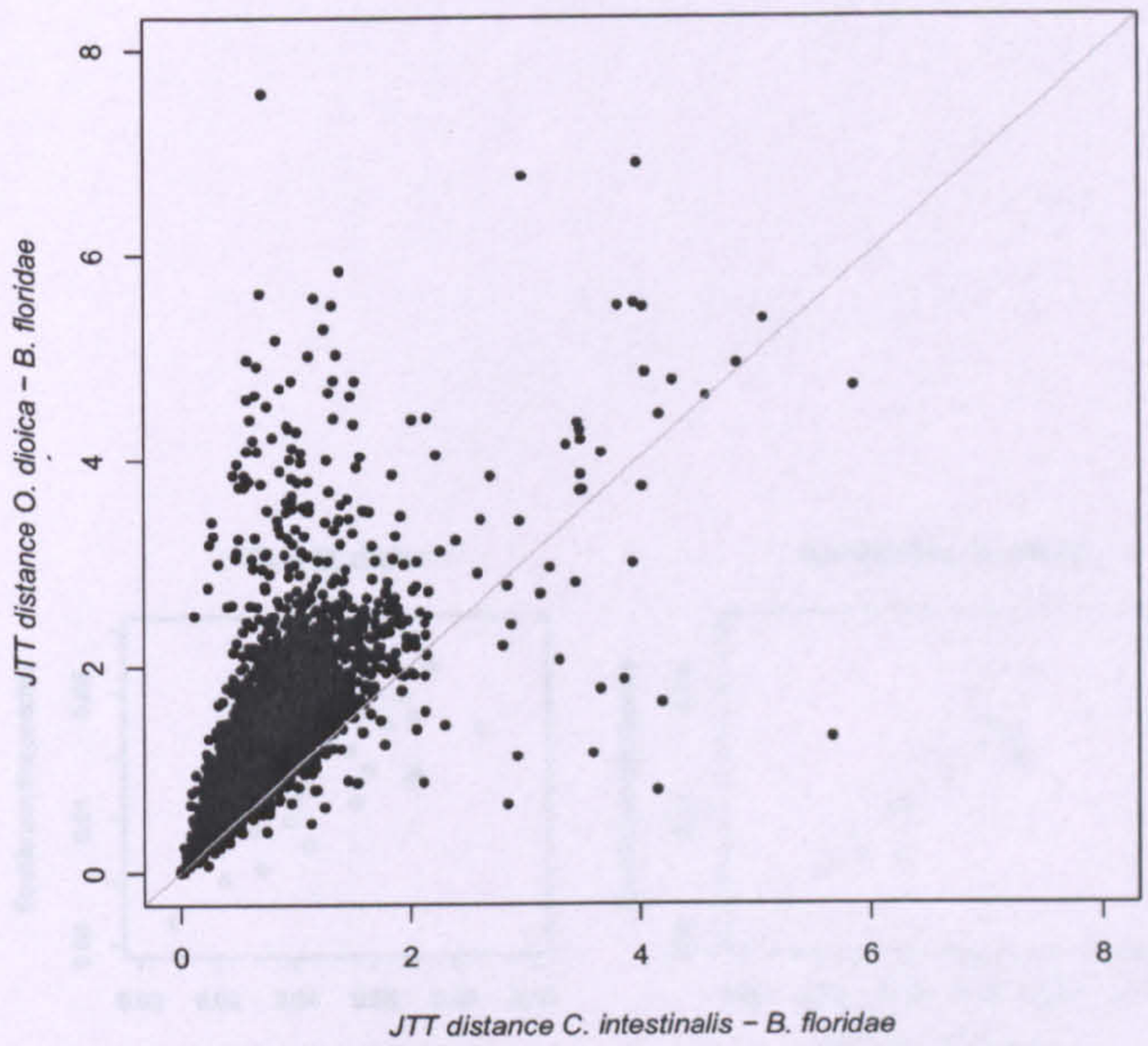
Phe	UUU	200746	1.17	Ser	UCU	127384	1.16	Tyr	UAU	140524	1.05	Cys	UGU	115487	1.17
	UUC	143243	0.83		UCC	81548	0.74		UAC	128036	0.95		UGC	82040	0.83
Leu	UUA	149970	1.21		UCA	166522	1.51	TER	UAA	1385	1.52	TER	UGA	912	1
	UUG	168451	1.36		UCG	72276	0.66		UAG	442	0.48	Trp	UGG	100523	1
	CUU	175136	1.42	Pro	CCU	111000	1.14	His	CAU	124600	1.19	Arg	CGU	77125	1.09
	CUC	81201	0.66		CCC	615380	0.63		CAC	84406	0.81		CGC	35968	0.51
	CUA	73242	0.59		CCA	163894	1.68	Gln	CAA	227549	1.32		CGA	86075	1.22
	CUG	93021	0.75		CCG	544400	0.56		CAG	116002	0.68		CGG	33423	0.47
Ile	AUU	201436	1.3	Thr	ACU	149244	1.17	Asn	AAU	201112	0.96	Ser	AGU	121891	1.11
	AUC	129912	0.84		ACC	98115	0.77		AAC	217841	1.04		AGC	90618	0.82
	AUA	134345	0.87		ACA	189953	1.49	Lys	AAA	332627	1.26	Arg	AGA	124819	1.77
Met	AUG	209392	1		ACG	71324	0.56		AAG	197240	0.74		AGG	65763	0.93
Val	GUU	213982	1.54	Ala	GCU	179779	1.45	Asp	GAU	279520	1.24	Gly	GGU	156832	1.3
	GUC	87372	0.63		GCC	76426	0.62		GAC	169946	0.76		GGC	75593	0.63
	GUA	105961	0.76		GCA	172634	1.39	Glu	GAA	342032	1.32		GGA	165206	1.37
	GUG	148503	1.07		GCG	66912	0.54		GAG	176475	0.68		GGG	85713	0.71

Table 8.1: Codon usage calculated for 19697 concatenated sequences of *C. intestinalis*.

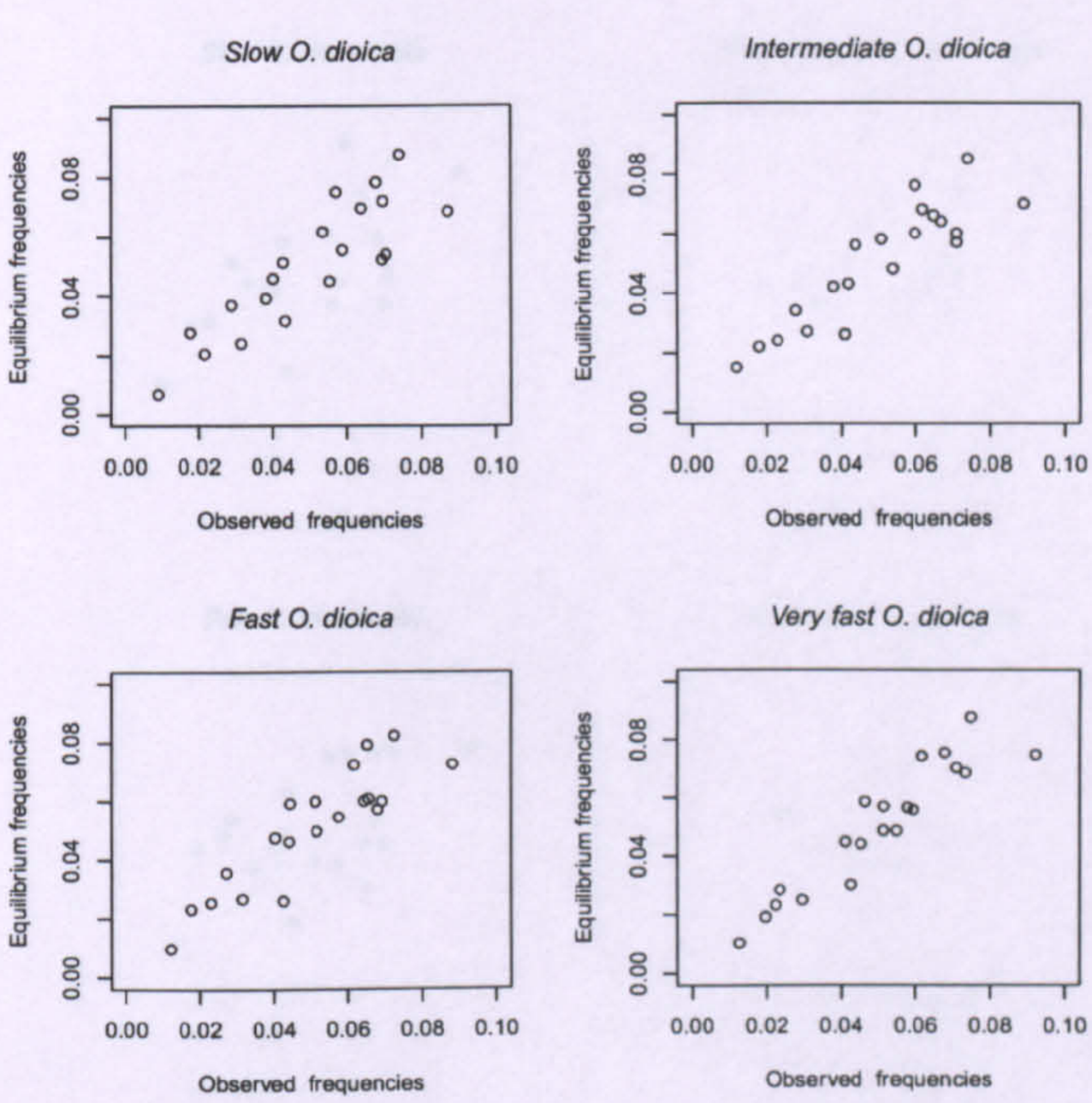
C.savignyi

Phe	UUU	200198	0.97	Ser	UCU	118268	0.95	Tyr	UAU	138218	0.87	Cys	UGU	116537	0.99
	UUC	213325	1.03		UCC	124027	0.99		UAC	178876	1.13		UGC	118129	1.01
Leu	UUA	123100	0.84		UCA	156274	1.25	TER	UAA	1653	1.41	TER	UGA	1331	1.13
	UUG	200342	1.37		UCG	107281	0.86		UAG	543	0.46	Trp	UGG	121778	1
	CUU	172772	1.18	Pro	CCU	108957	0.95	His	CAU	121179	0.98	Arg	CGU	70008	0.85
	CUC	142232	0.97		CCC	91833	0.8		CAC	125011	1.02		CGC	55421	0.67
	CUA	81450	0.56		CCA	171906	1.49	Gln	CAA	223490	1.14		CGA	108518	1.32
	CUG	159620	1.09		CCG	87600	0.76		CAG	168822	0.86		CGG	59567	0.72
Ile	AUU	209399	1.13	Thr	ACU	153517	1.07	Asn	AAU	208512	0.88	Ser	AGU	122733	0.98
	AUC	206060	1.11		ACC	153861	1.08		AAC	264046	1.12		AGC	119445	0.96
	AUA	140117	0.76		ACA	174085	1.22	Lys	AAA	338927	1.14	Arg	AGA	123468	1.5
Met	AUG	244894	1		ACG	90289	0.63		AAG	257325	0.86		AGG	76851	0.93
Val	GUU	191329	1.19	Ala	GCU	187211	1.26	Asp	GAU	291132	1.11	Gly	GGU	159272	1.08
	GUC	133251	0.83		GCC	119316	0.8		GAC	231383	0.89		GGC	111281	0.75
	GUA	99878	0.62		GCA	181649	1.22	Glu	GAA	347373	1.18		GGA	198138	1.34
	GUG	218026	1.36		GCG	106499	0.72		GAG	241806	0.82		GGG	123593	0.83

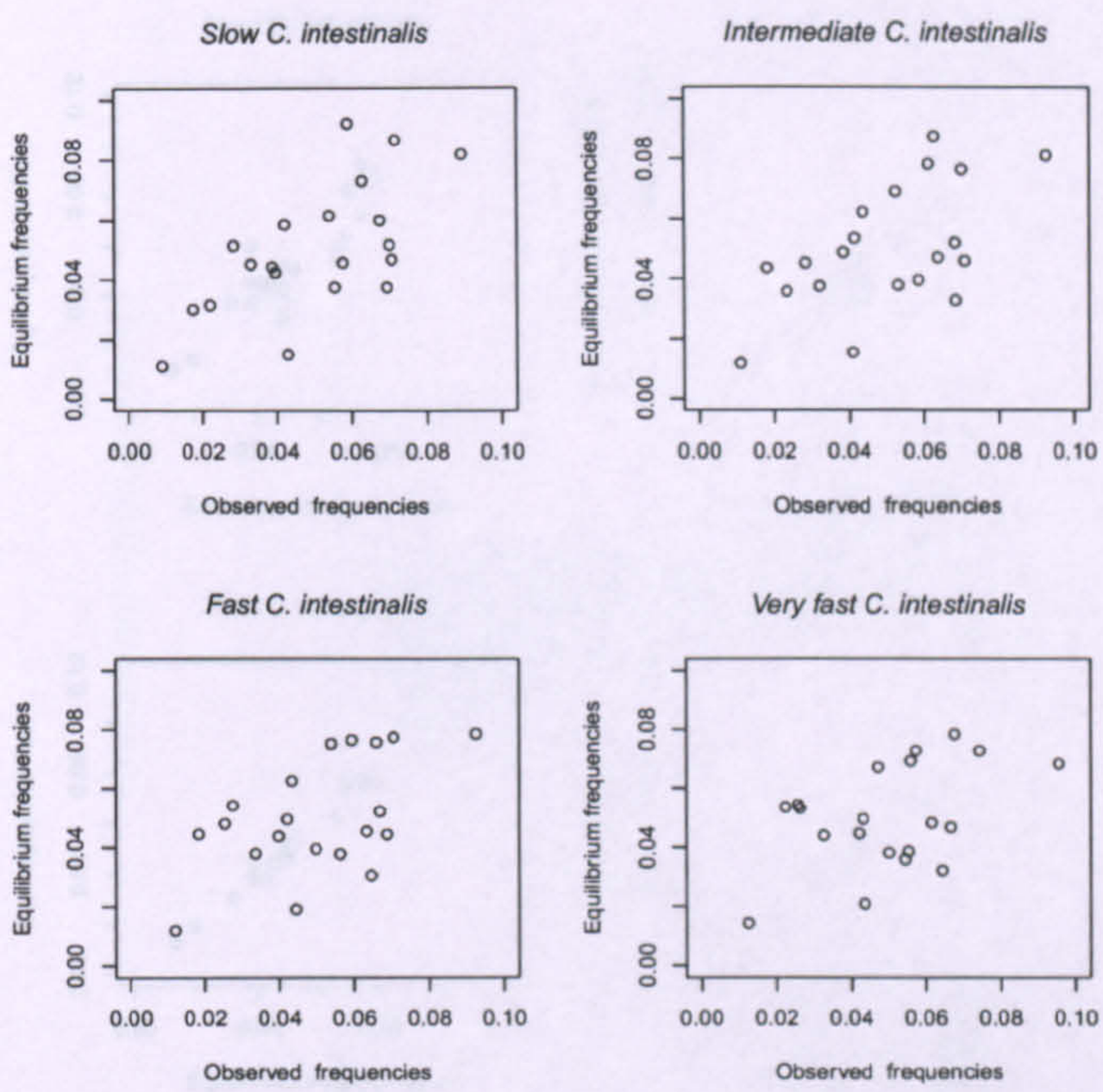
Table 8.2: Codon usage calculated for 20143 concatenated sequences of *C. savignyi*.



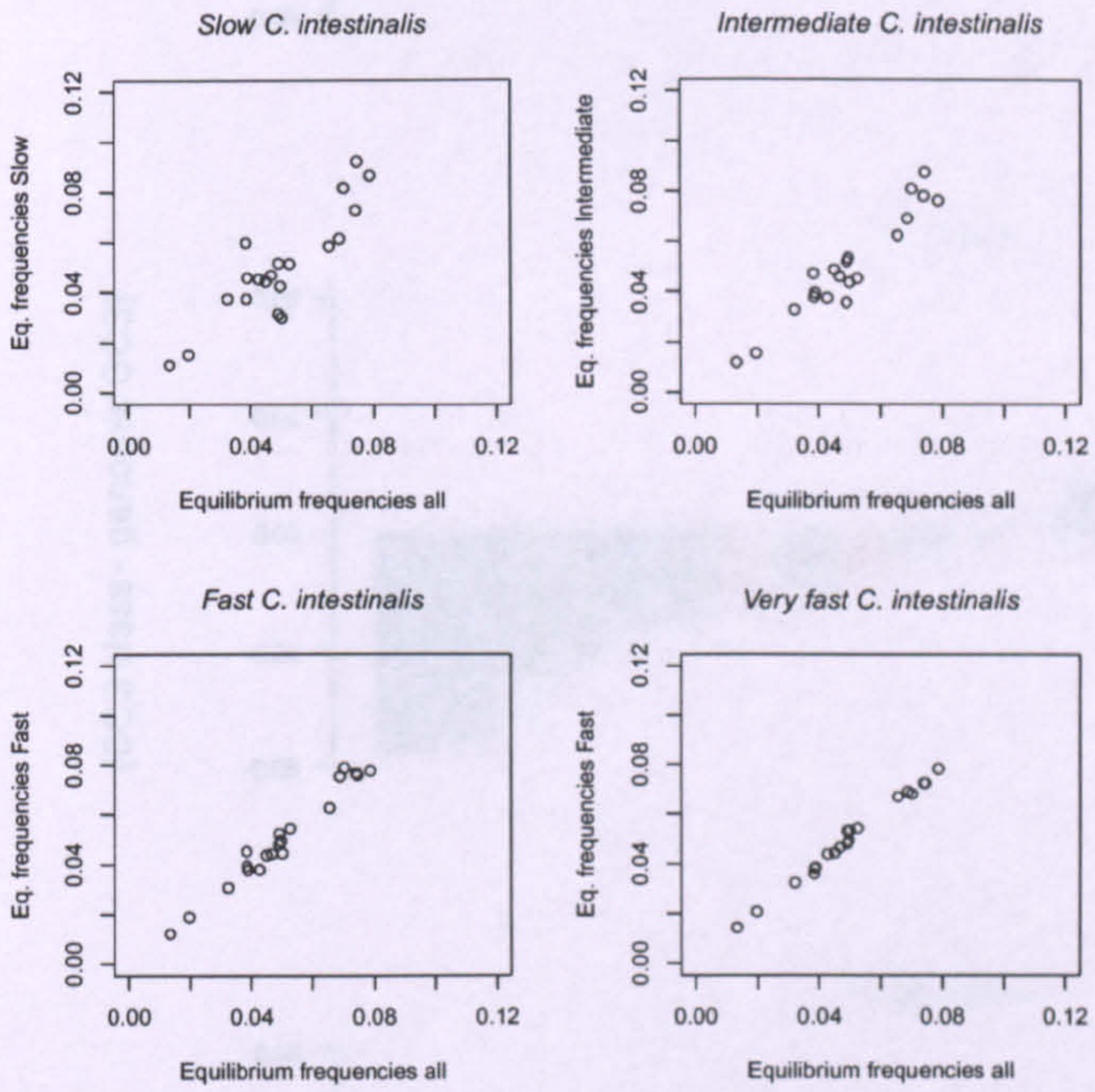
Sup.Fig 2: For each orthologs, the correspondent JTT distance of *C. intestinalis* versus *O. dioica* to the out-group *B. floridae*.



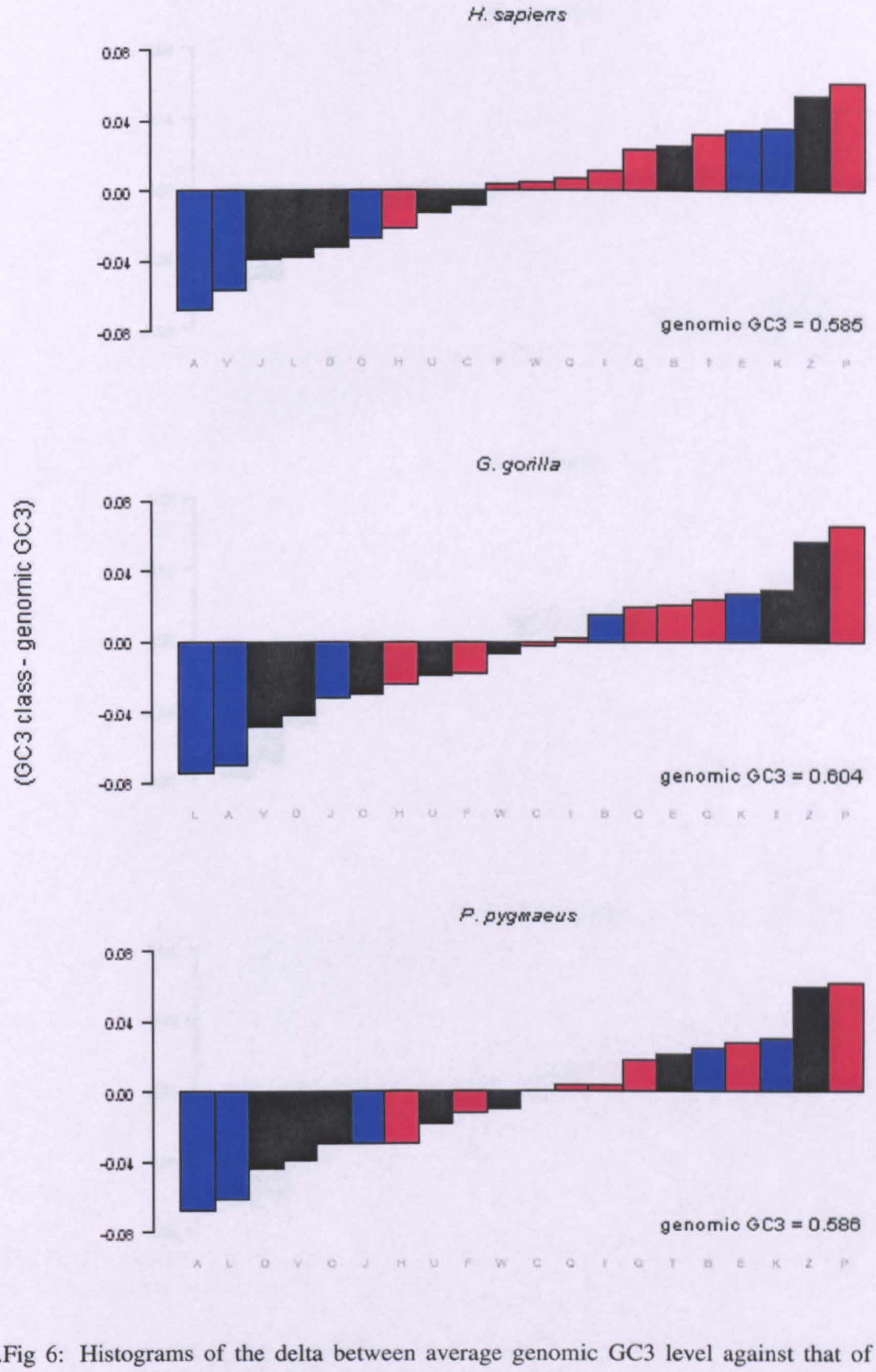
Sup.Fig 3: Equilibrium frequencies vs Observed frequencies for each groups of sequences analyzed of *O. dioica*



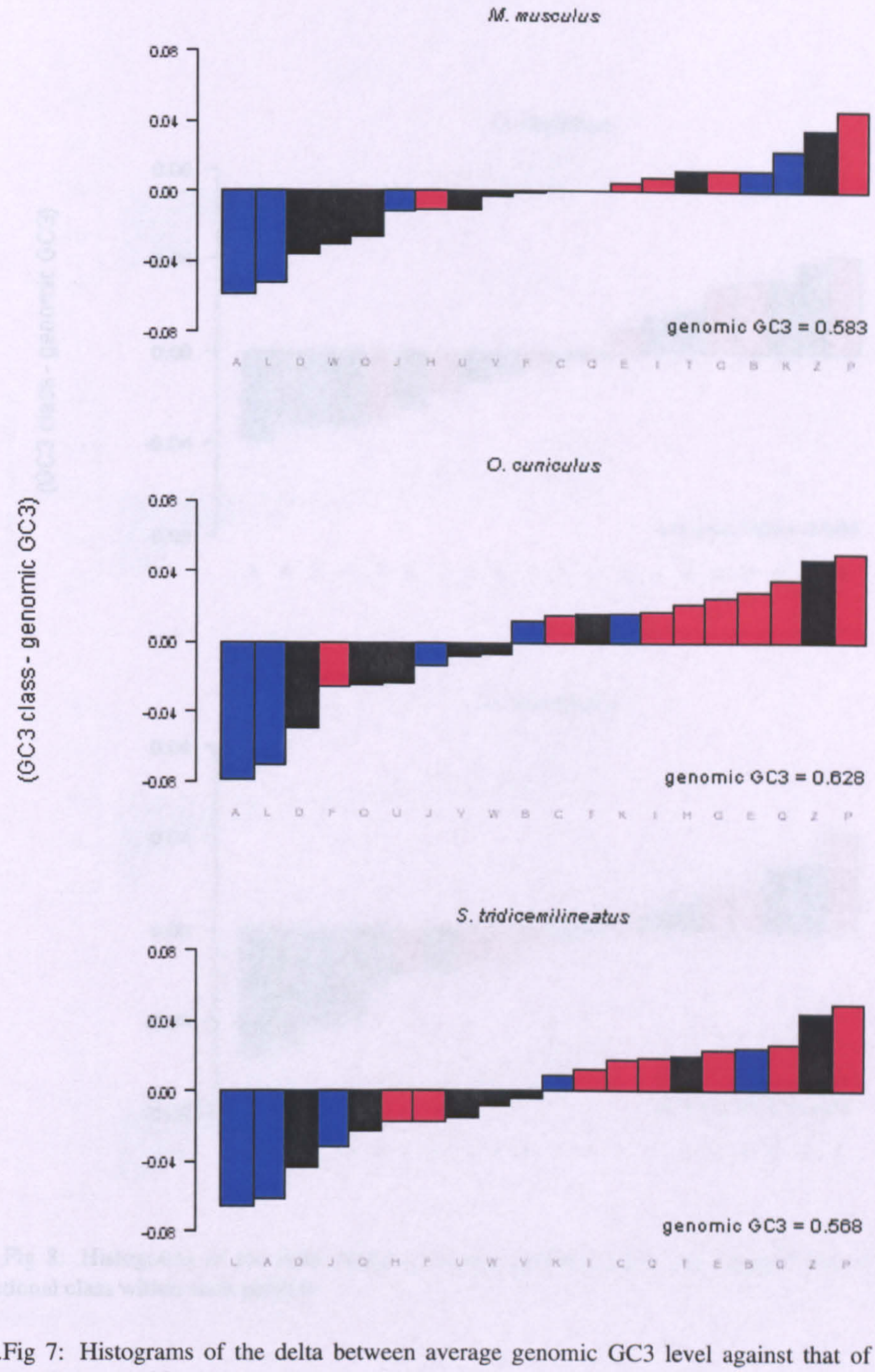
Sup.Fig 4: Equilibrium frequencies vs Observed frequencies for each groups of sequences analysed of *C. intestinalis*



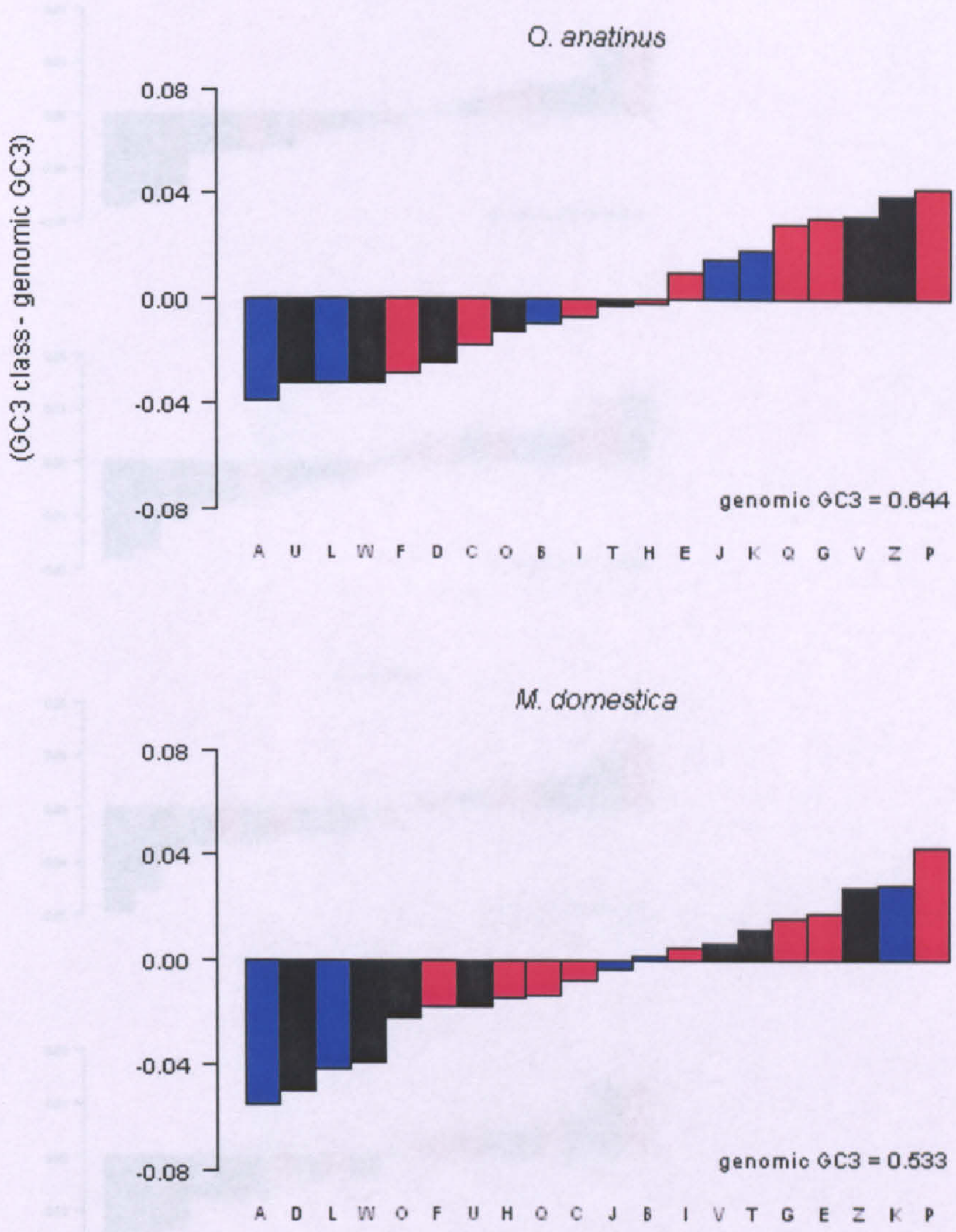
Sup.Fig 5: Equilibrium frequencies vs equilibrium frequencies for each groups of sequences analyzed of *C. intestinalis*



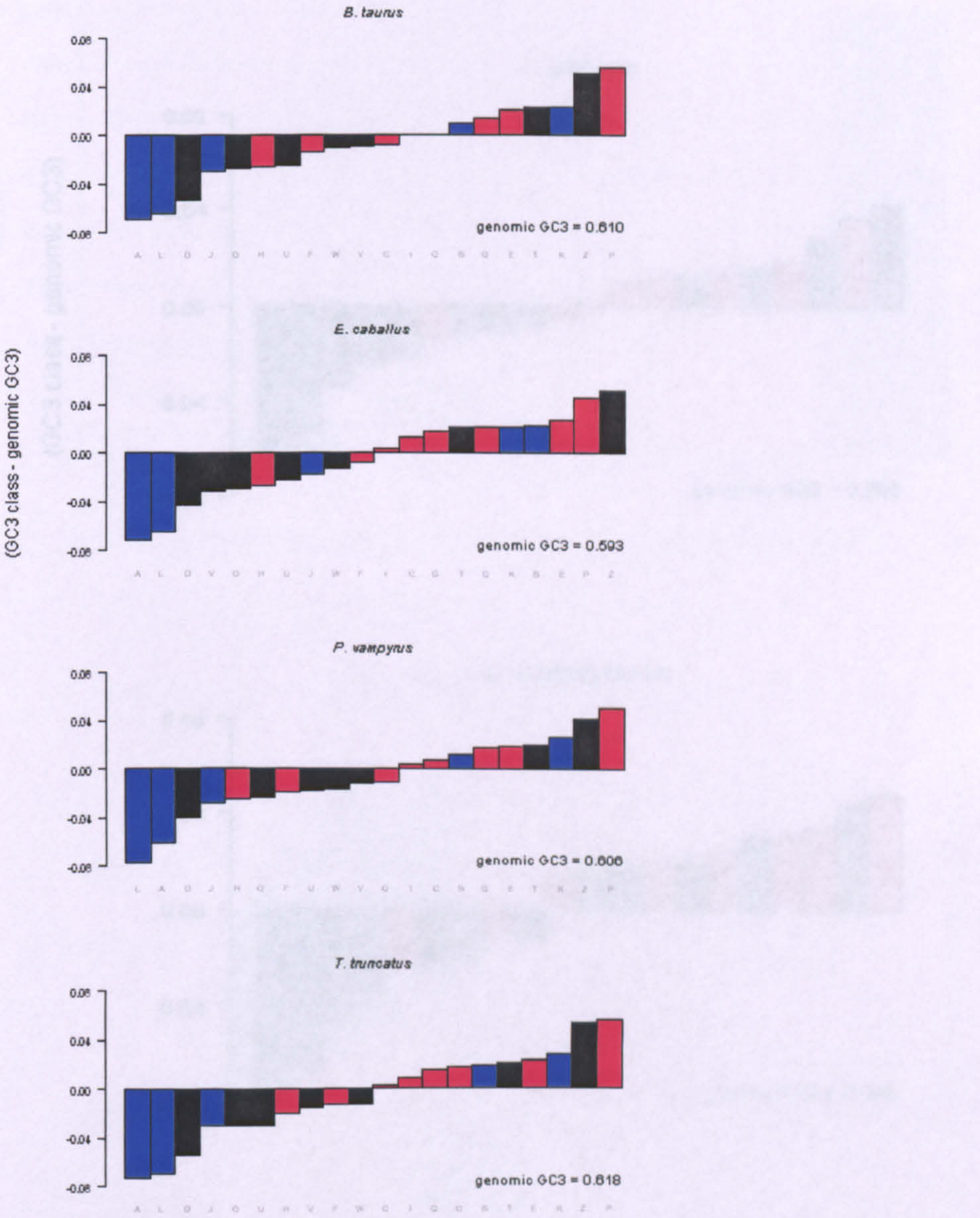
Sup.Fig 6: Histograms of the delta between average genomic GC3 level against that of each functional class within each genome.



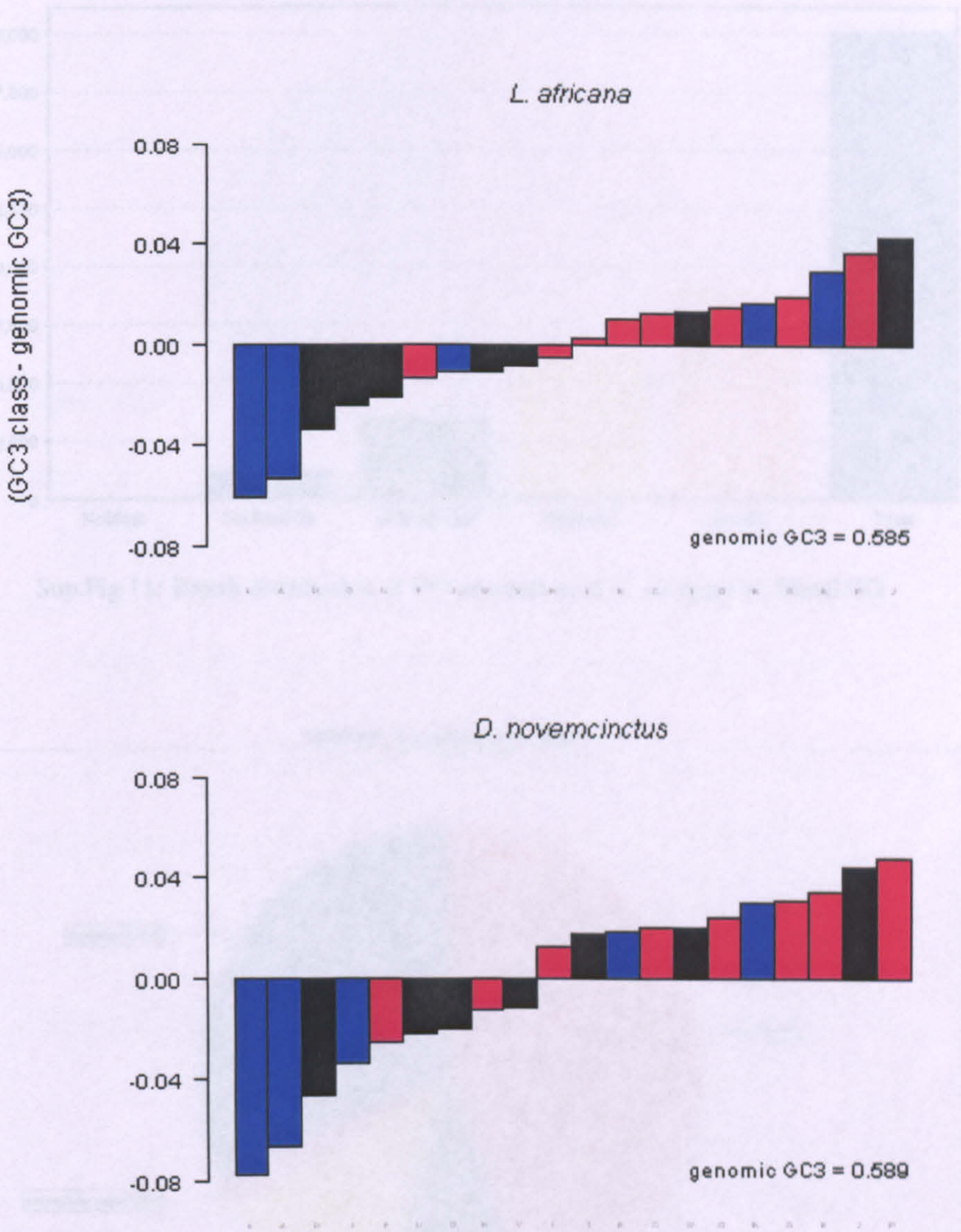
Sup.Fig 7: Histograms of the delta between average genomic GC3 level against that of each functional class within each genome.



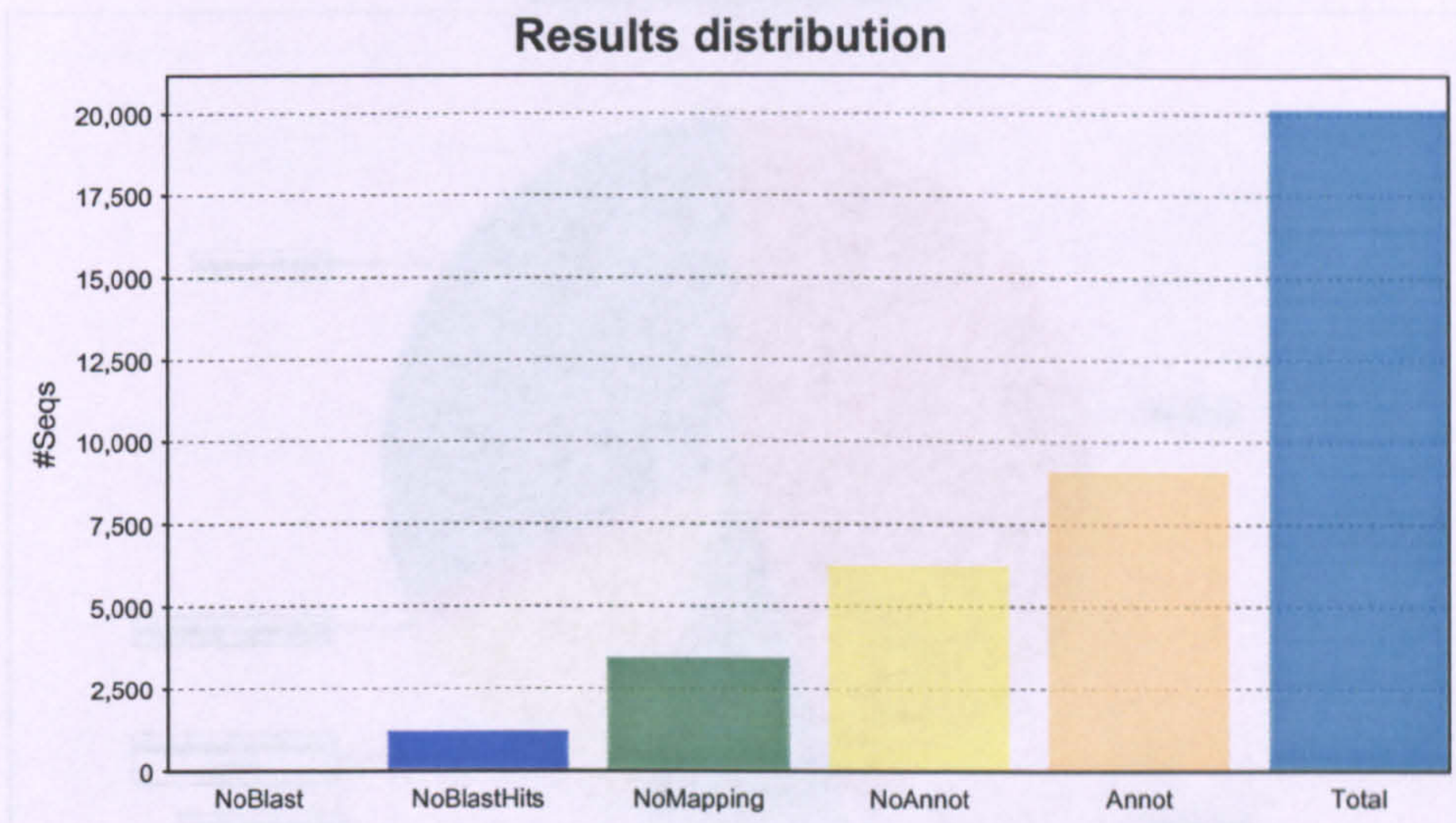
Sup.Fig 8: Histograms of the delta between average genomic GC3 level against that of each functional class within each genome.



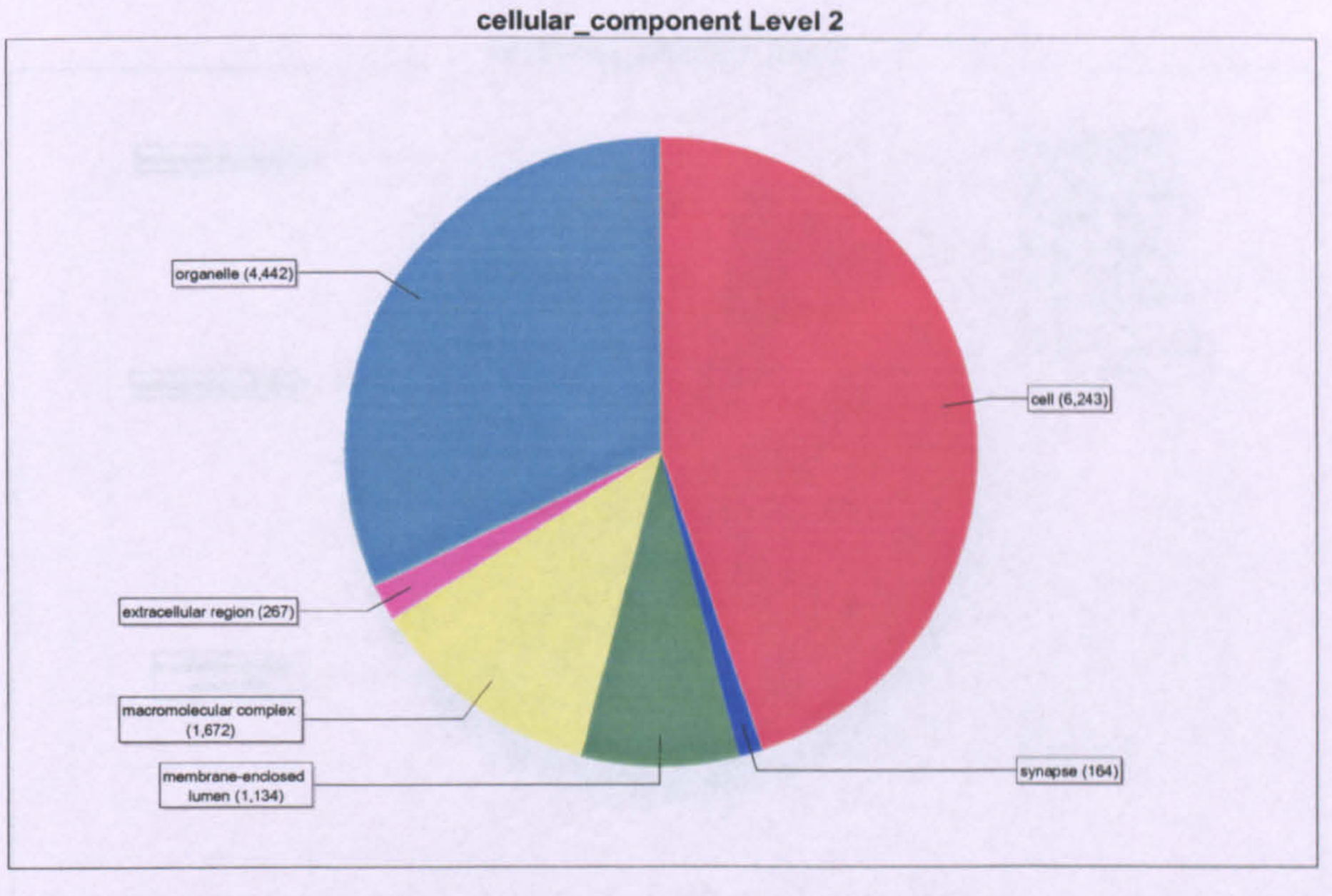
Sup.Fig 9: Histograms of the delta between average genomic GC3 level against that of each functional class within each genome.



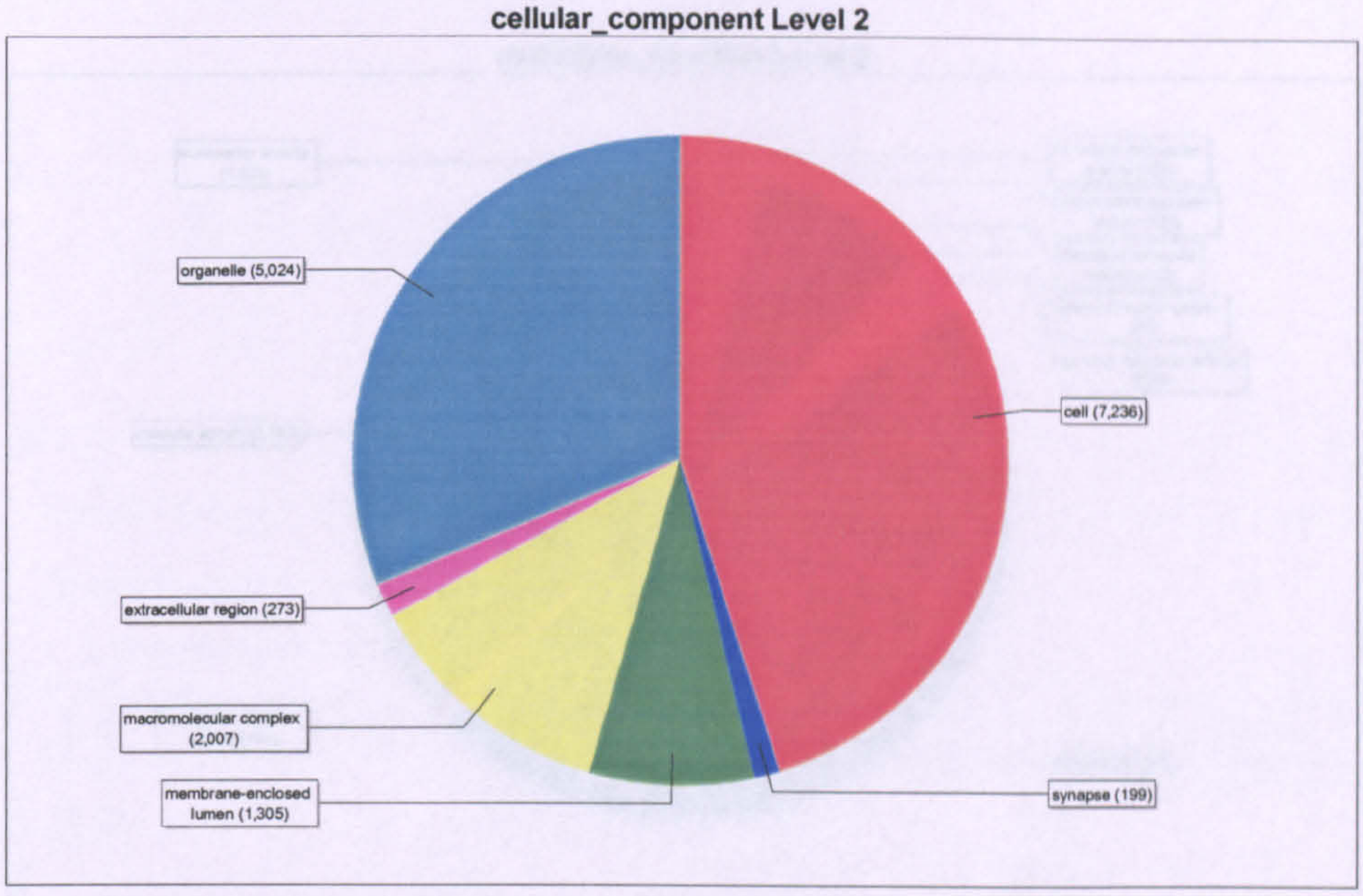
Sup.Fig 10: Histograms of the delta between average genomic GC3 level against that of each functional class within each genome.



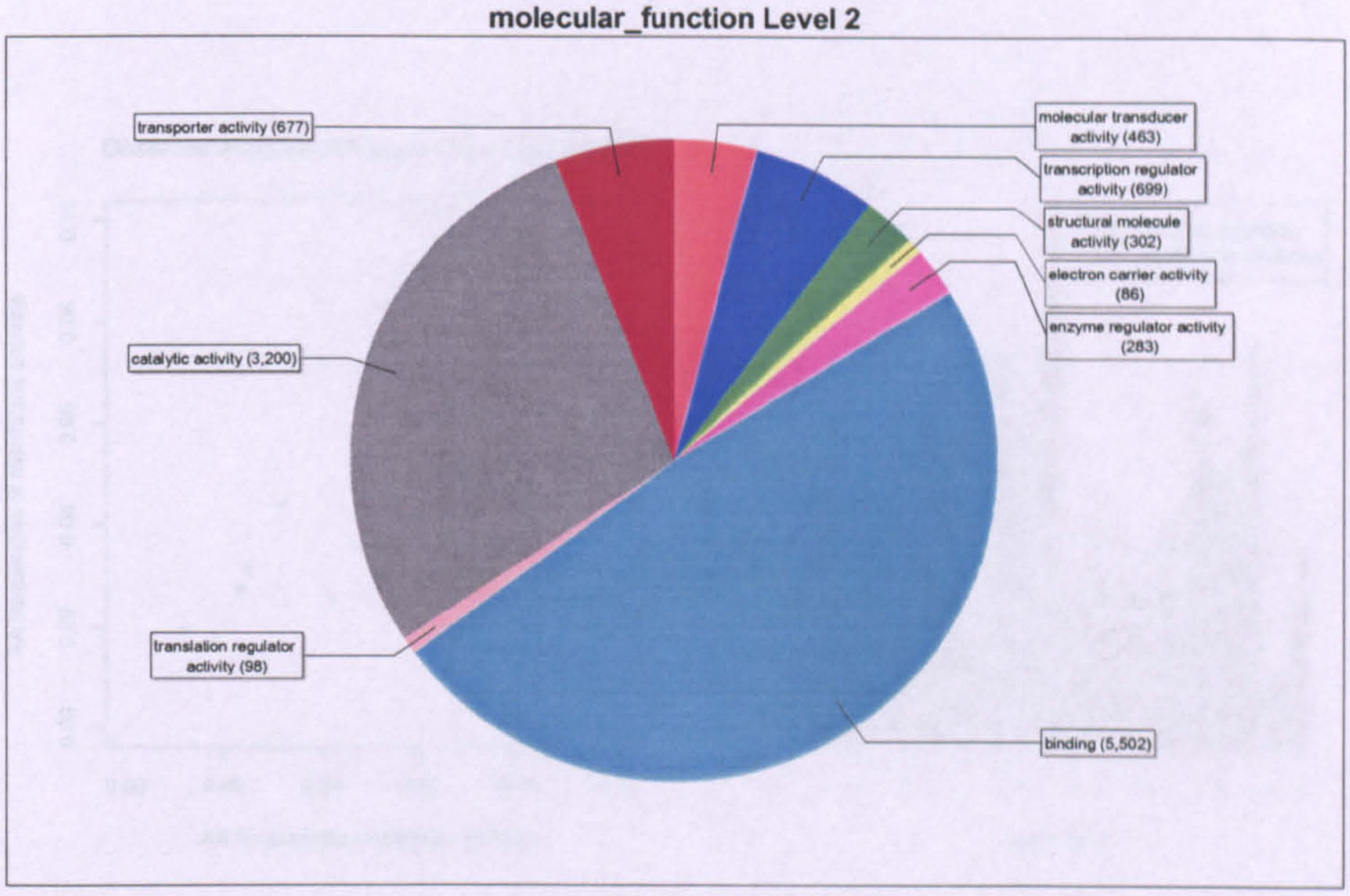
Sup.Fig 11: Result distribution of GO annotation of *C. savignyi* by Blast2GO



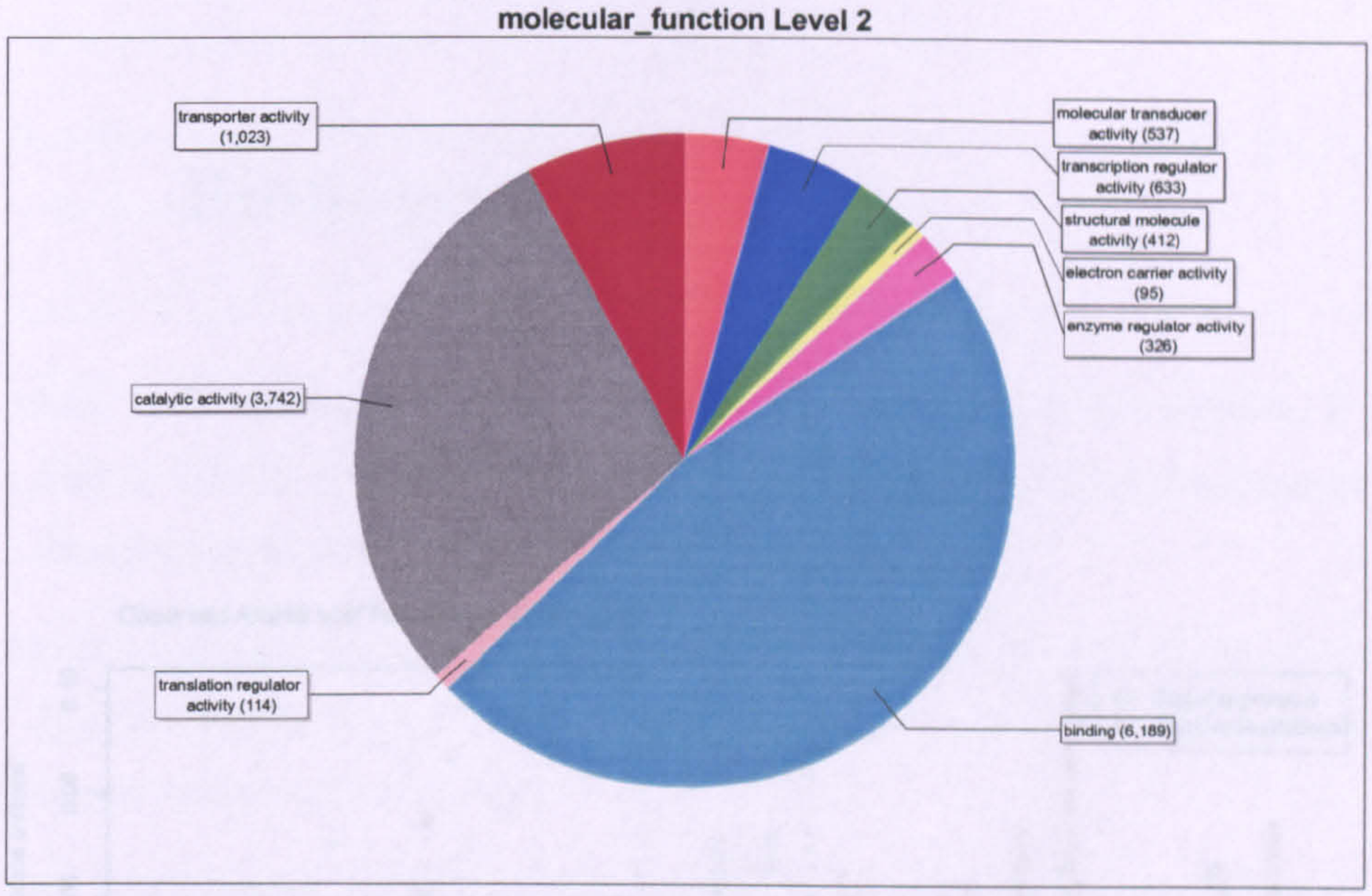
Sup.Fig 12: Biological process annotation by Blast2GO for *C. intestinalis*



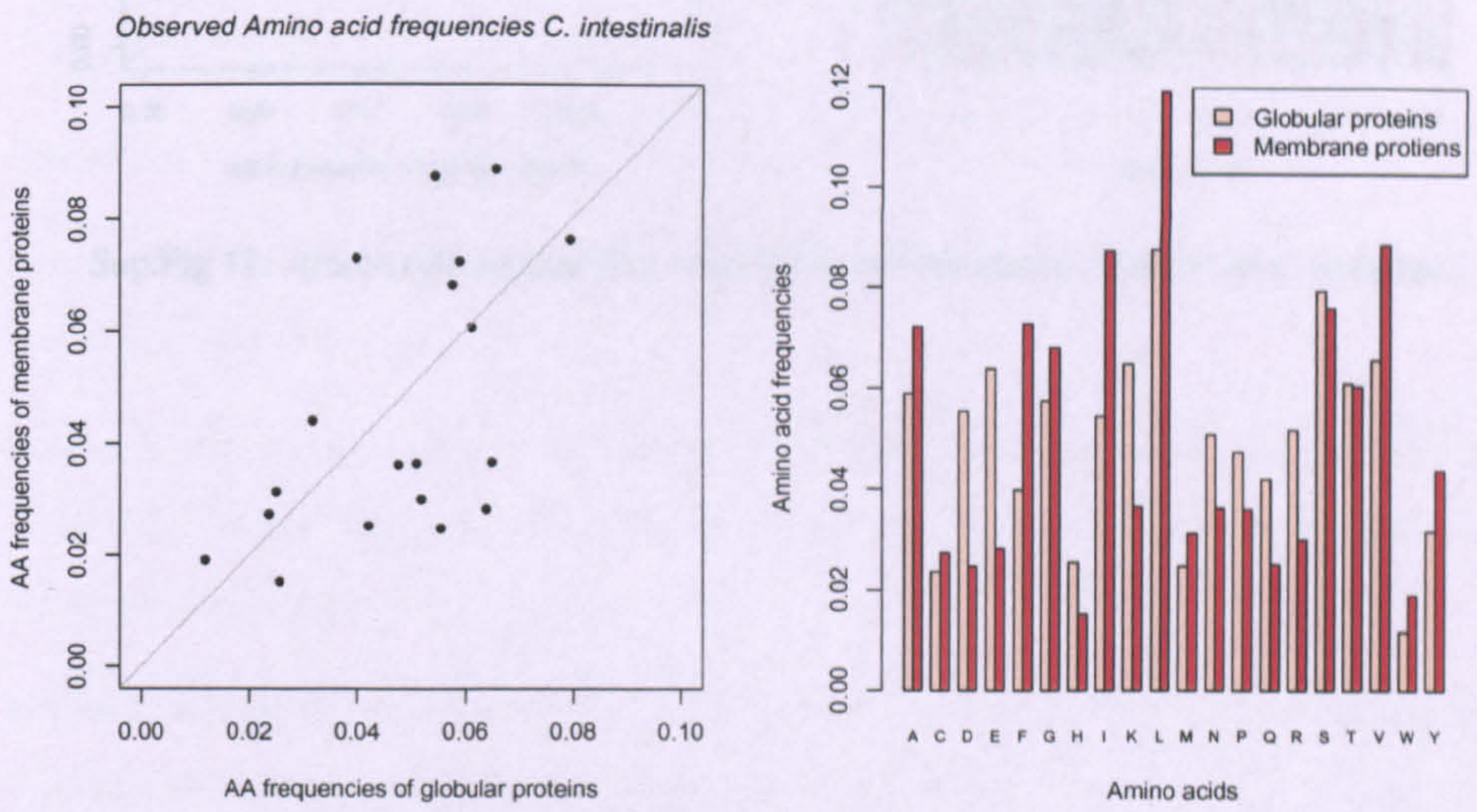
Sup.Fig 13: Biological process annotation by Blast2GO for *C. savignyi*



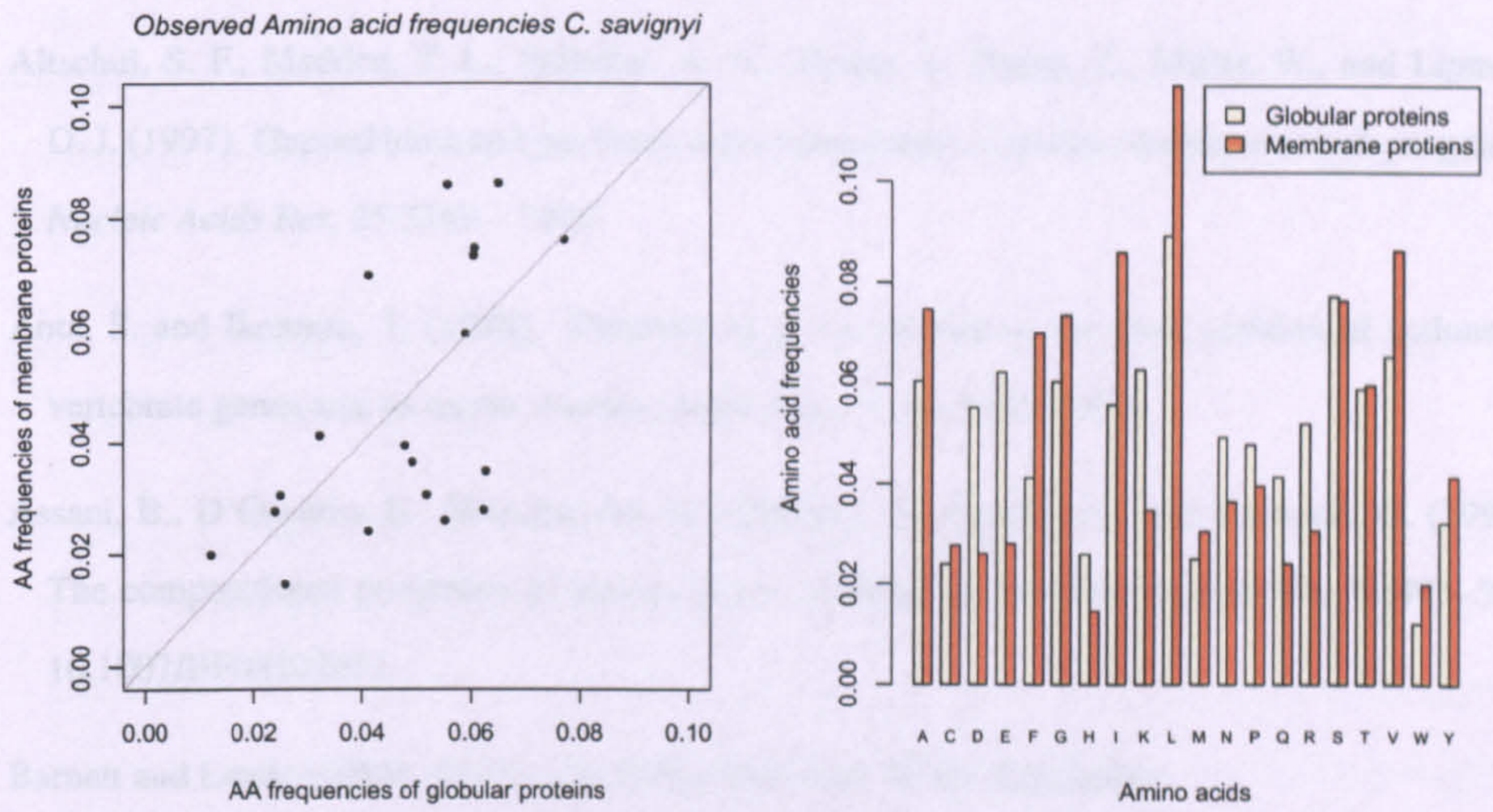
Sup.Fig 14: Biological process annotation by Blast2GO for *C. intestinalis*



Sup.Fig 15: Biological process annotation by Blast2GO for *C. savignyi*



Sup.Fig 16: Amino acid frequencies of globular and membrane proteins in *C. intestinalis*



Sup.Fig 17: Amino acid frequencies of globular and membrane proteins in *C. savignyi*

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., and Galle, R. F. (2000). The genome sequence of drosophila melanogaster. *Science*, 287:2185 – 2195.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389 – 3402.

Aota, S. and Ikemura, T. (1986). Diversity in g + c content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res*, 14(16):6345–6355.

Assani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., and Bernardi, G. (1991). The compositional properties of human genes. *Journal of Molecular Evolution*, 32:493–503. 10.1007/BF02102651.

Barnett and Lewis (1984). *Outliers in Statistical Data*. Wiley, Chichester.

Barton, N. H. (2010). Mutation and the evolution of recombination. *Philos Trans R Soc Lond B Biol Sci*, 365(1544):1281–1294.

Baudat, F. and Nicolas, A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome?iii. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):5213–5218.

Bellas, J., Beiras, R., and Vzquez, E. (2003). A standardisation of ciona intestinalis (chordata, ascidiacea) embryo-larval bioassay for ecotoxicological studies. *Water Res*, 37(19):4613–4622.

Benton, M. J. and Donoghue, P. C. J. (2007). Paleontological evidence to date the tree of life. *Mol Biol Evol*, 24(1):26–53.

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

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

Bernardi, G. and Bernardi, G. (1986). Compositional constraints and genome evolution. *J Mol Evol*, 24(1-2):1–11.

Bird, A. P. (1980). Dna methylation and the frequency of cpg in animal dna. *Nucleic Acids Research*, 8(7):1499–1504.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution*, 19(7):1181–1197.

Blair, J. E. and Hedges, S. B. (2005). Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol*, 22(11):2275–2284.

Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E. S., Thorndyke, M., Nakano, H., Kohn, A. B., Heyland, A., Moroz, L. L., Copley, R. R., and Telford, M. J. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum xenoturbellida. *Nature*, 444(7115):85–8. 1476-4687 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

Boussau, B. and Daubin, V. (2010). Genomes as documents of evolutionary history. *Trends Ecol Evol*, 25(4):224–232.

Bromham, L. and Penny, D. (2003). The modern molecular clock. *Nat Rev Genet*, 4(3):216–224. 1471-0056 10.1038/nrg1020 10.1038/nrg1020.

Bucciarelli, G., Bernardi, G., and Bernardi, G. (2002). An ultracentrifugation analysis of two hundred fish genomes. *Gene*, 295(2):153–162.

Bulmer, M. (1987). A statistical analysis of nucleotide sequences of introns and exons in human genes. *Molecular Biology and Evolution*, 4(4):395–405.

Byrd, J. and Lambert, C. C. (2000). Mechanism of the block to hybridization and selfing between the sympatric ascidians ciona intestinalis and ciona savignyi. *Mol Reprod Dev*, 56(4):541. 1040-452X (Print) Comment Letter.

Caestro, C., Bassham, S., and Postlethwait, J. H. (2003). Seeing chordate evolution through the ciona genome sequence. *Genome Biol*, 4(3):208.

Caputi, L., Andreakis, N., Mastrototaro, F., Cirino, P., Vassillo, M., and Sordino, P. (2007). Cryptic speciation in a model invertebrate chordate. *Proceedings of the National Academy of Sciences*, 104(22):9364–9369.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, 63:213–227.

Chen, J.-Y., Huang, D.-Y., Peng, Q.-Q., Chi, H.-M., Wang, X.-Q., and Feng, M. (2003). The first tunicate from the early cambrian of south china. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8314–8318.

Clarke, A. and Johnston, N. (September 1999). Scaling of metabolic rate with body mass and temperature in teleost fish. *Journal of Animal Ecology*, 68:893–905(13).

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–6. ¡p¿Conesa, Ana Gotz, Stefan Garcia-Gomez, Juan Miguel Terol, Javier Talon, Manuel Robles, Montserrat Research Support, Non-U.S. Gov't England Bioinformatics (Oxford, England) Bioinformatics. 2005 Sep 15;21(18):3674-6. Epub 2005 Aug 4.¡/p¿.

Consortium, T. C. E. S. (1998). Genome sequence of the nematode c. elegans: A platform for investigating biology. *Science*, 282(5396):2012–2018.

Corbo, J. C., Gregorio, A. D., and Levine, M. (2001). The ascidian as a model organism in developmental and evolutionary biology. *Cell*, 106:535–538.

Costantini, M., Auletta, F., and Bernardi, G. (2007). Isochore patterns and gene distributions in fish genomes. *Genomics*, 90(3):364 – 371.

Davidson, B. and Levine, M. (2003). Evolutionary origins of the vertebrate heart: Specification of the cardiac lineage in ciona intestinalis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11469–11473.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., Harafuji, N., Hastings, K. E., Ho, I., Hotta, K., Huang, W., Kawashima, T., Lemaire, P., Martinez, D., Meinertzhagen, I. A., Necula, S., Nonaka, M., Putnam, N., Rash, S., Saiga, H., Satake, M., Terry, A., Yamada, L., Wang, H. G., Awazu, S., Azumi, K., Boore, J., Branno, M., Chin-Bow, S., DeSantis, R., Doyle, S., Francino, P., Keys, D. N., Haga, S., Hayashi, H., Hino, K., Imai, K. S., Inaba, K., Kano, S., Kobayashi, K., Kobayashi, M., Lee, B. I., Makabe, K. W., Manohar, C., Matassi, G., Medina, M., Mochizuki, Y., Mount, S., Morishita, T., Miura, S., Nakayama, A., Nishizaka, S., Nomoto, H., Ohta, F., Oishi, K., Rigoutsos, I., Sano, M., Sasaki, A., Sasakura, Y., Shoguchi, E., Shin-i, T., Spagnuolo, A., Stainier, D., Suzuki, M. M., Tassy, O., Takatori, N., Tokuoka, M., Yagi, K., Yoshizaki, F., Wada, S., Zhang, C., Hyatt, P. D., Larimer, F., Detter, C., Doggett, N., Glavina, T., Hawkins, T., Richardson, P., Lucas, S., Kohara, Y., Levine, M., Satoh, N., and Rokhsar, D. S. (2002). The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, 298(5601):2157–67. 1095-9203 (Electronic) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439(7079):965–8. 1476-4687 (Electronic) Journal Article Research Support, Non-U.S. Gov't.

Delsuc, F., Tsagkogeorga, G., Lartillot, N., and Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis*, 46(11):592–604. 1526-968X (Electronic) Journal Article Research Support, Non-U.S. Gov't.

Donmez, N., Bazykin, G. A., Brudno, M., and Kondrashov, A. S. (2009). Polymorphism Due

to Multiple Amino Acid Substitutions at a Codon Site Within Ciona savignyi. *Genetics*, 181(2):685–690.

D'Onofrio, G., Ghosh, T. C., and Saccone, S. (2007). Different functional classes of genes are characterized by different compositional properties. *FEBS Lett*, 581(30):5819–5824.

D'onofrio, G., Habbari, K., Musto, H., Alvarez-Valin, F., Cruveiller, S., and Bernardi, G. (1999). Evolutionary genomics of vertebrates and its implications. *Annals of the New York Academy of Sciences*, 870:8194.

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

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–9. 1476-4687 (Electronic) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

Duret, L. and Galtier, N. (2000). The covariation between tpa deficiency, cpg deficiency, and g+c content of human isochores is due to a mathematical artifact. *Molecular Biology and Evolution*, 17(11):1620–1625.

Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10(1):285–311. PMID: 19630562.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. (2002). Vanishing gc-rich isochores in mammalian genomes. *Genetics*, 162(4):1837–1847.

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proceedings: Biological Sciences*, 252(1335):pp. 237–243.

Felsenstein, J. (2004). Phylip (phylogeny inference package) vesrion 3.67. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*.

Ferrier, D. E. K. and Holland, P. W. H. (2002). Ciona intestinalis parahox genes: evolution of hox/parahox cluster integrity, developmental mode, and temporal colinearity. *Molecular Phylogenetics and Evolution*, 24(3):412 – 417.

Forsdyke, D. R. (1996). Different biological species their dnas at different (g+c) *Journal of Theoretical Biology*, 178(4):405 – 417.

Forsdyke, D. R. (2007). Molecular sex: The importance of base composition rather than homology when nucleic acids hybridize. *Journal of Theoretical Biology*, 249(2):325 – 330.

Fraga, C. G., Shigenaga, M. K., Park, J. W., Degan, P., and Ames, B. N. (1990). Oxidative damage to dna during aging: 8-hydroxy-2'-deoxyguanosine in rat organ dna and urine. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4533–4537.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–752.

Fryxell, K. J. and Moon, W.-J. (2005). Cpg mutation rates in the human genome are highly dependent on local gc content. *Mol Biol Evol*, 22(3):650–658.

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

Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O., and Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11383–11390.

Gillooly, J. F., Brown, J. H., West, G. B., Savage, V. M., and Charnov, E. L. (2001). Effects of Size and Temperature on Metabolic Rate. *Science*, 293(5538):2248–2251.

Gissi, C., Pesole, G., Cattaneo, E., and Tartari, M. (2006). Huntingtin gene evolution in chordata and its peculiar features in the ascidian ciona genus. *BMC Genomics*, 7:288–304.

Goldstein, R. A. and Pollock, D. D. (2006). Observations of amino acid gain and loss during protein evolution are explained by statistical bias. *Mol Biol Evol*, 23(7):1444–1449.

Graur, D. and Li, W.-H. (2000). *Fundamentals of molecular evolution.* Sinauer Associates, Inc.

Gregorio, A. D. and Levine, M. (2002). Analyzing gene regulation in ascidian embryos: new tools for new perspectives. *Differentiation*, 70(4-5):132–139.

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

Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol*, 52(5):696–704.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M., and Lake, J. A. (1995). Evidence from 18s ribosomal dna that the lophophorates are protostome animals. *Science*, 267(5204):1641–1643.

Harrison, F. W. and Ruppert, E. E., editors (1997). *Microscopic Anatomy of Invertebrates*, volume 15, Hemichordata, Chaetognatha, and the Invertebrate Chordates. Wiley-Liss, Inc.

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nat Rev Genet*, 3:838 – 849.

Hill, M. M., Broman, K. W., Stupka, E., Smith, W. C., Jiang, D., and Sidow, A. (2008). The c. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Research*, 18:1369–1379.

Holland, L. Z. and Gibson-Brown, J. J. (2003). The ciona intestinalis genome: when the constraints are off. *Bioessays*, 25(6):529–32. 0265-9247 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review.

Hoshino, Z. and Nishikawa, T. (1985). Taxonomic studies of ciona intestinalis (l.) and its allies. *Publ Seto Mar Biol Lab*, 30(30):61–79.

Huang, S.-W., Friedman, R., Yu, N., Yu, A., and Li, W.-H. (2005). How strong is the mutagenicity of recombination in mammals? *Molecular Biology and Evolution*, 22(3):426–431.

Hughes, A. L. and Friedman, R. (2005). Loss of ancestral genes in the genomic evolution of ciona intestinalis. *Evol Dev*, 7(3):196–200. 1520-541X (Print) Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, P.H.S.

Hurst, L. D., Feil, E. J., and Rocha, E. P. C. (2006). Protein evolution: causes of trends in amino-acid gain and loss. *Nature*, 442(7105):E11–2; discussion E12.

Huttley, G. A., Wakefield, M. J., and Easteal, S. (2007). Rates of Genome Evolution and Branching Order from Whole Genome Analysis. *Molecular Biology and Evolution*, 24(8):1722–1730.

Iannelli, F., Pesole, G., Sordino, P., and Gissi, C. (2007). Mitogenomics reveals two cryptic species in ciona intestinalis. *Trends Genet*, 23(9):419–422.

Imai, K. S., Satoh, N., and Satou, Y. (2002). Early embryonic expression of fgf4/6/9 gene and its role in the induction of mesenchyme and notochord in ciona savignyi embryos. *Development*, 129(7):1729–1738.

Jabbari, K. and Bernardi, G. (2004). Cytosine methylation and cpg, tpg (cpa) and tpa frequencies. *Gene*, 333:143 – 149.

Jabbari, K., Cacci, S., de Barros, J. P. P., Desgrs, J., and Bernardi, G. (1997). Evolutionary changes in cpg and methylation levels in the genome of vertebrates. *Gene*, 205(1-2):109 – 118. Junk DNA: The Role and the Evolution of Non-Coding Sequences.

Johnson, D. S., Davidson, B., Brown, C. D., Smith, W. C., and Sidow, A. (2004). Noncoding regulatory sequences of ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Research*, 14:2448–2456.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8(3):275–282.

Jordan, I. K., Kondrashov, F. A., Adzhubei, I. A., Wolf, Y. I., Koonin, E. V., Kondrashov, A. S., and Sunyaev, S. (2005). A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433(7026):633–638.

Kano, S., Satoh, N., and Sordino, P. (2006). Primary genetic linkage maps of the ascidian, ciona intestinalis. *Zoological Science*, 23(1):31–39.

Kimbacher, S., Gerstl, I., Velimirov, B., and Hagemann, S. (2009). <i>drosophila</i><i>p</i> transposons of the urochordata <i>ciona intestinalis</i>. *Molecular Genetics and Genomics*, 282:165–172. 10.1007/s00438-009-0453-7.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–626.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

Lambert, C. and Lambert, G. (1998). Non-indigenous ascidians in southern california harbors and marinas. *Marine Biology*, 130(4):675–688.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948.

Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E., and Audit, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics*, 10(1):335.

Li, W.-H., Gojobori, T., and Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature*, 292(5820):237–239. 10.1038/292237a0 10.1038/292237a0.

Liu, X., Zhang, J., Ni, F., Dong, X., Han, B., Han, D., Ji, Z., and Zhao, Y. (2010). Genome wide exploration of the origin and evolution of amino acids. *BMC Evolutionary Biology*, 10(1):77.

Lobry, J. R. (1997). Influence of genomic g+c content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, 205(1-2):309–316.

Lobry, J. R. and Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 escherichia coli chromosome-encoded genes. *Nucleic Acids Res*, 22(15):3174–3180.

Loewe, L. and Hill, W. G. (2010). The population genetics of mutations: good, bad and indifferent. *Philos Trans R Soc Lond B Biol Sci*, 365(1544):1153–1167.

Marsolier-Kergoat, M.-C. and Yeramian, E. (2009). GC Content and Recombination: Reassessing the Causal Effects for the Saccharomyces cerevisiae Genome. *Genetics*, 183(1):31–38.

Martin, A. P. and Palumbi, S. R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *PNAS*, 90(9):4087–4091.

McDonald, J. H. (2006). Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation. *Mol Biol Evol*, 23(2):240–244.

Meinertzhagen, I. A. and Okamura, Y. (2001). The larval ascidian nervous system: the chordate brain from its small beginnings. *Trends in Neurosciences*, 24(7):401–10. 0166-2236 (Print) Journal Article Research Support, Non-U.S. Gov't Review.

Monniot, C. and Monniot, F. (1994). Additions to the inventory of eastern tropical atlantic ascidians - arrival of cosmopolitan species. *Buletin of marine science*, 54(1):71–93.

Montes, L., Le Roy, N., Perret, M., De Buffrenil, V., Castanet, J., and Cubo, J. (September 2007). Relationships between bone growth rate, body mass and resting metabolic rate in growing amniotes: a phylogenetic approach. *Biological Journal of the Linnean Society*, 92:63–76(14).

Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H., and Bernardi, G. (2001). Translational selection on codon usage in xenopus laevis. *Mol Biol Evol*, 18(9):1703–1707.

Nakatani, Y., Moody, R., and Smith, W. (1999). Mutations affecting tail and notochord development in the ascidian Ciona savignyi. *Development*, 126(15):3293–3301.

Nielsen, C. (2001). *Animal Evolution*. Oxford Univ. Press, 2nd edition edition.

Nishida, H. (1987). Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. iii. up to the tissue restricted stage. *Dev Biol*, 121(2):526–541.

Ohno, S. (1970). *Evolution of gene duplication*. Springer-Verlag, New York.

Ohta, T. (1993). An examination of the generation-time effect on molecular evolution. *PNAS*, 90:10676–10680.

Peterson, K. J., Lyons, J. B., Nowak, K. S., Takacs, C. M., Wargo, M. J., and McPeek, M. A. (2004). Estimating metazoan divergence times with a molecular clock. *PNAS*, 101(17):6536–6541.

Philippe, H. and Telford, M. J. (2006). Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol*, 21(11):614–20. 0169-5347 (Print) Journal Article Research Support, Non-U.S. Gov't.

Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A., and Pbo, S. (2005). Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 37(4):429–434.

Putnam, N. H., Butts, T., Ferrier, D. E., Fulong, R. F., Hellsten, U., Kawashima, T., and etc (2008). The amphioxus genome and the evolution of chordate karyotype. *Nature*, 453:1064–1072.

Rosenberg, M. S., Subramanian, S., and Kumar, S. (2003). Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. *Mol Biol Evol*, 20(6):988–993.

Roseto, G. d. L. i., Bucciarelli, G., and Bernardi, G. (2002). An analysis of the genome of ciona intestinalis. *Gene*, 295:311–316.

Saccone, S., C., F., and G., B. (2002). Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene*, 300:169–178(10).

Sarich, V. M. and Wilson, A. C. (1973). Generation time and genomic evolution in primates. *Science*, 179(78):1144–1147.

Satoh, N. (2001). Ascidian embryos as a model system to analyze expression and function of developmental genes. *Differentiation*, 68(1):1–12.

Satoh, N. (2003). The ascidian tadpole larva: comparative molecular development and genomics. *Nat Rev Genet*, 4(4):285–95. 1471-0056 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't Review.

Sharp, P. M. and Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution*, 24:28–38. 10.1007/BF02099948.

Shoguchi, E., Kawashima, T., Satou, Y., Hamaguchi, M., Sin, I. T., Kohara, Y., Putnam, N., Rokhsar, D. S., and Satoh, N. (2006). Chromosomal mapping of 170 bac clones in the ascidian ciona intestinalis. *Genome Res*, 16(2):297–303. 1088-9051 (Print) Journal Article Research Support, Non-U.S. Gov't.

Simmen, M. W., Leitgeb, S., Charlton, J., Jones, S. J. M., Harris, B. R., Clark, V. H., and Bird, A. (1999). Nonmethylated transposable elements and methylated genes in a chordate genome. *Science*, 283(5405):1164–1167.

Simmen, M. W., Leitgeb, S., Clark, V. H., Jones, S. J., and Bird, A. (1998). Gene number in an invertebrate chordate, ciona intestinalis. *Proc Natl Acad Sci U S A*, 95(8):4437–40. 0027-8424 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't.

Singh, T., Tsagkogeorga, G., Delsuc, F., Blanquart, S., Shenkar, N., Loya, Y., Douzery, E., and Huchon, D. (2009). Tunicate mitogenomics and phylogenetics: peculiarities of the herdmania momus mitochondrial genome and support for the new chordate phylogeny. *BMC Genomics*, 10(1):534.

Small, K., Brudno, M., Hill, M., and Sidow, A. (2007). A haplome alignment and reference sequence of the highly polymorphic ciona savignyi genome. *Genome Biology*, 8(3):R41.

Smit, A. and Hubley, R. (1996). Repeatmodeler open-1.0.

Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., Angerer, R. C., Angerer, L. M., Arnone, M. I., Burgess, D. R., Burke, R. D., Coffman, J. A., Dean, M., Elphick, M. R., Ettensohn, C. A., Foltz, K. R., Hamdoun, A., Hynes, R. O., Klein, W. H., Marzluff, W., McClay, D. R., Morris, R. L., Mushegian, A., Rast, J. P., Smith, L. C., Thorndyke, M. C., Vacquier, V. D., Wessel, G. M., Wray, G., Zhang, L., Elsik, C. G., Ermolaeva, O., Hlavina, W., Hofmann, G., Kitts, P., Landrum, M. J., Mackey, A. J., Maglott, D., Panopoulou, G., Poustka, A. J., Pruitt, K., Sapojnikov, V., Song, X., Souvorov, A., Solovyev, V., Wei, Z., Whittaker, C. A., Worley, K., Durbin, K. J., Shen, Y., Fedrigo, O., Garfield, D., Haygood, R., Primus, A.,

Satija, R., Severson, T., Gonzalez-Garay, M. L., Jackson, A. R., Milosavljevic, A., Tong, M., Killian, C. E., Livingston, B. T., Wilt, F. H., Adams, N., Belle, R., Carbonneau, S., Cheung, R., Cormier, P., Cosson, B., Croce, J., Fernandez-Guerra, A., Geneviere, A. M., Goel, M., Kelkar, H., Morales, J., Mulner-Lorillon, O., Robertson, A. J., Goldstone, J. V., Cole, B., Epel, D., Gold, B., Hahn, M. E., Howard-Ashby, M., Scally, M., Stegeman, J. J., Allgood, E. L., Cool, J., Judkins, K. M., McCafferty, S. S., Musante, A. M., Obar, R. A., Rawson, A. P., Rossetti, B. J., Gibbons, I. R., Hoffman, M. P., Leone, A., Istrail, S., Materna, S. C., Samanta, M. P., Stolc, V., Tongprasit, W., et al. (2006). The genome of the sea urchin strongylocentrotus purpuratus. *Science*, 314(5801):941–52. 1095-9203 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

Sordino, P., Belluzzi, L., Santis, R. D., and Smith, W. C. (2001). Developmental genetics in primitive chordates. *Philos Trans R Soc Lond B Biol Sci*, 356(1414):1573–1582.

Stach, T. (2008). Chordate phylogeny and evolution: a not so simple three-taxon problem. *Journal of Zoology*, 276(2):117 – 141.

Suzuki, M. M., Kerr, A. R. W., Sousa, D. D., and Bird, A. (2007). Cpg methylation is targeted to transcription units in an invertebrate genome. *Genome Res*, 17(5):625–631.

Suzuki, M. M., Nishikawa, T., and Bird, A. (2005). Genomic approaches reveal unexpected genetic divergence within ciona intestinalis. *Molecular Evolution*, 61:627–635.

Swalla, B. J., Cameron, C. B., Corley, L. S., and Garey, J. R. (2000). Urochordates are monophyletic within the deuterostomes. *Systematic Biology*, 49(1):52–64.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135(2):599–607.

Takahashi, H., Hotta, K., Erives, A., Gregorio, A. D., Zeller, R. W., Levine, M., and Satoh, N. (1999). Brachyury downstream notochord differentiation in the ascidian embryo. *Genes Dev*, 13(12):1519–1523.

Takashi, A. N., Chiyo, N., Shinji, M., Maki, H., Chihiro, H., Shonan, A., and Hirosuke, F. (1997). Embryonic thermosensitivity of the ascidian, ciona savignyi. *Zoological Science*, 14(3):511–515.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., and Nikolskaya, A. N. (2003). The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29:22 – 28.

Telford, M. J. and Holland, P. W. (1993). The phylogenetic affinities of the chaetognaths: a molecular analysis. *Mol Biol Evol*, 10(3):660–676.

Therriault, T. W. and Herborg, L.-M. (2008). Predicting the potential distribution of the vase tunicate ciona intestinalis in canadian waters: informing a risk assessment. *ICES J. Mar. Sci.*, 65(5):788–794.

Tsagkogeorga, G., Turon, X., Galtier, N., Douzery, E., and Delsuc, F. (2010). Accelerated evolutionary rate of housekeeping genes in tunicates. *Journal of Molecular Evolution*, 71:153–167. 10.1007/s00239-010-9372-9.

Tsagkogeorga, G., Turon, X., Hopcroft, R., Tilak, M.-K., Feldstein, T., Shenkar, N., Loya, Y., Huchon, D., Douzery, E., and Delsuc, F. (2009). An updated 18s rrna phylogeny of tunicates based on mixture and secondary structure models. *BMC Evolutionary Biology*, 9(1):187.

Uliano, E., Chaurasia, A., Bern, L., Agnisola, C., and D'Onofrio, G. (2010). Metabolic rate and genomic gc. what we can learn from teleost fish. *Marine Genomics*, 3(1):29 – 34.

Varriale, A. and Bernardi, G. (2006). Dna methylation in reptiles. *Gene*, 385:122–127.

Vienne, A. and Pontarotti, P. (2006). Metaphylogeny of 82 gene families sheds a new light on chordate evolution. *Int J Biol Sci*, 2(2):32–37.

Vinogradov, A. E. (2001). Intron length and codon usage. *J Mol Evol*, 52(1):2–5.

Vinogradov, A. E. (2005). Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Research*, 33(2):559–563.

Vinson, J. P., Jaffe, D. B., O'Neill, K., Karlsson, E. K., Stange-Thomann, N., Anderson, S., Mesirov, J. P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J. E., and Lander, E. S. (2005). Assembly of polymorphic genomes: algorithms and application to ciona savignyi. *Genome Res*, 15(8):1127–35. 1088-9051 (Print) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

Wada, H. (1998). Evolutionary history of free-swimming and sessile lifestyles in urochordates as deduced from 18s rdna molecular phylogeny. *Mol Biol Evol*, 15(9):1189–1194.

Wei, W., Davis, R. E., Jomantiene, R., and Zhao, Y. (2008). Ancient, recurrent phage attacks and recombination shaped dynamic sequence-variable mosaics at the root of phytoplasma genome evolution. *Proceedings of the National Academy of Sciences*, 105(33):11827–11832.

Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J., and Mallatt, J. (2002). Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new lsu and ssu ribosomal dna data. *Mol Biol Evol*, 19(5):762–776.

Zvyagintsev, A. Y., Sanamyan, K. E., and Kashenko, S. D. (2007). On the introduction of the ascidian ciona savignyi herdman, 1882 into peter the great bay, sea of japan. *Russian Journal of Marine Biology*, 33:133–136.