U.PORTO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

universidade de aveiro
theoria poiesis praxis

Evolutionary characterization of genes involved in development and adaptation in vertebrates under differential environmental conditions of selective pressure

João Paulo Rodrigues Machado

Tese de doutoramento em Ciências do Mar e do Ambiente

2014

João Paulo Rodrigues Machado

Evolutionary characterization of genes involved in development and adaptation in vertebrates under differential environmental conditions of selective pressure

Tese de Candidatura ao grau de Doutor em Ciências do Mar e do Ambiente Especialidade em Oceanografia e Ecossistemas Marinhos;

Programa Doutoral da Universidade do Porto (Instituto de Ciências Biomédicas de Abel Salazar e Faculdade de Ciências) e da Universidade de Aveiro.

Orientador – Professor Doutor Agostinho Antunes

Categoria – Investigador Auxiliar e Professor Auxiliar Convidado

Afiliação – Centro Interdisciplinar de Investigação Marinha e Ambientar e Faculdade de Ciências da Universidade do Porto.

Coorientador – Professor Doutor Vitor Vasconcelos

Categoria – Professor Catedrático

Afiliação – Centro Interdisciplinar de Investigação Marinha e Ambientar e Faculdade de Ciências da Universidade do Porto.

Acknowledgements

First I would like to thank my supervisor Dr. Agostinho Antunes and my co-supervisor Prof. Dr. Vitor Vasconcelos, for their acceptance to guide me in the course of the work presented here. I would like to thank them also for their cheer up words in my upset moments, and particularly for giving me the privilege of sharing their experience, inside and outside science, reminding me the long walk that I need to endeavor. I would like also to thank all the co-authors of each chapter for their persistence and patience to read and correct, building improved versions step by step. A special thanks to Warren E. Johnson for the critical review in several chapters of this thesis proposal.

I would like also to thank everyone that one way or another has helped me during this period, namely LEGE and PROMAR colleagues. Despite all the help and cheer up that you can provide to others, the time period span in thesis proposal ends almost all the time with enormous amount of emotional debts to others. Particularly for their patience in supporting me in the darkest points of this road. A special thanks to Zé Carlos, Marisa Silva, Siby Phillip, Flávio Alves, Rui Borges and Ana Rocha for their words. To my family, particularly for their comprehension and patience to be with me, despite my absence… I surely would not complete this alone.

I would like to acknowledge the financial support provided from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism during the first years and later the Portuguese Fundação para a Ciência e Tecnologia (FCT) under the grant SFRH/BD/65245/2009. I would also like to acknowledge the project PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490) from FCT.

Abstract

Natural selection is a mechanism that leads to genetic change among species and therefore, understanding the molecular basis of adaptation is a central goal of evolutionary biology. The comparison among vertebrate species allows the detection of natural selection acting on their genes and genomes. Those observed differences may therefore explain the genetic basis of their adaptation to the surrounding environment and also reveal mechanisms of evolutionary novelty.

This thesis was focused in the detection of natural selection in genes, and later extended to processed pseudogenes, mirroring species under different evolutionary pressure. For this purpose several different processes were surveyed, namely bone-associated genes changes through flight or genes involved in mammalian tooth diversification. Later, the contribution of gene duplications was studied in species adaptation through a gene involved in iron homeostasis, and also implicated in temperature adaptation in teleosts. This thesis was mainly conducted to detected natural selection mirroring coding regions, however, in light of the most recent work in molecular biology it was extended to search signatures of natural selection beside coding genes, namely processed pseudogenes.

The original contributions of this thesis to understand the adaptive evolution of vertebrates included:

- Detection of selection signatures in Matrix extracellular phosphoglycoprotein (MEPE)outside the previous known functional motifs (dentonin and the ASARM) that might be of crucial relevance for MEPE function in vertebrates;

- Characterize the genes responsible for the diversification of mammalian dentition, finding that “new” genes have a higher evolutionary rate, therefore likely having a major contribution in the mammalian dentition diversification;

- Impact of flight in bone associated genes in birds and bats, showing an extensive selection in bone-remodeling genes, likely associated with flight adaptation. Furthermore their involvement in other functions, suggests that the presence of positive selection may also influence other key features for flight adaptation such as hyperglycemia tolerance or muscle development;

- Importance of positive selection in the functional diversification of the two WAP65 copies present in some teleosts, showing the role of the functional divergence in the retention of the two copies and their functional distinctiveness;

- The potential adaptive value of processed pseudogenes, since they remain potentially active and subjected to similar selective events as the coding sequences counterparts of fully active genes.

Summarizing, the results presented here show that the diversification of life habits, diets and mode of locomotion present signatures of positive Darwinian selection on several gene protein encoding regions in Vertebrates. Although, besides coding regions, is shown the role of intronic regions and pseudogenes in the adaptation to different selective regimes.

Keywords:

Adaptive Evolution, Vertebrates, Positive Selection, Gene Duplication, Evolutionary Novelty

Resumo

A Seleção Natural é um mecanismo que leva às diferenças genéticas entre espécies, sendo a compreensão da base molecular da adaptação um objetivo primordial da Biologia Evolutiva. A comparação entre diferentes espécies de vertebrados permite a deteção de assinaturas de seleção natural nos genes e genomas. Essas diferenças podem explicar a base genética da adaptação ao meio ambiente e também os mecanismos responsáveis pelas novidades evolutivas.

Esta tese focou-se na deteção da ação da seleção natural em genes e por último em pseudogenes processados em espécies sob diferentes pressões evolutivas.

Para este efeito foram estudados vários processos diferentes, tais como, alterações genéticas ósseas relativas ao voo nas aves ou genes envolvidos na diversificação dos dentes dos mamíferos. Mais tarde, foi estudado a contribuição de duplicações de genes na adaptação de espécies através de um gene envolvido na homeostasia do Ferro, que também está implicado na adaptação à temperatura em teleósteos. Esta tese teve como objetivo a deteção de regiões codificantes alvo de seleção natural. No entanto, tendo em conta recentes trabalhos em evolução molecular ampliamos a busca da seleção natural aos pseudogenes processados além das regiões codificantes.

A maior contribuição da tese para o entendimento da evolução adaptativa em vertebrados foi:

- Deteção de assinaturas de seleção do MEPE fora dos motivos funcionais conhecidos anteriormente (“dentonin” e o ASARM), que podem ser relevantes para a função fundamental do MEPE em vertebrados;

- Compreender os genes responsáveis pela diversificação da dentição em mamíferos, mostrando uma taxa de evolução maior em genes “novos”, e desse modo potencialmente associados à variedade de fenótipos observada na dentição dos mamíferos;

- Impacto do voo em genes associados à ossificação em aves e morcegos, mostrando fortes evidencias de seleção em genes associados à ossificação, e desse

modo associados a adaptações ao voo. Além disso o envolvimento em outras funções sugere que a seleção positiva também pode estar associada a outras características da adaptação ao voo, tais como o desenvolvimento do músculo e a tolerância a níveis altos de glicémia.

- Envolvimento da seleção positiva na distinção funcional das duas cópias do WAP65 presentes em alguns teleósteos, tornando evidente o papel da divergência funcional na retenção das duas cópias;

- O valor do potencial adaptativo dos pseudogenes processados, pois parecem poder continuar ativos e sujeito a eventos de seleção similar aos observados em regiões codificantes.

Resumindo, os resultados aqui apresentados demonstram que a diversificação de hábitos de vida, dietas e modos de locomoção apresentam assinaturas de seleção positiva, ou darwiniana, em várias regiões codificantes de proteínas nos vertebrados. De igual forma é demonstrado o papel das regiões intronicas e pseudogenes na adaptação a diferentes regimes seletivos.

Thesis Publications

This Thesis includes two scientific papers published in international journals and three manuscripts in preparation resulting from part of the obtained results:

- *Adaptive evolution of the Matrix Extracellular Phosphoglycoprotein in mammals,* BMC Evolutionary Biology, Machado JP, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A., 2011, 11:342, doi:10.1186/1471-2148-11-342. (published)

- *Convergent selection in bone-associated genes modeled through the evolution of flight in birds and bats* (will be submitted for publication)

- *Role of positive selection and recent gene duplication on generation of novelty in Mammalian dentition patterns* (will be submitted for publication)

- *Adaptive Functional Divergence of the Warm Temperature Acclimation-Related Protein (WAP65) in Fishes and the Ortholog Hemopexin (HPX) in Mammals*. Journal of Heredity, Machado JP, Vasconcelos V, Antunes A., 2014, 105 (2): 237-252, doi: 10.1093/jhered/est087 (published)

- *Processed pseudogenes close to parent gene, or under a favourable expression context have higher chances to be functionally relevant* (will be submitted for publication)

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Pre-Darwian Perspective

Early evolutionary theories emerged during classical Greek period when Anaximander suggested that ancestors of humans had been born in water, and therefore their origin was from fishes (Cleve 1965). Later was proposed by Empedocles that extant species evolved by elements combinations, and was natural selection that led to the extinction of "monstrous" organisms (Zirkle 1941). Although the evolutionary thinking, stating that species may change over time, had its roots in antiquity, the occidental biological thinking spread more the theories of Plato and Aristotle. Their ideas were opposed to evolution, becoming more influential in western during the middle ages, where Creationism was the essentialist dogma stating that species were fixed and created by a divine design.

Before Charles Darwin, there was an unformed or unstated evolutionary thinking, with the first staid attempt being brought by a French naturalist Jean-Baptiste Lamarck, with his work *Philosophie Zoologique* in 1809. Lamarck argued that species change over time leading to a new species. Conceptually was a challenge to enrooted idea of creationism or “fixism”. Despite the similarity in the core idea, the way that Lamarck described his idea was fundamentally different from the modern Darwinian view of evolution. As Lamarck proposed that lineages persists indefinitely and changes over time with a sort of unknown force, an “internal force”, within the organism to produce offspring slightly different from itself, and after many generations the lineage would be visibly different from the ancestors. Through generations and accumulating enough differences relatively to the ancestors, then would be eventually considered as the emergence of a new species. Since he pictured evolution as a successive accumulation of transformations in “living-forms” from the very simple to a more and more sophisticated ones, his idea is often described as “transformism”. These ideas were mainly based on two key mechanisms: use and disuse (loss of futile characteristics and development of useful ones) and the inheritance of traits. He has although reintroduced an important idea in the occidental perspective about evolution, the mechanism of inheritance of acquired characters as a response to changing environment. The term “character” is as short-hand for “characteristic”, meaning any property distinguishable of an

organism. The classical example of Lamarck idea is the giraffe's neck, where he explains that ancestral giraffes have extended their necks to reach leaves in up trees; exertion caused their neck to become slightly longer over generations. The inherited longer necks in offspring, were stretch longer over generations resulting in longer necks. Despite the current caricatured of Lamarck idea that evolution happen by some sort of "will of the organism", he have not proposed that. Instead he proposed that evolution occurs given the flexibility in individual development and the inherence of the acquired characters, and not driven by the conscious striving on an organism, or part of organism (e.g. giraffe's neck). Despite conceptually wrong Lamarck, since he argued that species does not extinguish but rather evolved in new species, he have challenge the idea that species are fixed, and brought to his time the older idea that species may change over time from Aristotle. His ideas were opposed by the anatomist George Cuvier, which stated the "fixity" of species but also claimed that they may become extinct. The Cuvier's idea was deeply embedded in western thought, were the majority of biologist and geologist accept his idea that species had a separated origin and then remained constant until they may went extinct.

1.2 Darwin and Post-Darwin ideas

A new chapter in the study of evolution of species was opened by a young naturalist, Charles Darwin, while observing animals and plants that inhabited the Galapagos Islands on board of Beagle (1832-37). Darwin noticed how animals, such as tortoise, mockingbird, and finch, differed morphologically from island to island. Years after the Beagle voyage, Darwin was compiling his observations, formulating the idea that for instance finches were initially the same species, and the colonization of different islands leaded to different specializations from island to island, and then finches became distinct species, despite sharing a common ancestry. Additional clear evidence was from the ostrich-like birds, rhea, different from one region to another in South America. These geographic observations were probably the first evidence for Darwin to accept that species can change.

Darwin refused contemporary explanations for the species change over time, including Lamarckism, since all fail to explain adaptation. In his search for a plausible explanation he formulated the theory that species, in struggle for existence, those forms better adapted tend to leave more offspring and lead to an increased frequency in next generations. Since the environment may change, then species that have better capability of adaptation to those new environmental conditions will leave more offspring, leading to and increased frequency in those forms, and in opposite, those poorly adapted will leave less offspring. Over generations this would lead to the formation of "new species". The publication of Charles Darwin's

book “On the Origin of Species" (Darwin 1859) was a landmark in evolutionary biology answering the question how species change over time and how is related with changes in environment, connecting two theories, evolution and natural selection. Simultaneously, Alfred Wallace has independently reached similar conclusions through his own work on natural selection and therefore is often considered co-discoverer of evolution by natural selection (Morrison 2014).

Although the theory of evolution and natural selection led to an intense debate in the following years after the initial formulation underwent many revisions and modifications. Darwin’s theory of evolution was generally and initially almost immediately accepted among biologists but not his explanations of the natural selection and several different alternative explanations were in debate. The most claimed reason for the objection was the theory of heredity, since the works from Gregor Mendel seem incompatible with Darwin’s theory. Although years later the rediscovery of Mendel’s laws showed that the variation in populations was mainly caused by mutation. For the first time, R.A. Fisher described a statistical model for quantitative inheritance in 1918, and during 1920s and 1930s, R.A. Fisher, J.B.S. Haldane and S. Wright develop the fundamental principles of quantitative genetics. Based on a rigorous mathematical approach they demonstrated that Mendelian heredity and natural selection framework were compatible. From the conciliation of both theories arose the Neo-Darwinism or the synthetic theory of evolution, recognizing mutations as the ultimate source of genetic variation, and natural selection as the major force shaping the genetic makeup. In the following years this conception become gradually more widespread in all areas of biology and widely accepted leading to the provocative essay of Theodosius Dobzhansky, "Nothing in Biology Makes Sense Except in the Light of Evolution" (Dobzhansky 1973).

The theory of evolution proposed by Darwin *in sensu lato* states that the successful establishment of an organism to a specific habitat and/or environment may be attributed to traits (i.e. phenotypes). It favors the survival of those traits better adapted, during the course of generations increasing their frequency over generations (fitness). In an evolutionary perspective, these traits are called as adaptive if they are favorable for a given environmental condition (e.g. temperature) or new life habit (e.g. flight).

1.3 Molecular Evolution: a new Era

The discovery of the DNA (Deoxyribose Nucleic Acid) structure by Watson and Crick in 1953 revealed the molecule carrying hereditary information and the advances of molecular techniques enabled the beginning of a new era in biological research. Technical advances allowed to accumulate more and more empirical molecular data, used to examine the Genetical theory, and to search for evidences of adaptation. This allowed studies of

adaptive mechanisms in the organisms at the molecular level. The DNA molecule is a physical mechanism that holds the “life code” and allows the vertical transmission to the descendants of that information. Structurally the DNA is composed by nucleotides that consist in a phosphate and a sugar group with a base attached (A, T, G or C). In a broad perspective, proteins are the organism’s building blocks that are encoded by DNA, which is organized in genes (coding sequences and noncoding regions) and intergenic regions. Genes encode trough an intermediate (mRNA) the genetic instructions, and are molecular unit of heredity, transmitted, “packed”, in chromosomes to offspring. Since there are only four nucleotides and 20 amino acids, a one-to-one code would be impossible. The triplet of bases, also called codon, encodes one amino acid, and the relation between the triple and the coded amino acid is called the genetic code.

In Molecular Evolution the majority of the studies are conducted in coding sequences of the DNA, since is easier to functional characterized them and their predictable higher conservation than noncoding sequences, resulting in an easier annotation and comparability between species. Although most of the DNA is either intragenic or intergenic “noncoding”, being only around 2.8% of the total amount of it is exonic (coding) (Alexander, Fang et al. 2010), recently increasing arguments towards non coding functionality and adaptive value are raising (Andolfatto 2005; Ponting and Lunter 2006; Haygood, Babbitt et al. 2010) and the old term “junk” DNA is now slowly been abandoned as more functional evidences come out (Wells 2013). Another reason for the higher amount of studies conducted in coding regions is the larger amount of comparative methods/tools for coding sequences based on the degeneracy of the genetic code. Evolution, at the molecular level, is observable as nucleotides changes in DNA and amino acid changes in proteins. The term substitution is often used to refer an evolutionary change, and is detected when comparing different species. Mutations are referred to a base change in one individual and happen at DNA level. The mutations (evolutionary changes) within protein coding gene can alter the amino acid (also called non-synonymous substitution), or left the amino acid unchanged (synonymous or silent mutation), therefore the effects of mutations are accessed at protein level. Alternatively the molecular evolution can also be studied within populations, looking at polymorphisms that can be either results of natural selection or drift.

The fixation of adaptive mutations in evolving populations was a central theme in Darwin’s theory, and until the late 1960’s, most evolutionary biologists supported the idea that the majority of a population variations were maintained through balancing selection, a form of selection through which two or more alleles for the same locus are maintained in a population. However, the increasing protein sequences availability from several mammals was progressively changing a few paradigms. The number of amino acid differences between different lineages in hemoglobin protein sequences was roughly proportional within

time, estimated from fossil evidence (Zuckercandl and Pauling 1965), to explain this linearity, given a constant rate for amino acid substitution variation within time. This linearity observation together with observation of genetic equidistance in cytochrome c (Margoliash 1963), resulted in the formal postulation of the molecular clock hypothesis in the early 1960's (Kumar 2005). Later, protein electrophoresis studies (Harris 1966; Lewontin and Hubby 1966) demonstrated a widespread presence of polymorphisms within populations. Although both observations were highly controversial, given the difficulty to accommodate an evolutionary mechanism that on one hand allows a constant substitution rate and on the other whereby natural selection could maintain high levels of polymorphisms within populations (Haldane 1957).

The early evidences lead Kimura (Kimura 1968) and King & Jukes (King, Jukes et al. 1969) to independently formulate, what the former author called neutral theory of evolution. The neutral theory does not suggest that random drift explains all evolutionary changes, and natural selection is still need to explain adaptation. The neutral theory of evolution is although opposed the Darwinian in the concept of natural selection as major force for evolutionary change and variability at the molecular level. Kimura's theory suggests that at molecular level the random fixation of mutations is selectively neutral (later refined to nearly neutral), and that the effect of natural selection was insignificant. Consequently in this theory is stated that evolution at DNA and protein level are dominated by random processes, being most evolution at molecular level non-adaptive. If the majority of nucleotide and amino acid substitutions were selectively neutral, then substitution may occur at a fairly constant rate and high levels of polymorphism could be maintained since the selective cost would be low.

Neutral theory had a progressive influence on the current state of molecular evolution, since it suggests that at the molecular level, stochastic processes are dominant rather than deterministic processes, and therefore differ greatly from "selectionist" theory (Figure 1-1).

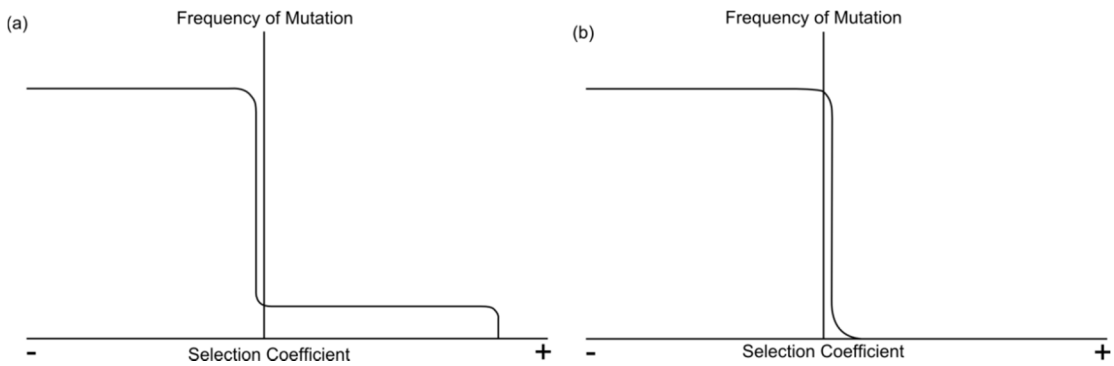

Figure 1-1. Comparison between neutral and selection theory. Mutations and the beneficial measures as selection coefficient from deleterious (-) to advantageous (+): (a) selection theory and (b) neutral theory.

Nowadays many aspects of the neutral theory are still accepted, and there is a consensus that both weak deleterious selection and occasional positive selection are important evolutionary factors. Therefore adaptation should not be just assumed, but should be rigorously tested and proved. Detecting natural selection in genomes became an efficient strategy for finding causes of interspecific differences or identifying genomic regions of prime functional relevance for species adaption.

1.4 Adaptive Selection

As previously explained, species change over time, and it is consensual that the theory of evolution proposed by Darwin among scientific community, opened the question in the severity of the impact that natural selection have on genes and/or genomes. In an evolutionary view, adaptation can be defined as the functional modifications of structure, physiology, and behavior of an organism that increases species fitness, i.e. increase its chances of survival in a particular ecological niche. Therefore species adaptation can be considered by the conformity between an organism and selective pressure of the environment. Adaptive evolution, often called positive selection, is the process by which a beneficial allele, translating either in higher reproduction or survival rates, that increases its frequency, given the augmentation in fitness for individuals carrying that allele. The term positive selection (also referred as Darwinian Selection) is often used at interspecific level, but at population level is called selective sweep, when the occurrence of a new mutation increases the fitness of the carrier relative to other members of a certain population, and over time gradually becomes dominant.

Recent advances in genomic sequencing and computational analyses (either available tools or computational capacity) have changed the perception on molecular basis of adaptation. Previously was rare to uncover evidence that a gene had been subjected to adaptive evolution at the molecular level, but now this is becoming increasingly common (Swanson 2003). Nowadays, this is more easily tested given the recent advances in the field, like the large influx of genomic sequences databases available in public repositories, such as Ensembl or GenBank, generated from genome sequencing projects. These new resources combined with the development of new statistical analyses provide a diversified amount of possibilities to test adaptive selection. More recently the scale mirroring a single gene in evolutionary analysis is gradually increasing to a larger scale, in projects using supercomputers to create databases such as Selectome (Proux, Studer et al. 2009) or reduction in costs of sequencing to compare entire genomes creating informative high-resolution genome maps (Lindblad-Toh, Garber et al. 2011). The identification of genes and gene regions subjected to positive selection can lead to predictions regarding the putative

functionally important regions within genes but also their prime role in species adaptation. Several examples explaining the relation between the statistical tests and selective constrains are available, from mammalian secondary adaptations to aquatic environment (Sun, Zhou et al. 2013; Zhou, Sun et al. 2013), to adaptations in flight species (Zhang and Edwards 2012) or modifications in diet (Zhang, Zhang et al. 2002). All mentioned results where changes in life habits that lead detectable trails in genomes, referred as adaptive selection. Although the more remarkably examples of such alterations in the selective pressure regimes shaping genomes are those obtained for parallel evolution, where isolated population of fishes reach a similar traits facing alteration of salinity (Colosimo, Hosemann et al. 2005), or the examples of convergent evolution of echolocation in mammals (Parker, Tsagkogeorga et al. 2013). These examples constitute a good lesson how adaptive evolution operates in shaping the gene/genomes, where the organism found independently similar strategies, showing a clear relation between molecular evolution and the environment or life habit.

1.4.1 Detecting adaptive selection

1.4.2 Theoretical conception

The methods to distinguish selection are broadly of two types: (1) those that focus on divergence of genes between orthologues; and (2) those methods applied to test alteration in polymorphisms frequency within population of a given species. At population level there are various methods to test positive selection, Tajima's D test (Tajima 1989), Hudson–Kreitman–Aguade (or HKA test) (Hudson, Kreitman et al. 1987) and HKA successor McDonald-Kreitman test (McDonald and Kreitman 1991). For this thesis was primarily employed a model to study lineages and sites (not at population level) evolving under positive selection in a phylogenetic perspective.

Genes and genomes are therefore not immutable units, and there are four typical changes that may occur in the DNA, broadly called mutations, such as: insertions, deletions, inversions and substitutions. Despite all being phylogenetically informative and of key relevance in molecular evolution, the work here was more focused on substitutions, and alterations of the mutation rate within coding regions and non-coding regions of genes.

Given the nature of the substitutions in amino acid coding genes they are classified into two different categories: 1) if it does not change the amino acid which the codon codes for (synonymous) or; 2) if alters the amino acid that the codon codes for (nonsynonymous) due to the degeneracy of the genetic code. If the replacement increases the fitness of the

carrier then will be transmitted to the future generations, increasing its frequency within population and become established. This is the most direct way to obtain evidences for adaptive molecular evolution on a protein-coding gene, by identifying amino acid sites where the rate of nonsynonymous (or replacement) substitutions (d_N) suppress the rate of the synonymous (or silent) rate (d_S). Since synonymous mutations have no effect in protein sequence, they are assumed to be not subjected to natural selection. This is only partial true since in mammals synonymous mutations have been found to have an effect on species fitness. Such mutations can disrupt splicing, in alternatively spliced exons, and interfere with mRNA binding (Hurst 2006) or even modify structure and activity of the protein (Kimchi-Sarfaty, Oh et al. 2007). However, synonymous substitutions influence more easily low-expressed genes, as well as a very small proportion of these substitutions may affect mRNA stability (Hurst 2006).

The detection of adaptation has been a subject of prime relevance and intense interest in evolutionary molecular Biology and the increasing availability of DNA sequences from closely related organisms allow comparisons of their encoded protein sequences. The current methods do detect positive selection at interspecific level, inhabiting different ecological niches does not accommodate Single Nucleotide Polymorphisms (SNPs), this implies that is used a single nucleotide sequence for each species, assuming that its representative of a certain gene in a species. Although this is a simplification since several genes show intraspecific variability (Renaut, Nolte et al. 2010). On the other hand, d_N/d_S approach is also frequently incorrectly applied to genetic sequences sampled from a single population, but the differences among the population represent polymorphisms and not established substitutions (Kryazhimskiy and Plotkin 2008).

Thus, the comparison of the fixation rates of synonymous and nonsynonymous mutations can be used to understand the action of selective pressure on protein coding sequences between species. Selective pressure at the gene encoding level is a more reliable approach to measure the ratio dN/dS (also called ω). When $\omega < 1$ indicates that the amino acid changes are deleterious, therefore under purifying selection and the fixation replacement rate is low. Although if a replacement on the protein has no effect on fitness, nonsynonymous substitutions are assumed to occur at the same rate as synonymous substitutions tend to present an $\omega = 1$ and therefore suggestive of neutral evolution. Only when amino acid changes are advantageous the rate of nonsynonymous changes fixed may occur at a higher rate than synonymous changes do. Thus provides also an evidence of adaptive molecular evolution given the higher rate in nonsynonymous substitution rate than synonymous substitutions, i.e. $\omega > 1$.

A fundamental principle in genetics of adaptation is to associate environment and the effect mutations, since genes involved in adaptation have an input of beneficial mutations

that are associated with environment (Martin and Lenormand 2006; Chevin and Beckerman 2012). Therefore, if there is an optimum phenotype set by the environment, then will increase the proportion of beneficial mutations and the rate of adaptive evolution will increase, suggestive that the amount of positive versus negative selection in any gene depends on the genetic and environmental context (Martin and Lenormand 2006; Chevin and Beckerman 2012).

A simplified illustration can be obtained from the classical example of the Peppered moth (*Biston betularia*) (Cook 2000). Initially the majority had light (of peppered) colored wing patterns, but gradually after industrial revolution, the “melanic" or black forms became much more common. Given rest with their wings open on tree bark, this swap to darker forms allow them to be more camouflaged, closer to the prevailing background color in polluted areas of Britain. This provides a sort of camouflage, because bird predators would be able to find easily the light colored forms moths in polluted areas. This example illustrates how the environmental context may changes the selective pressure (here simplified to birds predation), a random mutation previously unfavourable changed to favorable. Those non-synonymous mutations in the genes responsible for the darker melanism in the peppered moth change swiftly from deleterious to beneficial and those carrying that allele become prevalent in the population. If this example is extended to different species habiting different dark color in different location, then we may conclude that those nonsynonymous mutations altering the color in the peppered moth will largely suppress the synonymous, and is therefore expected departure from neutrality, leading to the so called selection signatures. On the other hand, and inverting the process, gathering sequences from different moths inhabiting different locations, turn possible to either associated the genes responsible for the melanism and the sites that might be functionally associated with that color change (Figure 1-2).

The example is although an oversimplification of the adaptive process; a protein can be coded by more than one gene, a trait (or phenotype) is normally complex involving multiple genes and a gene can be associated with more than one biological/molecular function.

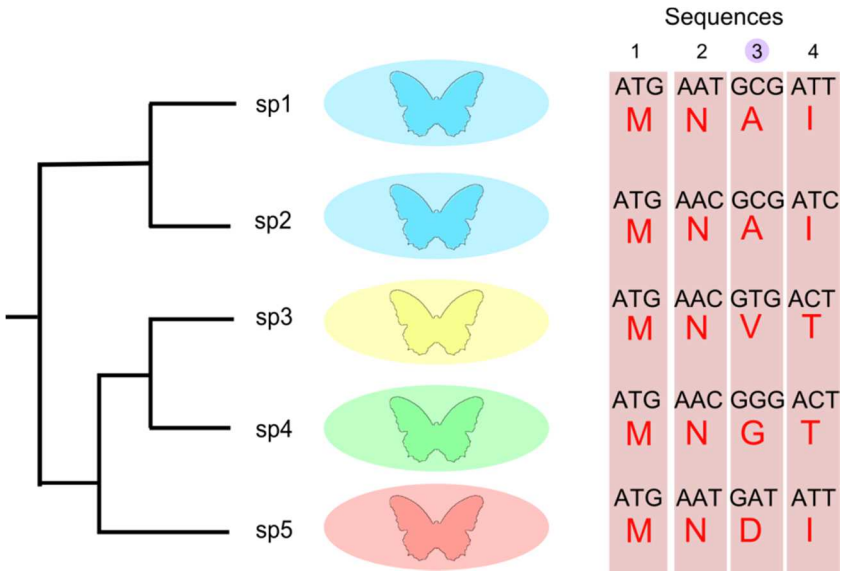

Figure 1-2. Hypothetical gene (coding sequence and the corresponding amino acid) with signatures of positive selection and involved in coloring development. Those species inhabiting a hypothetical a similar color to the habit have a higher fitness, by reducing their predation by birds. The columns 1 and 2 represent the sequences amino acid under negative selection. The column 4 represent a nearly neutral site, while the column 3 represent a positively selected site that might be implicated in color development.

1.4.2.1.1 Homologous sequences retrieval

Several evolutionary analyses often follow a common path, starting with the finding of a set of homologous (orthologs) sequences. Homology is defined as similar sequences that arose after speciation and therefore inherited from a common ancestor. This is a critical point, since the accurate identification and analysis of homologies are relevant for the statistical support and validity in phylogenetic systematics studies but to conduct positive selection studies. In a simplified approach, homologous genes are generally subdivided into orthologues and paralogues. While orthologous genes are homologous in two or more organisms that are the result of speciation, paralogous genes arose after duplication events. One often misconception about orthology is the isofunctionality (Nehrt, Clark et al. 2011; Gabaldón and Koonin 2013), since orthology does not guarantee the isofunctionality among the orhologs (further discussed in chapter 2). The target group is also of prime relevance, mirror a specific taxonomic group (for example only mammals, fishes, or both) or expand it for all homologous sequences available. Variations at this point may change the methodology that can be applied, for instance this was considered in chapter 3, we used branch-site for the entire dataset (mammals and fishes), however in site models were applied only for the individual datasets avoiding sequences saturation (further discussed).

1.4.2.1.2 Detecting Natural Selection at Molecular Level

The choice of an alignment tool/algorithm should be careful, since is critical for the next analyses and therefore should be take into account the aim of the work (Markova-Raina and Petrov 2011). Each alignment column is expected to represent a statement of homology, descent from a common ancestry. Meticulous visual inspection is frequently required (applied in chapter 2 and 3), but less feasible in larger dataset (chapter 4, 5 and 6). Several alignment tools are available, from progressive methods like ClustalW (Thompson, Gibson et al. 2002) or T-Coffee (Notredame, Higgins et al. 2000), to iterative methods such as MUSCLE (Edgar 2004), or phylogeny-aware methods such as PRANK (Loytynoja 2014). A reliable alternative to the previous, and more often used methods, is BAli-Phy, since it uses a Markov chain Monte Carlo to calculate both the phylogenetic tree and the alignment. This model takes alignment uncertainty into account to estimate the phylogeny integrating all possible alignments (Suchard and Redelings 2006). However, BAli-Phy and PRANK are computationally demanding, and runs may take several hours or days. A good comprise between power and time consuming could be obtained with MUSCLE and filtered with GBLOCKS (Privman, Penn et al. 2012). For this thesis ClustalW was only used in the first chapter, and the remaining chapters was used MUSCLE with or without GBLOCKS.

1.4.2.1.3 Substitution Models and Phylogenetic Trees

Several different Markov models of DNA sequence evolution have been proposed (Posada and Crandall 1998). Those substitution models have different parameters to describe the rates at nucleotides replacement during evolution. This is particularly relevant since evolutionary processes vary between genes and may also be from coding and noncoding regions. These models mostly differ in the parameterization of the rate matrix and the variation rate. They are evaluated through a likelihood framework, searching for the best-fit model that is either reasonable explanatory of the mutations occurring in the alignment and using the lower amount of parameters. To find the best-fitting substitution model in nucleotides, we employed the frequently applied tool Modeltest (Posada and Crandall 1998). For amino acids we used Modelgenerator (Keane, Creevey et al. 2006).

After the best-fit model to the dataset determination then a phylogenetic tree can be build. Depending on the (n) number of sequences the possible bifurcating phylogenetic trees are:

$$(2n-3)!!\ \text{(rooted trees)};$$
$$(2n-5)!!\ \text{(unrooted trees)};$$

This means for n=30 there are respectively ~4.95×10^{38} and ~8.7×10^{36} possible rooted and unrooted trees respectively, for this reason the phylogenetic trees calculations to find optimal tree is NP-hard (Non-deterministic Polynomial-time hard), this is partially solved using heuristic search and optimization methods (e.g. SPR, NNI or TBR) in combination with tree-scoring functions (likelihood) to identify a reasonably good tree that fits the data. Several algorithms implementations to generate a gene-based are available, here we used frequently the distance-matrix based Neighbor-Joining and the character based approaches such as Maximum-Likelihood or under a Bayesian Framework. The branch support give the confidence in the phylogenetic gene-based tree reconstruction, which can be attain trough pseudo-replicas such as bootstrapping or posterior probabilities methods such as aLRT or Bayesian. The comparison between the phylogenetic tree to find within the previous methods the “best-tree” can be done, e.g. in Tree-Puzzle (Schmidt, Strimmer et al. 2002).

1.4.2.1.4 Selection Analyzes

In this thesis we employed primarily models to inspect lineages and sites (between species) evolving under positive selection from a phylogenetic perspective, thus each chapter will contain explanation about the models. The codon models can be used to calculate the d_N/d_S accurately (for more extensive details on codon models parameters and assumptions (Yap, Lindsay et al. 2010)). Tests of adaptive evolution use the neutral theory, since under neutral assumption the models will not tolerate a site-class $\omega > 1$. Generically the likelihood (*lnL*) of the null models is compared against the likelihood of a model that allows positive selection (a site-class with $\omega > 1$). The significance for the referred likelihood ratio tests (LRTs) is calculated using the chi-square approximation $2\Delta lnL$ and compare against a chi-square table.

The simplest form of d_N/d_S estimation is obtained in counting methods, measuring the proportion of nonsynonymous substitutions (n) and synonymous substitutions (s) weighted trough the pathway in each case. However these methods are known to underestimate of

the distance between the two sequences (i.e. expected number of substitutions) since multiple substitutions may occur at the same site and observed in the sequences. A latter approach, using an approximate method (*YN00*) incorporates the transition/transversion bias and unequal base frequencies in their algorithm assuming the substitution model HKY85 nucleotide. This method was shown to produce good estimates of d_N and d_S, closer to true values, even for datasets with strong transition/transversions and codon bias. We used this model on chapter 2 and 6 for pairwise estimations of the ratio of d_N/d_S.

Currently the more widely used models are complex methods using maximum likelihood modeling (PAML) since they use a phylogenetic framework for the estimative process, furthermore they provide a statistical post-hoc analysis to accommodate sampling error such as Bayes Empirical Bayes (BEB). The codon substitution models make possible to test either branch or sites under selection, specifying the probability that codon (i) changes to codon (j) during evolution along the segment of the tree of length t (in time units). A first-order Markov model, assumes that the state at time (t_n+1) depends only on the previous state (t_n) (Δt is often referred as a branch length), where 61 states in the Markov model correspond to the 61 codons in the Universal genetic code (excluding stop codons). For each codon there is a maximum of 9 neighbors, to which the codon (i) may change to codon (j). The substitution models are specified as follow:

$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

Figure 1-3. Basic model of codon substitutions [1]

The substitution rate is proportional to the equilibrium frequency (π_j) of the codon (i) to codon (j), calculated using the observed frequencies in the alignment and incorporating codon usage information in the analyses (Figure 1-3). Further complementary models were developed such as SLR (Massingham and Goldman 2005), that does not accept any distribution of ω across sequence sites, and assume that dS fluctuated along the sequences. Posterior models include site-wise methods developed using a similar strategy as the "M-

[1] http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf

models" of PAML such as SLAC (Kosakovsky Pond and Frost 2005). Later, and more computationally demanding, models such as REL the sites in a coding sequences are partitioned allowing the variation in nonsynonymous and synonymous rates along the protein (similarly to SLR) or FEL method that directly estimates nonsynonymous and synonymous substitution rates at each site(Kosakovsky Pond and Frost 2005).

Since positive selection may occur along the protein but also on specific branches (time-dependent). Additionally to "site-models", branches models can be applied to the phylogenetic tree, and could test the branches under positive selection. Under more complex methods (branch-site methods) can similarly be applied to the sequence alignment but testing site-specific positive selection in a branch (or branches) along the protein. For both "branch models" and "branch-site models" positive selection detection requires labeling specific clade, a priori biological hypothesis of selection is required for the specific branch. Branch-site models (Zhang, Nielsen et al. 2005) remove the problems of branch models where all ω values on sites are averaged along the alignment). Or site-models where ω values on all branches are averaged, and therefore a more reliable approach to test episodic events of "positive selection affection", concerning only a few sites in specific branch(es) (this approach was used in the chapter 3 to see the sites positively selected in WAP65, after the teleost-specific genome duplication).

Branch models can be ran without preliminary assumption of the lineages evolving at a dN/dS ratio greater than 1, doing two tests: one with constraining all lineages to evolve at neutrality (one-ratio) and another where ω values are allowed to vary in all (free-ratio). Free-ratio branch approach constitutes a good primer method to generate "a biological hypothesis", but since the use of free-ratio models are strongly discouraged, once free-ratio models are too parameters rich. Therefore the former model is a good approach to identify and to produce final results. Since is recommended to label specific braches, under less parameters rich comparison (such as one-ratio vs two-ratio). This approach allows the identification branches evolving under differential evolutionary rate in the phylogenetic tree. Later approaches such as Branch-Site REL (Kosakovsky Pond, Murrell et al. 2011) or TestNH (Dutheil, Galtier et al. 2012) use different approaches to partially solve the problem, on the need to know the branch to label, although these approaches are computationally exhaustive with time consuming runs.

Codon models are known to be sensible and ineffective as d_S reaches saturation, which is often measured through the transitions vs transversions plots. The principle to determine if a dataset present saturation is to observe a linear proportion between substitutions and the genetic distance, where the presence of a "plateau" suggest saturation. An alternative methodology is the possibility to check for the d_S value, defining an cut-off to exclude those presenting high level of d_S, that are probably saturated (Huerta-Cepas and

Gabaldon 2011). Additionally, the incomplete lineage sorting in the gene tree can also be problematic, and state-like tree can affect selection analyses (this was taken in consideration in the chapters 2, 3 and 4).

A radical alternative approach to calculate positive selection is attaining thought positive selection at the amino acid level. Two different approaches/tools are widely used (Woolley, Johnson et al. 2003; Dutheil 2008). Contrary to codon models, these methods incorporate information about the amino acids properties, and since there are 20 amino acids but only 4 nucleotides these models are less prone to be more resilient to data saturation. By constructing discrete probability distributions, in magnitudes of biochemical change for alternative physicochemical amino acid properties and compared these expectations with the observed biochemical changes where deviations from constrained randomness indicating positive or negative selection (Woolley, Johnson et al. 2003), or by applying a two rates methods to observe unexpectedly high conservation along the protein (Dutheil 2008). An later approach was obtained "mixing" both codon models and information about the amino acid properties such as PRIME implemented in recent version of HYPHY (Pond, Frost et al. 2005).

1.4.2.1.5 Evolutionary Novelty's: Gene Duplication

1.4.2.1.5.1 Gene Duplications

Mutations are a major source of variation, although despite the value of subtle genetic modifications on preexisting ancestral genes providing differences between the species (at DNA or amino acid level), endows great value to understand adaptive evolution, they cannot not explain all their diversity. Gene duplication is known to be a major driving force of evolvability, and therefore subject of a great value to understand adaptation. Furthermore evolution through gene duplication is a mechanism known to conduct the appearance of novel features (phenotype traits or genotypes) on living organisms. Early works from Haldane (Haldane 1933) and Muller (Muller 1935) brought the hypothesis that new gene functions may emerge from old genes, highlighting gene duplication process on the arose of new genes. It is now known that gene duplication significantly contributed to functional genomes evolution and phenotypic changes, with great value in adaptive evolution. Wherefore the "birth-death" of genes has attracted much attention from biologists as well the mechanism responsible for their retention.

1.4.2.1.5.2Functional Divergence

Gene duplication is thought to be an important evolutionary mechanism by which leading to evolutionary novelty. After duplication, there are two copies in the organism which share a similar function, generally redundant copies tend to be lost during evolution or became functional distinct. Although there are at least two exceptions: one if the increased "gene-dose" is beneficial, or if both copies mildly accumulate mutations that reduce their activity. Evolutionary mechanisms such as positive selection or an increased evolutionary rate (and relaxed purifying selection) could modify one copy, leading to the functional distinction of both copies (neofunctionalization) or if both of the copies share functions and typically but expressed in different tissues (subfunctionalization). Functional shift often is categorized in Type-I or II, depending on the behavior of the evolutionary rate.

1.5 Thesis Outline

This thesis was developed with the aim to evaluate the involvement and the presence of signatures of selection, remarking positive selection as an evidence of natural selection shaping the vertebrates genes/genomes. For these purpose comparative evolutionary analyses were conducted using genes/genomes available in GenBank, Ensembl from diverse vertebrates and for chapter 3 extended with 45 avian genomes from BGI. The analyses were generically conducted in the more widely represented vertebrates in the public's repositories (Mammals, Fishes and Birds), encompassing the majority of the taxonomic groups within vertebrates. Our scope was to conduct evolutionary analyses associating the adaptive value for several genes in respect to evolutionary novelties (e.g. duplications) and occupation of different ecological niches (e.g. flight ability).

Bones are part of an intricate and complex organ involved in wide range of functions from locomotion to ion homoeostasis. In vertebrates despite the skeleton components show a remarkably similarity in basic plan, they are associated with specific adaptations for particular habits or environments within each class. Bones are structurally highly adapted to be strong yet light for movement (e.g. fly, run, swim), and thus adapted to a diversity of life habits within vertebrates. The organic part that constitutes bones encompasses extracellular matrix (ECM) primarily composed by collagen type I and several non-collagenous proteins (NCPs). Among the NCPs there is a group of proteins known as short integrin-binding ligand-interacting glycoproteins (SIBLINGs) that encompass five proteins. Preliminary analysis revealed that among the SIBLING family MEPE is the protein with the higher evolutionary rate and ergo the "best-suited" candidate to study the role in the adaptive evolution. On **chapter 2,** was aim the study of MEPE adaptive evolution in 26 Eutherian mammals and three birds. We highlighted that, besides the previous known functional domains (e.g. dentonin and ASARM) several other residues revealed to be of prime relevance. Technically we showed that applying several codon models and amino acids models, aided to reveal "functional motifs" that in opposition to the conventional methods (detects highly conserved regions) is possible to detect highly variable regions, evolving statistically faster than neutrality will predict. In this chapter is showed that rodentia and scandentia have a distinct substitution rates when compared with the other mammals, raising the question about the isofunctionallity in orthologs. The gene MEPE shows a high number of selection signatures (either nucleotide or amino acid level), revealing a crucial role of positive selection in the evolution of this SIBLING member.

The flight ability has presumably triggered differences between flying and non-flying vertebrates. On chapter 3, 89 ossification genes were studied in detail in birds and mammals, since the former are typically described as flyers and mammals are described as terrestrial. Although there are few exceptions, as several bird lineages lost the flight ability independently and within mammals, bats develop the flight as main way of locomotion. Since flight ability is correlated with bone structure, the study of close related groups, flying and non-flying, turned possible to assess the impact of flight in bones genes. Based on this, was conducted an evolutionary study in 39 mammals and 47 birds to depict the impact of flight in bone associated genes. Flight requires lightweight, compact, fused and denser bones. We detected higher evidences of selection on birds relatively to mammals that were associated with adaptive selection in ossification genes, suggesting an adaptation mechanism to improve the efficiency of flight in birds. Prevalence of positive selection on bone remodeling genes, which is similarly observed on bats, suggests that flight ability impose signatures of adaptive selection in bone-associated genes.

On chapter 4, the evolution in mammal's dentition was studied, since is a major component of the vertebrate feeding apparatus and therefore playing a crucial role in species adaptation. We focused mainly on the mammalian dentition since in this taxonomic group teeth are similar in basic components, yet exhibit great diversity in number, size and shape, through the diversity of diets in mammals from omnivorous, carnivores or herbivores. In this purpose, we surveyed 236 genes in 39 mammalian genomes, trying to depict signatures of natural selection on those genes. The analyses revealed an age relation with the evolutionary rate, since the more recent genes, having fewer interactions, have more evidence of positive selection, thus probably being those involved in the diversification of the mammalian dentition. Additionally we find evidences of a correlation between the evolutionary rate of introns and exons, since genes under positive selection have more evidences of substitutions departing from neutrality in introns.

Duplications are major contributors for evolutionary novelty, in teleost a specific genome duplication (TGD) contributed largely to their success and complexity. The protein WAP65 presented is a multifunctional protein but is mainly related with iron homeostasis. This gene was duplicated in teleost, were two paralogs diversified in terms of function. On chapter 5 we have evaluated the molecular evolution of HPX and WAP65 in 66 vertebrates. We have conducted evolutionary analyses to understand the evolution and functional distinctiveness of the WAP65 paralogs and the mammalian orthologue HPX, characterizing in detail signatures of positive selection acting on these protein-coding genes. We have depicted the selection signatures that may have been responsible for the functional divergence

between WAP65 in fishes and HPX in mammals, particularly by testing the branch immediately after duplication.

Pseudogenes were primarily described as “junk” DNA, dead copies of an ancestral fully active gene that are present in the genome of different species. Two main processes may lead to the form pseudogenes from functional copies, retro-transposition (processed) or inactivation of an early functional gene mainly from duplicated copies but also from non-duplicated copies (non-processed). On chapter 6, processed pseudogenes were studied in 18 species mirroring their potential adaptive value. The observation of non-random process on non-coding regions enlighten the potential role of noncoding regions on the adaptive value on those species, particularly 5 species that were surveyed in detail, looking for the retention pattern when the processed pseudogenes are allocated to a different chromosome

Chapter 2 *- Adaptive evolution of the Matrix Extracellular Phosphoglycoprotein in mammals*

2.1 Abstract

Matrix extracellular phosphoglycoprotein (*MEPE*) belongs to a family of small integrin-binding ligand N-linked glycoproteins (SIBLINGs) that play a key role in skeleton development, particularly in mineralization, phosphate regulation and osteogenesis. MEPE associated disorders causes various physiological effects, such as loss of bone mass, tumors and disruptmepion of renal function (hypophosphatemia). The study of this developmental gene from an evolutionary perspective could provide valuable insights on the adaptive diversification of morphological phenotypes in vertebrates.

Here we studied the adaptive evolution of the *MEPE* gene in 26 Eutherian mammals and three birds. The comparative genomic analyses revealed a high degree of evolutionary conservation of some coding and non-coding regions of the *MEPE* gene across mammals indicating a possible regulatory or functional role likely related with mineralization and/or phosphate regulation. However, the majority of the coding region had a fast evolutionary rate, particularly within the largest exon (1467 bp). Rodentia and Scandentia had distinct substitution rates with an increased accumulation of both synonymous and non-synonymous mutations compared with other mammalian lineages. Characteristics of the gene (e.g. biochemical, evolutionary rate, and intronic conservation) differed greatly among lineages of the eight mammalian orders. We identified 20 sites with significant positive selection signatures (codon and protein level) outside the main regulatory motifs (dentonin and ASARM) suggestive of an adaptive role. Conversely, we find three sites under selection in the signal peptide and one in the ASARM motif that were supported by at least one selection model. The MEPE protein tends to accumulate amino acids promoting disorder and potential phosphorylation targets.

MEPE shows a high number of selection signatures, revealing the crucial role of positive selection in the evolution of this SIBLING member. The selection signatures were found mainly outside the functional motifs, reinforcing the idea that other regions outside the dentonin and the ASARM might be crucial for the function of the protein and future studies should be undertaken to understand its importance.

2.2 Introduction

Dentin, one of the major mineralized tissues of teeth, is deposited by odontoblasts, which synthesize collagenous and non-collagenous proteins (NCPs) (Butler 1998; Thesleff 2003). Among the NCPs, there is a family of small integrin-binding ligand N-linked glycoproteins (SIBLINGs) consisting of dentin matrix protein 1 (*DMP1*), dentin sialophosphoprotein (*DSPP*), integrin-binding sialoprotein (*IBSP*), matrix extracellular phosphoglycoprotein (*MEPE*, also known as *OF45*) and osteopontin (*SPP1*) (Chen, Chen et al. 2008). These genes share common genetic and structural features, including a small non-translational first exon, a start codon in the second exon and a large coding segment in the last exon (although exon number varies among the different genes) (Fisher and Fedarko 2003). The entire SIBLING protein family likely arose from the secretory calcium-binding phosphoprotein (SCPP) family by gene duplication, since this cluster of genes encodes proteins with similar molecular–structural features and functions (Kawasaki and Weiss 2006).

Members of this gene family are encoded by a compact tandem gene cluster (located on chromosome 4q in Human and 5q in mouse) characterized by: (*i)* common exon–intron features, (*ii*) the presence of the integrin-binding tripeptide Arg-Gly-Asp (RGD) motif that mediates cell attachment/signaling via interaction with cell surface integrin (Fisher and Fedarko 2003), and (*iii*) post-translational modifications of conserved phosphorylation and N-glycosylation sites (Fisher and Fedarko 2003). In humans, the MEPE protein (525 amino acids) is encoded by four exons with a 1960 bp transcript with two N-glycosylation motifs (at residues 477–481), a glycosaminoglycan (SGDG) attachment site at residues 256–259, and the RGD cell attachment motif at residues 247–249 (Yamada 1991). The RGD motif has a similar function in other members of the SIBLING´s (DSPP, DMP1, IBSP, and SPP1) (Rowe, de Zoysa et al. 2000). The protein MEPE has several predicted phosphorylation sites/motifs for protein kinase C, casein kinase II, tyrosine kinase, and cAMP–cGMP-dependent protein kinase and a large number of N-myristoylation sites that appear to be also a feature of the RGD-containing proteins (Rowe, de Zoysa et al. 2000). The acidic serine-aspartate-rich MEPE-associated motif (ASARM motif) occurs at the C-terminus in MEPE (residues 509 to 522) (Rowe, de Zoysa et al. 2000) and when phosphorylated this small peptide can bind to hydroxyapatite and inhibit mineralization (Addison, Nakano et al. 2008).

The basic MEPE protein was first cloned from a tumor resected from a patient with tumor-induced osteomalacia (OHO) (Rowe, de Zoysa et al. 2000; Ogbureke and Fisher 2007), which is associated with hypophosphatemia and is caused by a renal phosphate wasting. The *MEPE* gene is also up-regulated in X-linked hypophosphatemia rickets (XLH or HYP-osteoblasts) and OHO-tumors (Petersen, Tkalcevic et al. 2000; Rowe, de Zoysa et al. 2000; Argiro, Desbarats et al. 2001; De Beur, Finnegan et al. 2002; Gowen, Petersen et

al. 2003; Rowe 2004). Under normal conditions it is expressed primarily in osteoblasts, osteocytes, and odontoblasts (Rowe 2004).

Targeted disruption of the *MEPE* gene in mouse causes increased bone formation and bone mass, suggesting that MEPE plays an inhibitory role in bone formation and mineralization (Gluhak-Heinrich, Kotha et al. 2004). In humans, MEPE inhibits mineralization and is also involved in renal phosphate regulation (Dobbie, Shirley et al. 2003; Rowe, Kumagai et al. 2004). The inhibition of mineralization and phosphate uptake are related with the protease resistant small peptide ASARM motif located near the end of the protein (Rowe, de Zoysa et al. 2000; Rowe, Kumagai et al. 2004). However, the MEPE protein has dual functions depending on the proteolytic processing. When the protein is cleaved by cathepsin B or D into several fragments, the small peptide ASARM is released (Rowe, Matsumoto et al. 2005) and when phosphorylated, this small peptide can bind the hydroxyapatite crystal and inhibit mineralization (Addison, Nakano et al. 2008). By contrast, when fragments containing the RGD motif are released and the ASARM is not degraded by proteases, mineralization is accelerated (Hayashibara, Hiraga et al. 2004). The influence of MEPE-ASARM peptides in the modulation of mineralization is due to a protein-protein interaction with PHEX, an X-linked phosphate-regulating endopeptidase homolog (also called the minhibin model) (Rowe, Kumagai et al. 2004). PHEX is also expressed in osteoblasts, osteocytes and odontoblasts and the protein interacts with MEPE, protecting it from the proteolytic process (from cathepsin-B) and preventing ASARM from being released into blood circulation (Addison, Nakano et al. 2008). Most of the disorders associated with MEPE result from a malfunction of this PHEX-MEPE interaction, which in turn leads to an increase of ASARM blood levels.

The majority of mammalian genes are strongly conserved in the coding sequence (Lander, Linton et al. 2001; Waterston, Lindblad-Toh et al. 2002). Genes carrying signatures of selection may be involved in adaptation and functional innovation, and often have elevated ratios of nonsynonymous/synonymous nucleotide substitutions (d_N/d_S) in their coding regions (Ohta 1992). However, evolutionary rates of nuclear and mitochondrial genes are not equal in all the mammalian lineages (Hasegawa, Thorne et al. 2003). For example, while rodents tend to accumulate more mutations in nuclear genes than humans (Wu and Li 1985), the differences between the rates in the two lineages seems to be smaller than the generation time difference (Hasegawa, Thorne et al. 2003).

Since MEPE protein has an important role in the regulation of the skeleton mineralization process and since the mineralized tissue is a critical innovation in vertebrate evolution, the evolutionary study of this developmental gene could provide valuable insights on the adaptive diversification of morphological phenotypes in mammals. As the *MEPE* gene has been suggested to be under selection (Bardet, Delgado et al. 2010), our objective was

to undergo a thorough analysis to evaluate signatures of positive selection using both a gene-level and protein-level approaches. We assessed the evolution of the MEPE protein-coding gene in 26 mammalian species, from Hyracoidea to Primates, showing that while four regions/motifs in the *MEPE* gene have a high degree of conservation, the majority of the coding region has a fast evolutionary rate, especially in rodents and tree shrews. Indeed, evidence of strong positive selection (gene and protein-level) was found in 20 amino acids that encompass MEPE protein, highlighting the role of molecular adaptation in the functionality of this gene.

2.3 Methods

Comparative genomic analyses

MEPE nucleotide sequences were retrieved from GenBank and ENSEMBL. We aligned 26 *MEPE* sequences representing eight orders of mammalian species and produced two different alignments, one including all species and another excluding rodents and the tree shrew due to its nucleotide saturation bias. Given the low similarity between the avian and the mammalian sequences, the avian sequences were excluded from phylogenetic and selection analyses. BLAST searches were used to retrieve non-annotated sequences from several mammalian genomes. All the alignments were performed after the translation of nucleotides to amino acids and the corresponding alignments were back-translated to nucleotides. The alignment were performed in ClustalW (Thompson, Higgins et al. 1994) implemented in BIOEDIT v7.05 (Hall 1999), MEGA4 (Tamura, Dudley et al. 2007) and LAGAN (Brudno, Do et al. 2003). Sliding-window percent amino acid and nucleotide identity, and % GC content were calculated in Swaap 1.0.3 (Pride and Blaser 2002). Saturation plots (including or excluding the third-coding position) and the estimated pI (excluding indels) were assessed in DAMBE (Xia and Xie 2001). Conservation in the coding and the non-coding regions was assessed using mVISTA (Frazer, Pachter et al. 2004).

Phylogenetic analyses

We used Modelgenerator version 0.85 (Keane, Creevey et al. 2006) to determine the optimal model of sequence substitution for our protein dataset, employing the Jones-Taylor-Thornton (JTT+I+G) substitution model. MrModeltest 2.3 (Nylander 2004) was employed to determine the optimal model of sequence substitution for our coding sequence dataset, employing the General-Time-Reversible (GTR+I+G) substitution model with the invariant site plus gamma options (five categories). Bayesian inference methods with Markov

chain Monte Carlo (MCMC) sampling were performed in MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). The analysis was run for 5,000,000 generations with a sample frequency of 100 and burn-in was set to correspond to 25% of the sampled trees. The Maximum-Likelihood (ML) phylogenetic tree was constructed in PHYML (Guindon, Delsuc et al. 2009), under the best-fit model for nucleotides and amino acids, 1000 bootstrap replicates and the NNI branch search algorithm. The parameters used in the tree reconstructions were set to: (i) Nucleotides: GTR+I+G with 6 substitution rate parameters and gamma-distributed rate variation with a proportion of invariant sites; (ii) Amino acids: JTT+I+G. A neighbor-joining tree was conducted in MEGA 4 (Tamura, Dudley et al. 2007) using the complete deletion of ambiguous data and the maximum composite likelihood option. The topologies were tested in TREE-PUZZLE (Schmidt, Strimmer et al. 2002) to identify the tree that best fits the alignment, using three tests: KH, SH and ELW. A phylogenetic signal test was performed in TREE-PUZZLE (Schmidt, Strimmer et al. 2002) using the implemented methodology (Strimmer and von Haeseler 1997).

Detection of positive selection

Gene-level analyses

Positive selection analyses were performed in the Eutherian mammals (the closely-related taxa) to avoid nucleotide saturation and base-compositional bias. We assessed positive selection using primarily a gene-level approach (Antunes and Ramos 2007) based on the ratio (ω) of nonsynonymous (d_N) to synonymous (d_S) substitutions rate (i.e., $\omega = d_N/d_S$), implemented in PAML v4.3 (Yang 2007) and in the web-based program SELECTON (Doron-Faigenboim, Stern et al. 2005; Stern, Doron-Faigenboim et al. 2007), PAML uses LRT to compare two nested models, a model that does not allow, and a model that allows, sites categories > 1 (null versus positive selection, respectively). Here, we used three LRTs based on site-specific models comparing the nested models: M1a-M2a, M7-M8 and M8a-M8. The first LRT was performed comparing M1a (nearly neutral: p0, p1, ω0 < 1, ω1 = 1, N_S sites = 1) against M2a (positive selection: p0, p1, p2, ω0 < 1, ω1 = 1, ω2 < 1, $N_S$sites = 2); the second LRT was comparing M7 (beta: p, q, NS sites = 7) with M8 (beta & ω: p0, p1, p, q, ω_s > 1, N_S sites = 8). The third LRT was between M8a (beta & ωs = 1: fix omega = 1, omega = 1, N_S sites = 8) and M8. However, a significant LRT only demonstrated that the selection model is more suitable than the neutral model; it does not provide any indication of the sites under selection (Osorio, Antunes et al. 2007). This can be accomplished through an Empirical Bayes (EB) approach to calculate the posterior probability (PP) that a given site comes from the class with ω>1. Sites presenting a PP above the defined cut-off value

(e.g. p>95%) (Yang, Wong et al. 2005) are inferred to be under positive selection. A robust method was used to accommodate the uncertainties in the MLEs of parameters in the ω distribution, designated by BEB (Yang, Wong et al. 2005). This approach was shown to be reliable in both small and large data sets, and also to have a good resolution power for identifying individual sites under positive selection, especially in large data sets or with strong selective pressure. We also performed an analysis using the branch-site model A (Zhang, Nielsen et al. 2005) (model = 2 N_S sites =2), including and excluding the rodents and tree shrew as foreground branch, allowing the ω ratio vary both among sites and among lineages. The branch-site test 2 was performed using the null model, $\omega_2 = 1$ fixed (using the parameters fix omega = 1 and omega = 1). The sites under selection in the foreground branches were obtained after calculating probabilities of site classes using the BEB procedure.

Although the PAML models (Yang 2007) allow for variation in the non-synonymous substitution rate, the synonymous rate is fixed across the sequence. To overcome that specificity, we used SLAC and FEL (Pond and Muse 2005) for detecting positive selection while allowing variation in synonymous rate. SLAC is a heavily modified and improved derivative of the Suzuki–Gojobori counting approach (Kosakovsky Pond and Frost 2005; Pond and Frost 2005) that maps changes in the phylogeny to estimate selection on a site-by-site basis. SLAC calculates the number of non-synonymous and synonymous substitutions that have occurred at each site using ML reconstructions of ancestral sequences (Kosakovsky Pond and Frost 2005; Pond and Frost 2005). The FEL model estimates the ratio of nonsynonymous to synonymous not assuming *a priori* distribution of rates across sites substitution on a site-by-site analysis (Kosakovsky Pond and Frost 2005). The SLAC and FEL methods were implemented using the web interface Datamonkey (Pond and Muse 2005). Since recombination in the gene can bias the analysis (Anisimova, Nielsen et al. 2003), we also rerun SLAC and FEL in Datamonkey using the GARD method (Kosakovsky Pond, Posada et al. 2006), allowing each calculated partition to have its own phylogenetic tree.

Additionally, we used the LRT based analysis as implemented in the SLR (Sitewise Likelihood-Ratio) software package (Massingham and Goldman 2005). This method assumes that substitutions (both synonymous and non-synonymous) can occur independently with every other site, modulating substitution rates as a continuous-time Markov process. The LRT on a site-wise basis is performed testing a null model (neutrality, $\omega = 1$) against an alternative model $\omega \neq 1$.

Protein-level analyses

We performed multiple analyses to differentiate the different types of selective pressures acting in *MEPE*: (i) positive versus negative selection, and (ii) stabilizing (selection that tends to maintain the overall biochemistry of the protein) versus destabilizing selection (selection that results in radical structural or functional shifts in local regions of the protein). These analyses provided insight into the structural and functional consequences of the residues under selection (McClellan, Palfreyman et al. 2005). We used TreeSAAP v3.2 (Woolley, Johnson et al. 2003) and CONTEST (Dutheil 2008) implemented in IMPACT (Maldonado, Dutheil et al. 2011) to detect selection signatures at the amino acid level. In TreeSAAP positive destabilizing-selection is detected based on the properties changes with significantly greater amino acid replacements than would be expected under neutrality for magnitude categories +7 and +8 (i.e., the two most-radical property-change categories). Within TreeSAAP, 31 amino acid properties were evaluated across the phylogenetic tree to identify the specific amino acid residues within each region that showed evidence of positive destabilization for each property. The *Baseml* implemented PAML (Yang 2007) is used in TreeSAAP (Woolley, Johnson et al. 2003) to reconstruct ancestral character states at the nodes on the *MEPE* phylogeny.

To test if evolutionary rates varied between lineages we used the relative-rate test, weighting by the predefined tree topology, as implemented in RRTree (Robinson-Rechavi and Huchon 2000). To detect directional selection over the tree or a large number of substitutions towards a particular residue in a maximum likelihood context we used the directional evolution in protein sequences (DEPS) analysis to identify statistically significant directional changes in amino acid residue frequencies (Kosakovsky Pond, Poon et al. 2008).

MEPE Three-Dimensional Structure Modeling

To determine the position of the positive selected amino acids when the protein is folded, we modeled the three-dimensional (3D) structure of MEPE. Protein structure prediction can be approached in three ways: (i) comparative modeling, (ii) threading, and (iii) ab-initio folding. For MEPE, the first two methods, which build a protein model by aligning query sequences onto solved template structures, were not feasible. Thus, the only practical strategy was to run the I-TASSER (Zhang 2008) to obtain an ab-initio 3D model of MEPE. The model obtained using the *Homo sapiens* sequence had a TM Score of 0.46±0.15 and a C-Score=-2.18. To accurately infer the correct topology, the model should have a C-score above -1.5, varying from [-5; 2] (Zhang 2008). A TM score above 0.5 means that the obtained topology is not random [86]. Results using the sequences of the rock hyrax (out-

group), the dog (i.e. one of the species showing differences in the pI) and the mouse (which demonstrates accelerated evolution) all had similar C-scores and the 3D structures similar to the results retrieved for the human MEPE, suggesting that the biochemical differences in the composition of the amino acids that constitutes the different orthologues are not imposing significant differences in the folding of the protein.

Structural analyses

To assess the surface exposure of the amino acids in the protein structure, we used the GETAREA 1.1 (Fraczkiewicz and Braun 1998) web-based program based on the atom coordinates of the PDB file. This provides an estimate of the solvent exposure based on the ratio of the side-chain surface area to "random coil" value per residue, performing an analytical calculation of solvent accessible surface area residues. These are considered to be solvent exposed if the ratio value exceeds 50% and to be buried if the ratio is less than 20% (Fraczkiewicz and Braun 1998). Since MEPE has been described as an intrinsic unfolded protein, we also used the Protein DisOrder prediction System (PrDOS) server (Ishida and Kinoshita 2007) to predict natively disordered regions of a protein chain based on the composition of the amino acid sequence. Protein stability was calculated with the PoPMuSiC 2.1 web server (Dehouck, Kwasigroch et al. 2011) using the MEPE PDB file previously obtained in I-TASSER to calculate the sites Γ considering all the possible mutations in each site. The secondary structure was visualized in POLYVIEW (Porollo, Adamczak et al. 2004).

2.4 Results

Presence of the MEPE in vertebrates

Twenty-six mammalian *MEPE* sequences were retrieved from the GenBank and Ensembl databases, comprising eight different mammalian Orders (Appendices II: Table S1). In addition, sequences of the putative *MEPE* orthologue, Ovocleidin-116, were obtained from the available bird genome projects (*Gallus gallus, Taeniopygia guttata, Meleagris gallopavo*) for comparative purposes. For the majority of the mammals considered in this work, the *MEPE* gene encompasses four exons that encode a transcript that varied from 1272 bp in *Ochotona princeps* to 2030 bp in *Pan troglodytes*. Some of the smallest reported transcripts may be incomplete, as in the case of *O. princeps*, which is missing a stop codon. The absence of the ASARM motif in the MEPE´s C-terminal in some species (*Equus caballus*, *Ochotona. princeps, Otolemur garnetti and Pteropus. vampyrus*) also suggests that those genes were not fully annotated. Thus, we performed a detailed search in databases

for those species using TBLASTN (Altschul, Gish et al. 1990), which led to the identification of the ASARM in *E. caballus,* but not in *O. princeps*, *O. garnetti* and *P. vampyrus* (in these cases, the missing end portion of the protein corresponds to the end of the contig available in the database). However, several stop codons are present between the end of the present sequence and the putative ASARM motif in the *E. caballus* sequence and therefore it was not included in subsequent analyses.

We performed blast searches (TBLASTN and TBLAST) to determine if *MEPE* is present in non-mammalian or non-avian vertebrates (such as fish and amphibians), but we were not able to detect an orthologue in those lineages, suggesting that this gene may be considerably differentiated or even absent. In chicken (*G. gallus)*, a similar protein has been already described, MEPE/OC116 (Hillier, Miller et al. 2004) (i.e. Ovocleidin 116), and it is likely a homologue of *MEPE*. This orthologue is also present in two other birds (*T. guttata, M. gallopavo*). Although our initial BLAST searches did not return a significant hit in reptiles, a recent study suggests the presence of *MEPE* in *Anolis carolinensis* (Fisher 2011). Blast searches for the *MEPE* gene in teleost fishes (e.g. *Takifugu rubripes*, *Oryzias latipes* and *Danio rerio*) did not retrieve a significant hit. Even searching synteny blocks between Human and Zebrafish (results not shown), did not provide evidence of MEPE. This result is concordant with previous studies (Sollner, Burghammer et al. 2003; Kawasaki and Weiss 2006; Kawasaki and Weiss 2008; Kawasaki 2009; Ramialison, Bajoghli et al. 2009) that show the likely presence of two genes belonging to the SIBLING family in teleost fishes but not a *MEPE* orthologue. Mammals and reptiles are the only tetrapod lineages with all five SIBLING family genes (Figure 2-1), as previously suggested (Kawasaki 2009; Fisher 2011).

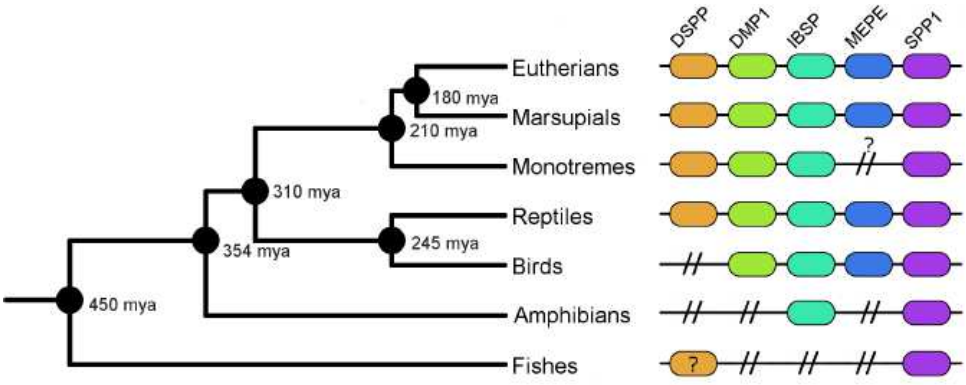

Figure 2-1. SIBLING (*DSPP*, *DMP1*, *IBSP*, *MEPE* and *SPP1*) presence in vertebrates. Illustrative representation of the SIBLING (*DSPP*, *DMP1*, *IBSP, MEPE* and *SPP1*) genes presence/absence in vertebrates. The estimated divergence time of the different groups are placed near the nodes.

Sequence analyses

At the protein level MEPE is highly variable, especially in the region encoding the last exon, with pairwise amino acid similarity among mammals varying from 99% to 28%. Nevertheless, four important regions within *MEPE* had high amino acid conservation (>80%): the signal peptide, the RGD and SGDG regions (the glycosaminoglycan attachment site), and the ASARM motif (Figure 2-2). Moreover, the protein is also highly conserved from positions 887 to 1091 bp of the human sequence, a region associated with a putative regulatory region (*Ensembl* annotation). Exon 2, only 54 bp long, encodes mainly the signal peptide and is highly conserved. Remarkably, two alanines (hydrophobic residues) are conserved in 25 of the 26 mammalian species studied (Figure 2-2). The fourth exon (that encodes most of the protein) comprehends the RGD, SGDG, and ASARM motifs and the putative regulatory region. GC content was similar along most of the coding sequences, with a few segments above 50% (Figure 2-2).

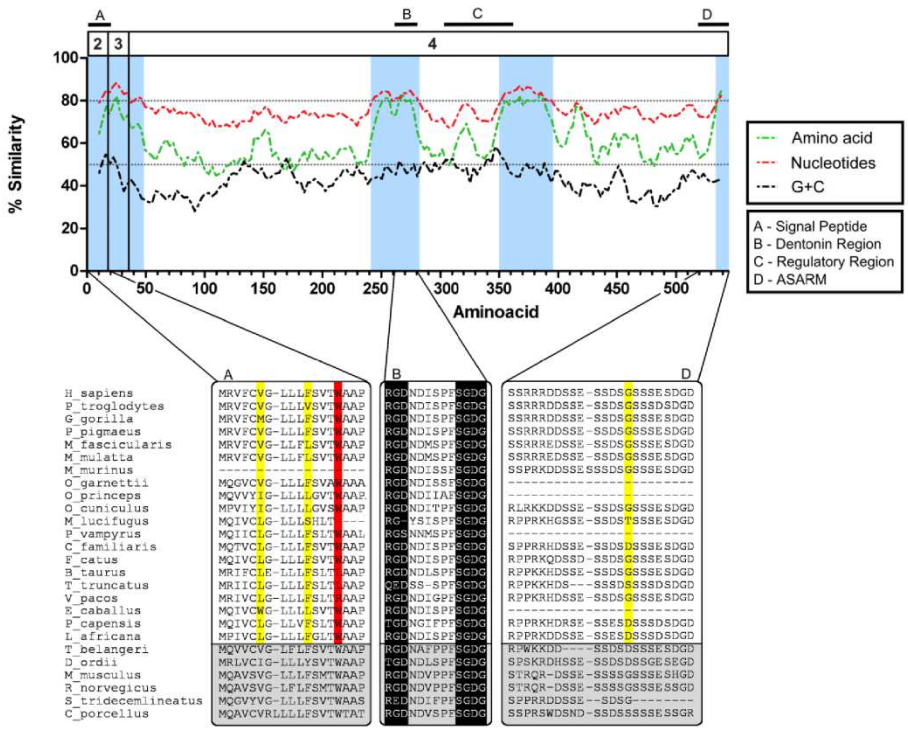

Figure 2-2. Sliding window plot and motifs comparison of MEPE across the 26 mammalian species. Sliding window plot of GC-content and nucleotide and amino acid conservation among the 26 mammalian MEPE coding sequences (exons 2, 3 and 4) that were used in this evolutionary study. The plot was calculated after pairwise deletion of ambiguous sites and the windows were adjusted to correspond to the same scale. The blue shading identifies conserved regions (>80% nucleotide or amino acid similarity), the red line tracks nucleotide similarity, the green line amino acid similarity and the black line %GC content. The three motifs/regions are represented within the boxes A-Signal Peptide-, B-Dentonin (SDGG, RGD), C- Putative regulatory regions and D-ASARM. The yellow and red shadow represents selection at codon level and amino acid level, respectively, while the grey shadow correspond to the species excluded from the positive selection analyses at site level.

Phylogenetic analyses of the mammalian MEPE protein sequences showed similar overall topologies with the three reconstruction methods used: Neighbor-Joining (NJ), Bayesian (BY), and Maximum Likelihood (ML) (Figure 2-3).

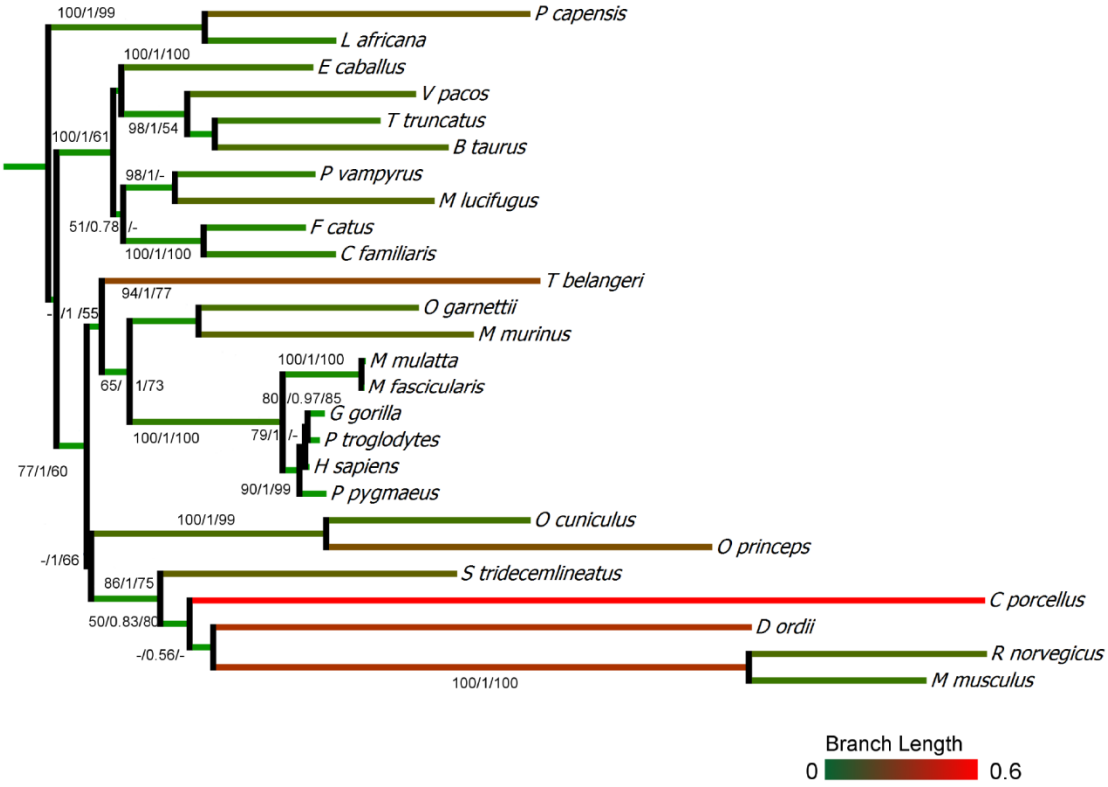

Figure 2-3. Phylogenetic tree of MEPE. Depiction of MEPE protein phylogeny constructed using Bayesian inference (Bayes), Maximum Likelihood (ML) and Neighbor-Joining (NJ) algorithms. Support for each node is summarized on the branch prior to the node (ML/Bayes/NJ). For the NJ and ML analysis the bootstrap values <50 are represented with the symbol (-). Branches are shaded with a gradient based on the branch length, from green (short) to red (longer).

The topologies were also consistent with those retrieved when using the *MEPE* nucleotide sequences (results not shown), and all were mostly compatible with the accepted phylogeny of mammals (Springer, Murphy et al. 2003; Nishihara, Hasegawa et al. 2006; Meredith, Janecka et al. 2011; Perelman, Johnson et al. 2011). However, Rodentia and Scandentia had long branches, suggesting higher mutation rates (increased number of synonymous and non-synonymous substitutions). We performed the two-sided Kishino-Hasegawa test (KH), the Shimodaira-Hasegawa test (SH), and the Expected Likelihood Weights (ELW) in TREE-PUZZLE to determine the best-fitting tree. The test of the three resulting phylogenetic trees suggests that the ML tree best fit the multiple sequence alignment (values of KH and SH were 1, and therefore were highly significant and ELW=0.7771), although the Bayesian tree was not significantly worse than the ML tree (Appendices II: Table S2). Conversely, after removing the rodents and tree shrew the three methods produced similar

topologies and therefore no significant differences were obtained in the tests implemented in TREE-PUZZLE. The best-fitting trees for the two alignments were then used in subsequent analyses. Likelihood mapping, implemented in TREE-PUZZLE to inspect the phylogenetic signal of the alignment (Appendices II: Table S2), showed a relevant value for both alignments that was slightly reduced when rodents and tree shrew were included. Phylogenies based only on transversions or only on the first and the second coding positions showed the same patterns (data not shown).

In the non-coding gene regions, the nucleotide similarity plots illustrate that the human sequence is highly conserved relative to the other primates, *Pan troglodytes*, *Gorilla gorilla* and *Macaca mulatta* (Figure 2-4A).

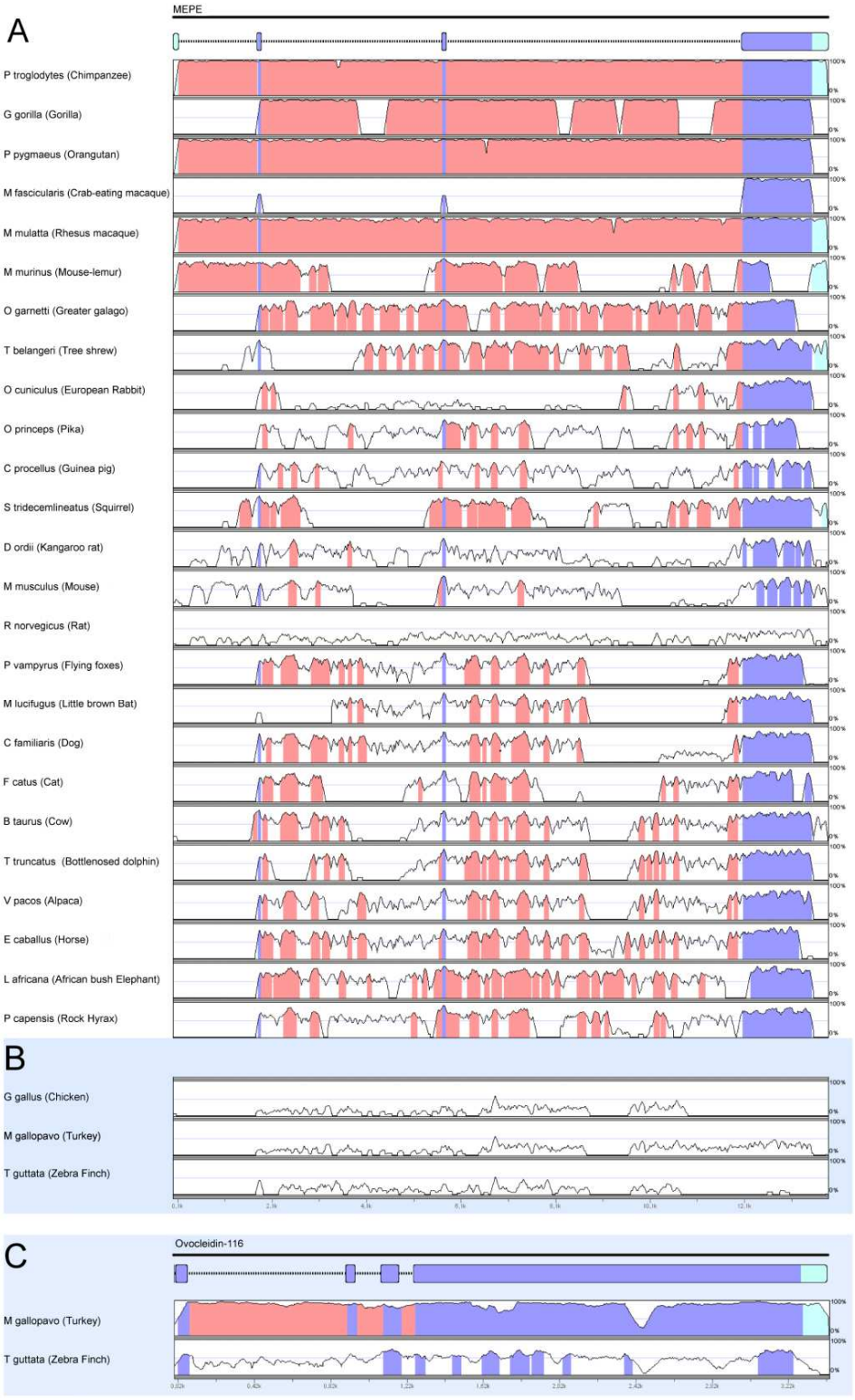

Figure 2-4. Nucleotide conservation of *MEPE* in mVISTA (A) MEPE gene conservation between 25 mammalian species orthologues compared with the human MEPE sequenced portrayed in an mVISTA plot with the 100 bp window with a cut-off of 70% similarity. The Y-scale represents the percent identity ranging from 0 to 100%. (B) Human MEPE compared with the three bird orthologues. (C) Pairwise comparison of the two birds ovocleidin-116 with the *G. gallus* orthologue. Exons are highlighted in blue, nontranslated regions in green-blue, and conserved non-coding sequences (CNS) in pink.

At a lower level the comparison of the *MEPE* non-coding regions across all species showed several Conserved Non Coding Sequences (CNS) after pairwise comparisons with the human sequence across all species. This intronic conservation is particularly important since CNS have been associated with transcriptional regulation (Majewski and Ott 2002). The length of CNS decreases when the Human *MEPE* is compared with homologues from more distantly related species, but not necessarily in a direct association with phylogenetic distance (Figure 2-4A). For instance, the dog (*Canis lupus familiaris*) and cattle (*Bos taurus*) are phylogenetically more distant from human than the mouse (*Mus musculus*) and rat (*Rattus norvegicus*), but showed a higher conservation both in coding and non-coding regions of the gene (Figure 2-4A). By contrast, in the Order Lagomorpha there is less conservation in the intronic regions but high conservation in the coding regions, and in rodents, there are high numbers of differences both in coding and non-coding regions (Figure 2-4A). As expected, birds showed low similarity in both coding and non-coding region with mammals (Figure 2-4B), although they exhibited high similarity in the coding regions in pairwise comparisons with *G. gallus* (Figure 2-4C). Furthermore, the two Galliform species also were similar in the non-coding regions while the *G. gallus* and the *T. guttata* did not present high intronic conservation.

Given the large difference in average length of CNS (from 1.8kb in Lagomorpha to 8.4kb in Primates) and their high similarity (from 71.4% in Lagomorha to 89.5% in Primates) (Appendices II: Figure S1), it is not surprising that introns have ample phylogenetic signal for gene-tree reconstruction. The alignment of the intronic regions comprehends 21120 bp and 856 of those sites were clean of ambiguity data in all the species (*Macaca mullata* was excluded since the intronic regions were not available). *MEPE* intronic sequences provided a significant phylogenetic signal across all the studied mammals, resulting in similar topologies as those trees reconstructed from coding regions and protein suggesting an appreciable level of evolutionary constraints in *MEPE* introns (Figure 2-5).

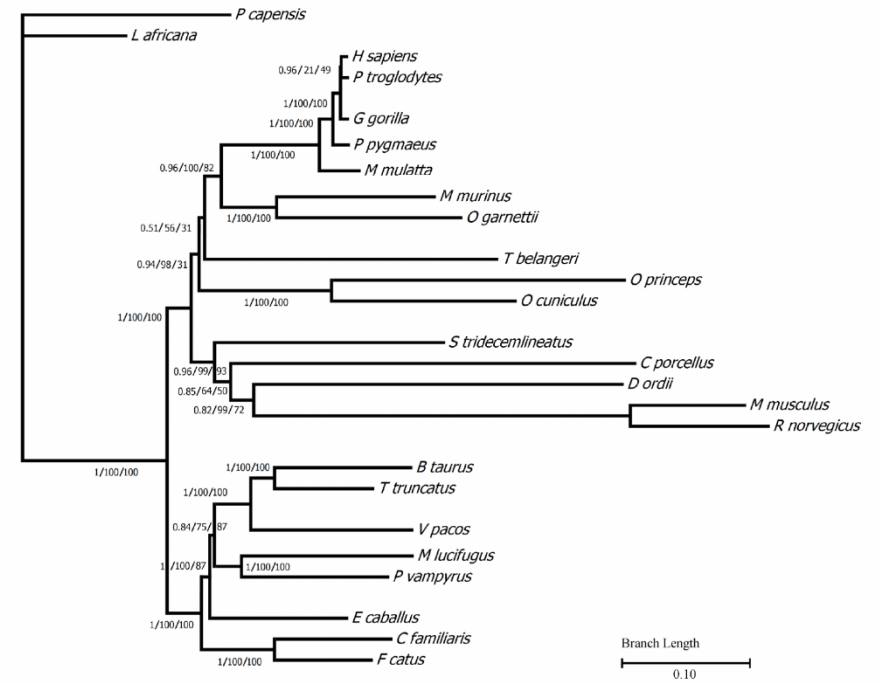

Figure 2-5. Phylogenetic tree of *MEPE* intronic regions.Phylogenetic depiction of the *MEPE* intronic region tree reconstructed using Bayesian inference (Bayes), Maximum Likelihood (ML) and Neighbor-Joining (NJ) algorithms. The labels are positioned near the branches supporting the tree and inside the brackets (Bayes/ML/NJ). The methodology was similar to the implemented in the coding regions. The alignment of the intronic regions comprehends 856 out of 21.120 sites completely clean of gaps in all the species (except for the *Macaca mullata* since the intronic regions was not available).

The MEPE protein is generally basic, with an average Isoelectric Point (pI) of 8.20 in the mammal species studied. Generally the pI was lower in Laurasiatheria, reaching 5.82 in *Felis catus* (Figure 2-6).

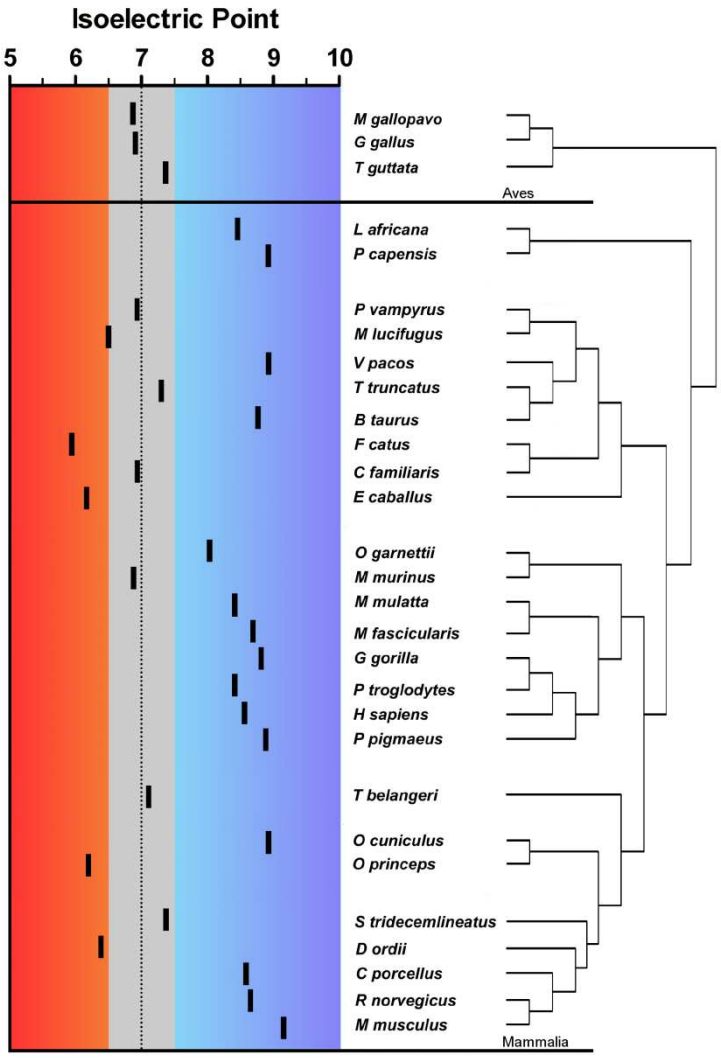

Figure 2-6. MEPE isoelectric points (pI) calculated for the 26 mammalian and 3 avian species. The red shadow represents the acid pI while the blue the basic pI, the grey shadow shows the nearly neutral proteins, from 6.5 to 7.5.

In the three available avian sequences pI was less than 7 in the two Galliformes and slightly above 7 in Passeriformes. These differences in pI may have dramatic effects on the protein folding, as those changes are caused by significant differences in the polarity of the amino acids that compose the protein.

Functional motifs

The cell attachment region, RGD, situated near the center of the MEPE protein, is fully conserved in 20 of the 26 mammalian species (Figure 2-2). However, some changes are observed in *Tursiops truncatus*, *Procavia capensis*, the bats *Pteropus vampyrus* and *Myotis lucifugus*, and in the rodents *Dipodomys ordii* and *Spermophilus tridecemlineatus*

(Figure 2-2) and it is likely that such amino acids changes in the RGD motif may have functional relevance. Moreover, the RGD motif is also present in other genes of this gene-cluster family.

The SDGD is completely conserved among all the mammals, reinforcing the premise that this peptide region is, along with RGD, important to the MEPE function. These two motifs constitute the dentonin region, which was not detected in any of the others members of the SIBLING protein family. The chicken and the turkey *MEPE* orthologues appear to be exceptions, since they do not have the cell-adhesion motif, RGD, but contain the glycosaminoglycan-binding motif, SGDG. In these species we found a HGD near the SGDG motif, suggesting that RGD is replaced by HGD (Appendices II: Figure S2). A similar change from RGD has been described in other members of the DSPP orthologues (e.g. in rat, *Rattus norvegicus*, the RGD replaces the HGD) (McKnight and Fisher 2009). Nevertheless, in zebra finch (*T. guttata*) we found the RGD motif but not the SGDG region (Appendices II: Figure S2). The ASARM motif is highly conserved within the 21 mammals for which ASARM is annotated (average above 85%), although the Bottlenose dolphin (*Tursiops truncatus*) has a similarity of only 59.1%. Pairwise similarity among birds was 79.9% (among the three avian species), but on average only 27.3% similarity was observed between birds and the mammalian ASARM. Moreover, in birds this motif is capped at the C-terminal by 21 (*G. gallus, M. gallopavo*) to 24 (*T. guttata*) amino acids, and this region shows 77.2% similarity between *G. gallus* and *M. gallopavo* but less than 40% between these two species and *T. guttata*, showing that this region in birds is probably less constrained than the ASARM.

Rodentia and Scandentia selection signatures

The saturation plots (Figure 2-7A and 7B) showed that the rodents and the tree shrew have accumulated a very high number of transitions and transversions relative to other mammalian species (also apparent in the long branches of those species in the phylogenetic tree; Figure 2-3).

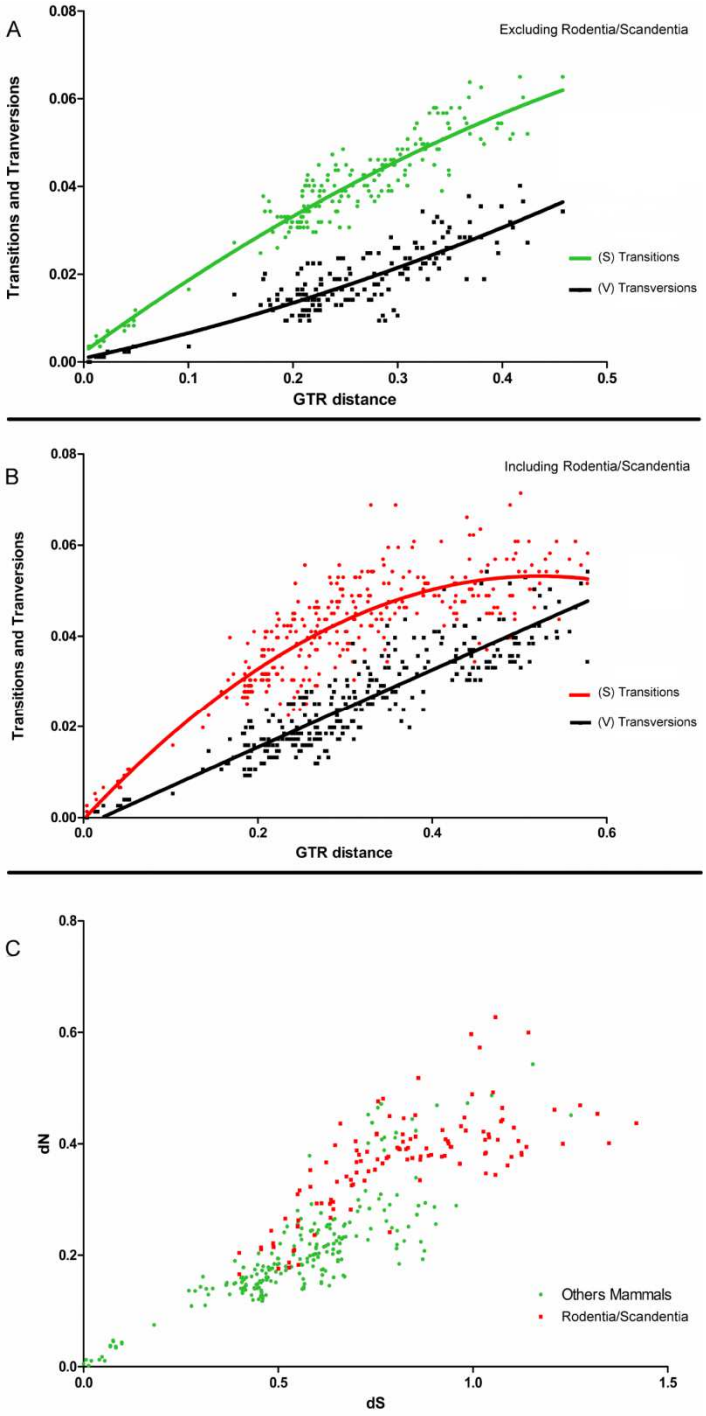

Figure 2-7. Accumulation of saturation and altered evolutionary rate in Rodentia and Scandentia compared with other mammals. (A) Nucleotide saturation plots excluding rodents and the tree shrew, showing transitions (S) and transversions (V) accumulated in the third position; and the same analysis (B) including the rodents and the tree shrew (C) Pairwise d_N/d_S comparison of rodents and the other mammals.

Saturation of synonymous mutations can bias the analysis of positive selection due to an underestimation of d_S that will increase ω (Lynn, Lloyd et al. 2004). Therefore, these species have been excluded from the codon and amino acid properties selection analyses (site models). When we grouped rodents and the tree shrew, and compared them with the

others mammals, the Relative Ratio Test (RRT) (Muse and Gaut 1997) showed that *MEPE* accumulated more mutations in the orders Rodentia and Scandentia (Table 2-1), with an average number of synonymous substitutions of 0.635 and non-synonymous substitutions of 0.304, in contrast with the other mammalian species with 0.527 and 0.235, respectively (both analyses being highly significant; P<0.025). The tree shrew and the rodents compared with the others mammals, had a higher GC percentage (49.9% versus 44.9%, respectively). This shows that Rodentia and Scandentia have accumulated more synonymous and non-synonymous substitutions (Figure 2-7C), which is consistent with the phylogenetic analyses that suggest that the rodents and the tree shrew have an accelerated rate of evolution.

Table 2-1. Results from the RRTree test comparing substitution rates in Rodentia, Scandentia and the other mammals.

Group	%GC	Ka	Ks
Rodentia and Scandentia (n=6)	49.9	0.304	0.635
Other Mammals (n=20)	44.9	0.235	0.527
p-value		<0.01	0.025

To evaluate if orders Rodentia and Scandentia have different sites under positive selection we compared the branch-site model A using the rodents (5 sequences) and tree shrew (1 sequence) as the foreground branch versus the other mammals as background branch (Appendices II: Table S3). The rodents had 12 sites under positive selection, with four of these being highly significant (PP>0.95) after the Bayes Empirical Bayes (BEB) analysis; 42-Tyr, 158-Lys, 239-Gly, 247-Asp (using the *Mus musculus* protein as reference). The likelihood ratio test (LRT) demonstrated that the branch-site analysis was statistically significant (P<0.04). Sliding window analysis using the Nei-Gojobori method also presented significant differences in the sites/regions under positive selection between the rodents/tree shrew and the others mammals (Figure 2-8). When we applied a window=15 and step=9, the rodents and the tree shrew showed eight regions with a $d_N/d_S>1$, while the others species had only two regions >1, suggesting that the rodents not only present an accelerated rate of evolution but also exhibited a different selection pattern in the protein (Figure 2-8).

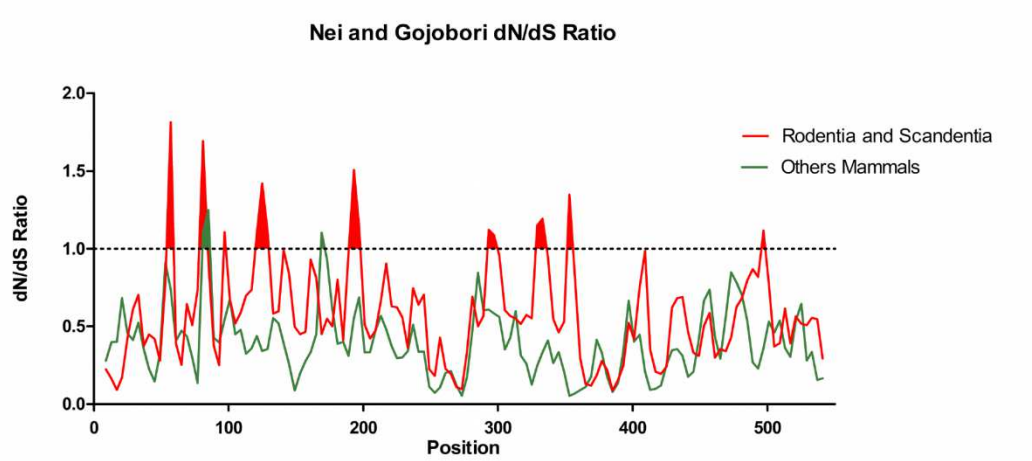

Figure 2-8. Differences in the selection pattern in Rodentia and Scandentia compared with other mammals. Sliding window analysis of the Ka/Ks ratio applying the Nei and Gojobori method for the rodents and tree shrew MEPE compared with the other mammalian species.

Selection signatures at the codon level

The *CODEML* test implemented in PAML was used to compare five different nested models in two situations, i.e. including or excluding the ambiguity data in the alignment. The MEPE protein had a global d_N/d_S ratio of 0.462, with 75 sites under negative selection and 17 sites under positive selection (Model 8 not removing the ambiguity data). When ambiguous data was removed the LRT's for the nested models, M1-M2, was rejected (Table 2), so the results of positive selection for M2 were not taken into account. In the LRT comparison between the more parameter-rich nested pairs of models (M8-M7), twice the log-likelihood difference was 7.1717 (Table 2-2), rejecting M7 and favoring M8 (Chi-square df =2; $p<0.05$). Under M8, 87% of the sites fit the ß distribution (1.584- 2.275), but 13% of the sites had a $\omega_1=1.30$. For posterior probabilities of $\omega>1$ using BEB with M8 vs. M7, nine sites were under positive selection (Table 2-2). However, none of these sites passed the stringent criterion of statistical significance PP>0.95 (using the method BEB as the statistical post-analysis). Additionally, the LRT between the M8 and the alternative null model M8a was 1.95, below the critical value (2.71 at $P<0.05$), and therefore not favoring the evolutionary model. However, it has been shown that in some cases this alternative LRT test has less power when the category of positively selected sites has a ω value that is only slightly larger than one (Swanson, Nielsen et al. 2003).

Table 2-2. PAML results of MEPE for the 20 mammalian species (excluding ambiguity data).

Model	Parameters	*LnL*	Test	LRT
Model 0	ω =0.46086	-7362.999747		
Model 1	p0=0.63812 p1=0.36188	-7308.340814		
Model 2	ω0=0.27631 ω1= 1.00000 ω2=1.00000			
	p0=0.63812 p1=0.27244 p2=0.08944	-7308.340814	M2 vs M1	0
Model 7	p= 1.06299 q= 1.09727			
		-7306.257659		
Model 8	p0=0.86974 p=1.58369 q= 2.27462			
	(p1= 0.13026) ω1= 1.29913	-7302.671784	M8 vs M7	7.1717

The evaluation of positive selection using the model implemented in Single Likelihood Ancestor Counting (SLAC) showed three sites under selection, one of those sites being similar to that retrieved with model M8 in PAML. Since SLAC tends to be quite conservative, we also estimated the selection signatures using the Fixed Effects Likelihood (FEL) model, which is assumed to be more powerful than SLAC (Pond and Frost 2005; Poon, Frost et al. 2009). The FEL model revealed a total of 23 sites under selection using this model, including a mutation in the highly conserved small peptide ASARM (from aspartate to glycine) (Figure 2-2). Such a radical change in ASARM was only observed in a few species and further studies are needed to better document its frequency across mammals. All the sites presented in the model implemented in Datamonkey have a significance value ($P<0.10$) in FEL and SLAC, which is an accepted level of significance for the test of those models (Pond and Frost 2005). When we use a significance threshold of 0.05, the number of positive selected amino acids decreased to 14 in FEL and zero in SLAC, meaning that 9 sites (out of the 23 detected with a significance level of 0.10) had less evidence of being under strong positive selection. However, these positions may still be indicative of selection signatures. Recombination can affect several analyses, including phylogenetic reconstruction and analysis of positive selection (Anisimova, Nielsen et al. 2003). Therefore, we assessed gene recombination using GARD implemented in the Datamonkey web-based server (Poon, Frost et al. 2009) and repeated the selection analysis including and excluding recombination in the dataset. Partitioning the data did not change the conclusions of the positive selection analyses (data not show), suggesting that recombination is not significantly affecting the *MEPE* gene evolution.

No additional sites were found using the SLR (Massingham and Goldman 2005), but six sites under selection in the previous analysis were also statistically supported. Overall, across MEPE, 32 of the 525 sites (referenced to the length of the human MEPE) were under positive selection; additionally six sites were supported by more than one codon analysis (9 in PAML, 23 in FEL, 3 in SLAC, and 6 in SLR) (Appendices II: Table S4).

Selection at the amino acid level

Selection models that use d_N/d_S ratios to detect selection are generally not sensitive enough to detect subtle molecular adaptations (McClellan, Palfreyman et al. 2005). It is therefore necessary to employ alternative criteria within generally conserved protein-coding genes or within proteins with strict motifs intermixed with regions under fast directional evolution. Therefore, we used TreeSAAP (Woolley, Johnson et al. 2003), which evaluates destabilizing radical changes at each site, and an empirical threshold of change in three properties was applied as evidence that a site is under positive (or negative) selection.

At the global protein level, eight of 31 amino acids properties were under strong positive selection in MEPE (P<0.001 for five and P<0.05 for the remaining three properties) (Table 2-3). Remarkably, pI is one of these eight properties under positive selection in *MEPE* which may also explain the high variability in pI observed across taxa (Figure 2-6).

Table 2-3. MEPE properties under positive selection determined in TreeSAAP.

Property	Category	Z-Score
Compressibility	7	3.783***
Equilibrium constant (ionization of COOH)	8	3.236***
Isoelectric point	8	3,418***
Power to be at the C-terminal	7	1.926*
Power to be at the middle of alpha-helix	7	3.757***
Power to be at the N-terminal	8	2.373**
Solvent accessible reduction ratio	7	3.953***
Turn tendencies	7	2.307*

List of properties under selection, the impact category and the level of significance (***:p<0.001; **:p<0.01; *:p<0.05).

At the amino acid site level, MEPE has 181 sites (33.8%) under positive selection in at least one property. Although applying the empirical threshold of at least three properties showing signatures of positive selection the number of sites is reduced to 41 (7.6%) (Appendices II: Table S5). The majority of these 41 sites are located in the N-terminal region of the protein and the dentonin region (68% of the positive selected sites). The alternative calculation method was performed using CONTEST and estimates of variation in amino acid charge and volume revealed 79 sites with signatures of positive selection for at least one of the amino acid properties (Appendices II: Table S6). However, after the Bonferroni and False Discovery Rate (FDR) correction, only one site showed positive selection. This site, located at position 354 in the alignment (position 349 in the human sequence), corresponds either to lysine or glutamate and was not detected by TreeSAAP. The ancestral

protein reconstruction in TreeSAAP, based on the baseml implemented in PAML, shows that glutamate is present in the common ancestor of non-Afrotheria mammals, suggesting that the radical change to lysine occurred in Cetartiodactyla, Perissodactyla and in at least one representative of the Lagomorpha.

Based on selection analyses at the protein level across MEPE, 42 of the 525 sites (human MEPE as reference) were under selection at the amino acid level (41 detected with TreeSAAP and 1 with CONTEST).

Selection at codon and amino acid level

We found 69 sites with signatures of positive selection, but there was concordance between codon and amino acid level methods for only 20 of these (Figure 2-9). More conservatively, the number of sites dropped to five if the most stringent and conservative criterion was used (requiring three properties under selection at amino acid level to be concordant with evidence from at least one codon-based-method).

		Position																			
		36	46	55	75	77	90	96	120	127	154	161	170	193	215	273	276	279	421	481	502
Analyses	Codeml	■					■	■			■			■	■			■			
	FEL	■	■	■	■	■		■	■	■		■	■	■		■					■
	SLAC	■			■									■							
	SLR				■			■									■		■	■	■
	TreeSAAP	■	■	■	■	■	■	+	+	■	■	■	■	+	+	+	■	■	■	■	■
Species	*H. sapiens*	R	G	N	S	A	F	L	S	E	I	I	E	D	R	G	L	K	L	R	F
	P. troglodytes	K	G	N	S	A	F	L	S	E	I	I	E	D	R	G	L	K	L	W	F
	G. gorilla	K	G	N	S	A	F	L	S	E	I	I	E	D	R	G	L	K	L	R	F
	P. pigmaeus	R	G	N	S	A	F	L	S	E	I	I	E	D	R	G	L	K	L	R	F
	M. fascicularis	K	G	Y	S	A	F	L	S	A	I	I	Q	D	R	G	L	K	L	R	L
	M. mulatta	R	G	Y	S	A	F	L	S	A	I	I	Q	D	R	G	L	K	L	R	L
	M. murinus	-	I	H	S	A	L	L	S	E	K	I	K	D	Y	G	P	V	V	R	F
	O. garnettii	P	T	N	S	S	L	L	S	D	V	M	K	D	H	G	L	M	-	-	-
	O. princeps	K	A	N	L	I	F	L	P	K	S	T	K	N	H	D	S	V	S	R	P
	O. cuniculus	-	A	N	L	T	S	L	P	D	T	T	K	E	H	G	P	I	V	Q	P
	M. lucifugus	R	A	S	L	T	S	T	S	E	G	T	E	G	L	D	L	T	S	-	-
	P. vampyrus	K	A	H	L	T	L	M	S	E	I	T	E	G	R	R	Q	T	S	W	F
	E. caballus	R	A	R	L	G	F	V	L	E	T	T	K	D	R	G	P	A	S	W	F
	C. familiaris	K	A	N	L	A	F	T	S	K	I	V	E	N	R	G	L	A	S	Q	F
	F. catus	R	A	N	V	A	F	T	S	E	I	T	E	D	H	G	L	T	S	W	A
	B. taurus	R	A	N	P	A	F	M	S	D	I	T	E	Q	L	G	R	T	-	-	-
	T. truncatus	K	A	N	P	V	F	M	S	E	I	I	D	I	H	G	R	A	V	W	V
	V. pacos	K	A	N	F	A	F	P	S	E	I	T	E	I	H	D	R	A	S	R	F
	P. capensis	K	A	K	L	A	P	R	S	E	R	N	K	Y	H	S	V	T	L	W	V
	L. africana	K	-	-	-	-	F	R	S	Q	L	T	K	G	R	G	L	T	L	-	-

Figure 2-9. Amino acids in the same evolutionary positions showing strong signatures of selection at the amino acids and the nucleotide level. Sites under positive selection confirmed by the different models used in this study for the dataset of 20 species (excluding rodents). The sites were numbered according to the Homo sapiens position [EMBL:ENST00000361056]. The results for SLAC, FEL, PAML (Model 8), TreeSAAP (at least one property under selection) and SLR are marked with a black box in the sites showing positive selection. The sites with more than three properties under selection in TreeSAAP are marked with a white plus symbol. The background colors represent the amino acids properties: polar positive (blue), polar negative (red and green), non-polar aliphatic (yellow), and P and G (pink).

Directed evolution analysis (DEPS)

MEPE evolution has disproportionally accumulated serines, threonines (potential phosphorylation target residues), arginines, alanines and valines, as all these amino acids showed directional evolution in the DEPS analysis (with a P-value<0.01) (Appendices II: Table S7). The MEPE protein had 14 sites under directional selection (Appendices II: Table S8), seven of which are amino acids that tend to increase the disorder/unstructured probability of the regions. Additionally, eight of these 14 sites had a tendency to change to amino acids that are potentially phosphorylated residues, particularly at positions 496 and 503 (505 and 512 positions in the alignment), since these sites are relatively near the ASARM motif and the cleavage site by cathepsin-B.

Selection Signatures and the MEPE structure

The MEPE protein belongs to a category of proteins classified as “intrinsically unstructured/natively disordered”, with 53.8% and 55.8% of the human and the mouse MEPE constituted by amino acids that are associated with disorder/unstructured regions, respectively. This is reinforced given that most of the protein (around 78.8%) is disordered at a 0.05% false positive rate. Interestingly, the ASARM motif has a high content of amino acids disorder promoters while the other functional motifs (such as RGD and SGDG) incorporate regions that are structured (Appendices II: Figure S3). The protein has a high percentage of the amino acid aspartate, which characterizes the proteins of the SIBLING family. Given the importance of disorder/order in MEPE, we analyzed the implications of selection signatures relative to the protein structural differences, and found that sites 75-Ser, 127-Glu, and 481-Arg (human MEPE as reference) are under positive selection and have a higher number of non-synonymous mutations towards codons that encode the amino acids disorder promoters.

The tertiary structure is similar to another extracellular matrix protein, anosmin-1 [PDB:1ZLG] with a Root Mean Square Deviation (RMSD) of 5.06. To determine if the spatial organization of these sites is associated with regions of functional importance, we plotted the positively selected sites (supported by at least two different inference methods) in the tertiary structure (Figure 2-10).

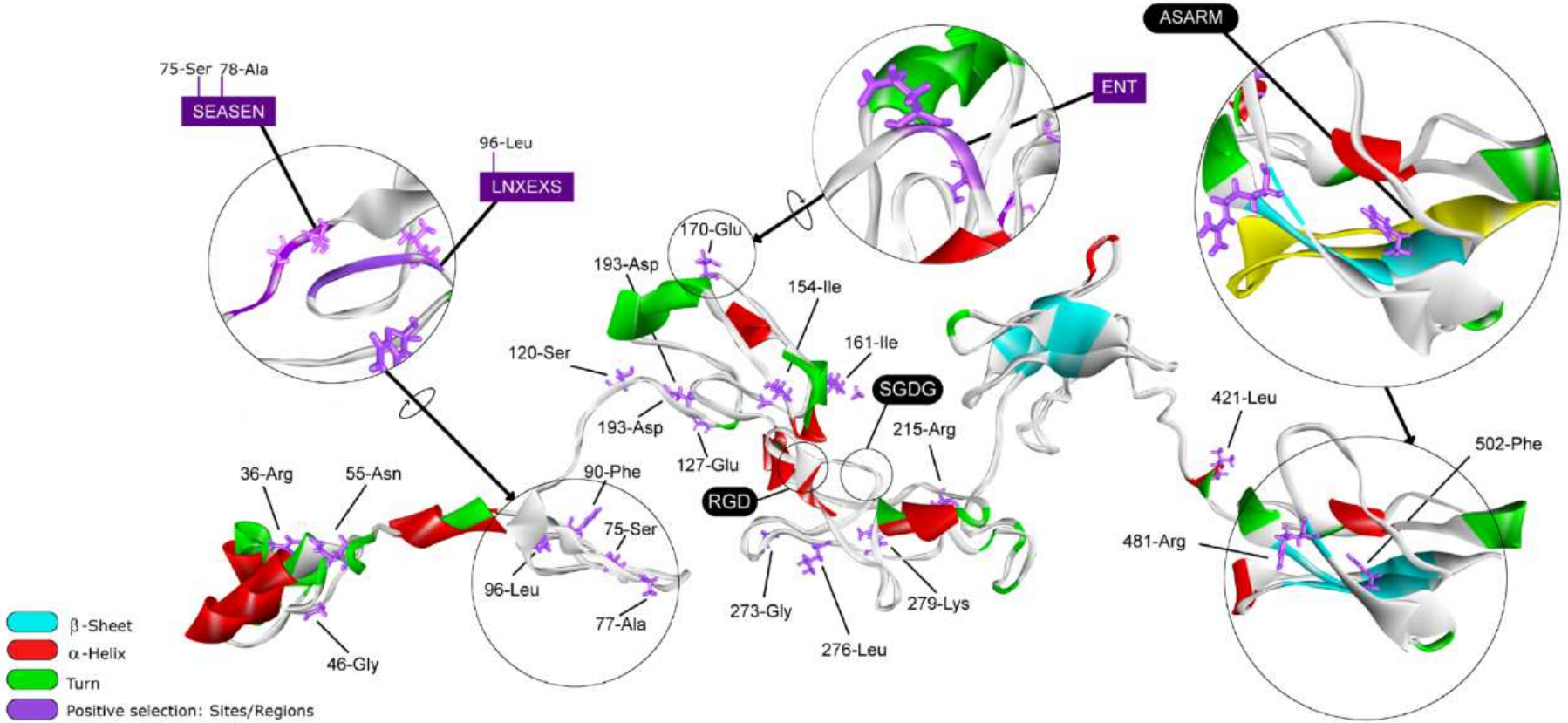

Figure 2-10. Tertiary structure of MEPE and the positive selected sites. The sites showing positive selection in at least two different analyses (purple sticks, three letters amino acid code) and the sites clustered showing positive selection in at least one analysis (close up circles, one letter amino acid code). The principal domains, RGD, SGDG and ASARM are also expanded and circled, as are the three "new" motifs (SEASEN, LNXEXS and ENT) showing positive selection. The secondary structure obtained is defined according to the code of colors described in the picture.

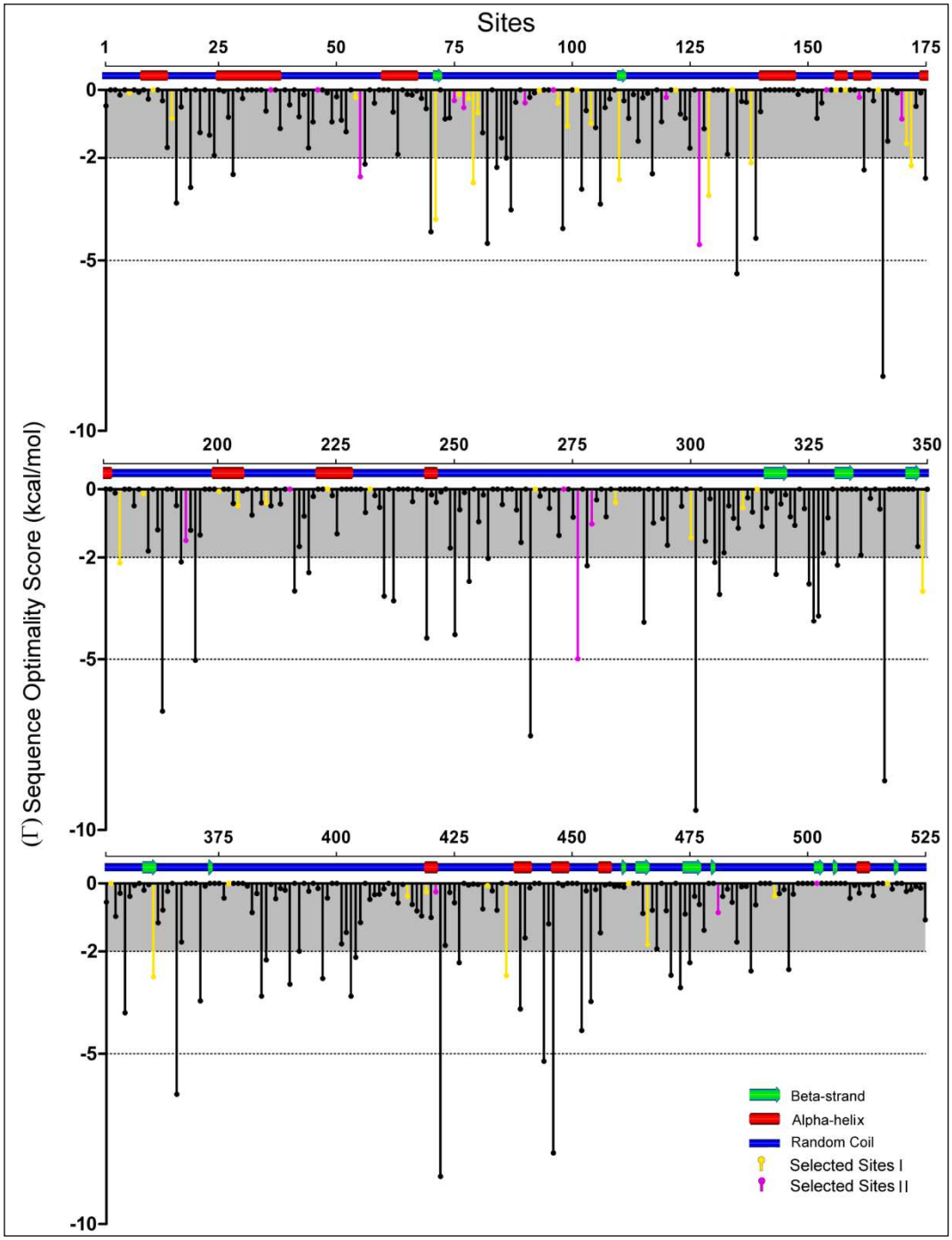

Figure 2-11. MEPE sequence optimality scores and the secondary structure. The sequence optimality scores (G) obtained in the Human MEPE, with the pink bars highlighting the sites under selection retrieved in both codon and amino acid level analyses and the yellow bars representing the sites showing selection in just one analysis (either codon or amino acid level methods). The secondary structure is represented in the top of the graph, with the nature represented: blue - random coil, green - ß-Strand and red - helices.

The sites showing selection signatures in both analyses are not restricted to any nature of the secondary structure (Figure 2-11) although most of the sites are located in random coils. In human MEPE, 69.3% of the amino acids are predicted to be found within

random coils, but when this analysis is restricted to the 69 sites under positive selection (retrieved considering either the codon or amino acid level method) the percentage increases to 71%. Of the 20 sites under selection (concordant sites retrieved simultaneously with codon and amino acid level methods) the percentage increases to 75%. This shows that the sites comprehending the random coils tend to have higher chances of being under selection. Similarly, the sites under positive selection tend to be in disordered regions, as 78.8% of the MEPE protein was “intrinsic disordered”. Of sites under selection in both analyses (codon and amino acid level), 90% were in disordered regions compared with 80% when considering all the sites under selection in at least one of the analyses. From the 20 sites under strong positive selection (concordant sites in both codon and amino acid level methods), 10 were solvent accessible, four were buried and the remaining six were in an intermediate category of neither buried nor exposed (Figure 2-12).

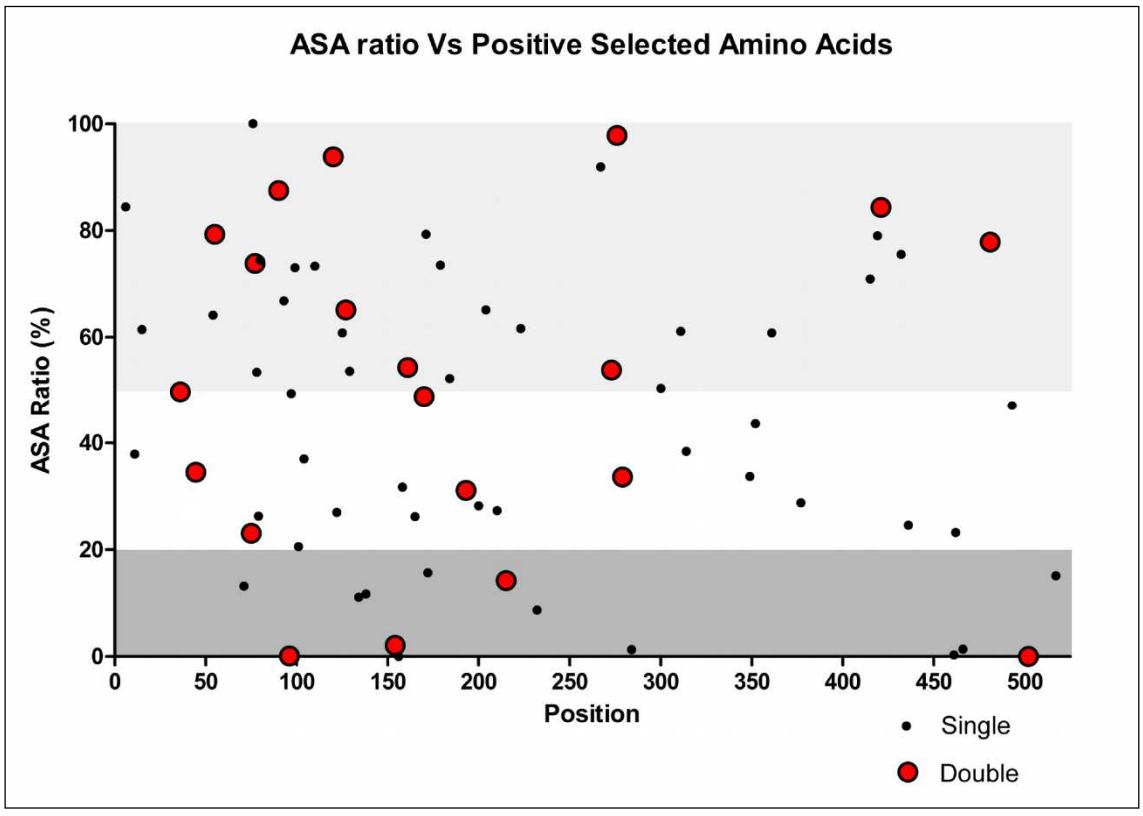

Figure 2-12. Exposure of residues to the exterior of the MEPE protein. Plot of the ASA ratio calculated between the side-chain and the 'random coil' value of each residue. Sites with a ratio above 50% (yellow box) are considered to be exposed to the outside of the protein whereas sites under 20% are considered to be buried (pink box). Sites under positive selection in both gene and protein-level analyses are marked with the red dots (double) and sites showing selection in at least one of the analysis is represented as black dots (single).

Estimates of protein stability revealed 11 sites of human MEPE with sequence optimality values (Γ) less than -5kcal/mol at positions 135, 166, 188, 195, 266, 301, 341, 366, 422,

444, and 446. While none of those sites correspond to a site with a signature of positive selection, when the Γ empirical cut-off is reduced to -2kcal/mol the number of sites with a non-optimal state increases to 75. Of these, three sites are under positive selection based on both codon and amino acid analyses, and 11 of these sites show evidence of being under positive selection at either the codon or amino acid level.

2.5 Discussion

MEPE in the Tetrapods

Given the absence of the *MEPE* gene in fishes and amphibians, its origin likely coincides with the divergence of amniotes, when mineralization (Gowen, Petersen et al. 2003; Rowe, Kumagai et al. 2003) and phosphate regulation (Quarles 2003) had a crucial role in species survival and diversification. *SPP1* diverged from *SPARCL1* (secreted protein acidic cysteine-rich like 1) and both are expressed in bone, participating in the bone formation (as an inhibitor of mineralization in *SPP1*) (Kawasaki, Suzuki et al. 2004). Therefore, the presence of SPP1 in fishes with a broader tissue expression pattern (Kawasaki, Buchanan et al. 2009) suggests that *SPP1* might also have similar functions to *MEPE*. Remarkably, after duplication, the genes were conserved during evolution and probably have differentiated to assume various functions related with tissue mineralization specificity. Recently, it was proposed that *SPP1* is a more-powerful inhibitor of mineralization than *MEPE* (Addison, Masica et al. 2010). This suggests that after the emergence of the complete SIBLING family in vertebrates, some functions were possibly shared among genes, notably because MEPE is absent in fishes.

The *MEPE* gene has similarities with other SIBLING genes, suggesting that it originated through a duplication event from another member of the gene family (Kawasaki and Weiss 2006), but different dynamics of gene duplication and gene loss have occurred among lineages (e.g. absence of *MEPE* - Figure 2-1). The five genes of the SIBLING family are present in therian mammals and reptiles, but birds only have four genes (IBSP, SPP1, *DMP1* and *MEPE/OC-116*), while fish only have two genes (*SPP1* and *DSPP*-like). The *DSPP* orthology in fishes is controversial (Kawasaki, Buchanan et al. 2009). However, despite the low similarity, DSPP *starmaker* was identified as a functional orthologue (Sollner, Burghammer et al. 2003) clearly associated with *DSPP* (Ramialison, Bajoghli et al. 2009). The presence/absence of various SIBLING family genes in vertebrates suggests that despite the crucial role of *MEPE* in mammals, birds and reptiles, its function may have been compensated in other taxa by other genes of the family. For example, in fishes a duplicated copy of *SPP1* has not been described, suggesting that the fish *SPP1* orthologue may have

had a similar function to *MEPE* since *SPP1* and *MEPE*, interact with *PHEX* (Addison, Masica et al. 2010). The release of the ASARM from the MEPE protein and the phosphorylation of this motif lead to an inhibition of mineralization (Rowe, Kumagai et al. 2003). Similarly, the ASARM from SPP1 inhibit the mineralization (Addison, Masica et al. 2010). Moreover, the ASARM from SPP1 is potentially phosphorylated and can interact with the hydroxyapatite crystals leading to a negative regulation of mineralization (Addison, Masica et al. 2010). Although SPP1 has an ASARM motif near the center of the molecule, it does not have the full dentonin region (just the RGD motif). Moreover, the SPP1-ASARM has been described as a more-potent mineralization inhibitor than the MEPE-ASARM (Addison, Masica et al. 2010). However, the knockouts of SPP1 and MEPE in mice have different phenotypes. MEPE knockouts have increased bone mass and inhibition of age-related bone loss (Gowen, Petersen et al. 2003) while SPP1 knockouts cause a resistance to bone loss and trabecular bone mass (Yoshitake, Rittling et al. 1999).

Functional Conservation

The functional motifs of MEPE (RGD, SGDG and ASARM) are highly conserved among the studied mammals. In the SIBLING proteins the first coding exon encodes the signal peptides (Fisher, Torchia et al. 2001; Fisher and Fedarko 2003), as is observed in MEPE. The RGD motif is a common feature of all member of the SIBLING family, remaining functionally preserved after the tandem duplication that gave rise to all the members of this gene family (Fisher and Fedarko 2003). Surprisingly, birds do not have a complete dentonin region (RGD and SGDG), although the high conservation observed among mammals suggests that this region has an important role in the function of the protein. In fact, in mammals the gene function apparently depends on the full dentonin region, as the RGD motif alone does not enhance an optimal adhesion on biomaterial surfaces in osteoblast (Dee, Andersen et al. 1998). However, when SGDG is close to RGD the mitogenic activity of dentonin increases, while the presence of only the SGDG motif promotes the cell proliferation (Liu, Li et al. 2004). In mammals, MEPE is involved in bone formation and osteoblast proliferation (Hayashibara, Hiraga et al. 2004; Liu, Li et al. 2004), while in birds it is involved in egg-shell formation (Hincke, Gautron et al. 1999). This functional divergence may explain the sequence differences observed between the two lineages, particularly reinforced by the absence of the full dentonin region in birds. The ASARM motif is also highly conserved among mammals, but shares less than 50% similarity with the avian ASARM. Moreover, we have not detected a similar cathepsin-B cleavage site near avian MEPE-ASARM and this motif is capped at the C-terminal by 21 to 24 amino acids. Amino acids towards the C-terminal after the ASARM motif are also observed in marsupials (Bardet, Delgado et al.

2010). Despite the lower similarity with the mammalian ASARM and its different position, the high conservation within birds suggests that this motif continues to have a crucial role. The changes are probably not due a relaxation of selection, but instead may have an adaptive role. In mammals, the cathepsin-B cleavage site is crucial for the function of MEPE, since this small peptide only interferes with hydroxyapatite crystals when released (Rowe, Kumagai et al. 2004). Therefore, birds are also expected to have a mechanism for cleavage of ASARM. MEPE has not yet been annotated in a monotremata, no significant matches were found in a representative species of this group, the platypus (*Ornithorhynchus anatinus*). Nevertheless, the discovery of this gene in egg-laying mammals would be of great relevance to understanding the functional differences between mammals and birds.

The coding region of MEPE that flank the motifs described above is less conserved, but retain considerable phylogenetic signal across species. The human MEPE sequence has a high similarity with the great apes and with the genus *Macaca* (Cercophitecidae), even in the non-coding regions (*M. mulatta*). To a lesser degree, human MEPE also has some significant similarities in the non-coding regions with the genes of the lower primates (*M. murinus* and *O. garnetti*). MEPE appears to be particularly conserved among primates, in both coding and non-coding regions. The intronic conservation could provide valuable information about the role of non-coding sequences in the regulation/functionality of this gene. Despite the accelerated evolution in rodents, intronic conservation allowed us to reconstruct a well-supported species phylogeny from intronic sequences (even including the rodents sequences) with similar results as those obtained from *MEPE* coding regions.

Several human diseases increase *MEPE* expression (Rowe, de Zoysa et al. 2000; De Beur, Finnegan et al. 2002; Brame, White et al. 2004), which may imply functional constraints in the gene even at the intronic level. Previous studies have demonstrated that highly conserved intronic regions are correlated with functional constraints and can be evidence of a hidden class of abundant regulatory elements (Hare and Palumbi 2003). Recently, a SNP in the region 7 kb 3' of the gene was associated with osteoporosis, a disease characterized by reduced bone mass and microarchitectural deterioration of bone tissue that reduces bone strength and leads to an increased risk of fracture (Rivadeneira, Styrkarsdottir et al. 2009). These findings suggest that intergenic regions can also be important in gene function and may cause significantly different phenotypes. We hypothesize that intronic regions can also lead to significant differences at the expression level and ultimately to differences in phenotypes. This is consistent with our findings that there are strong evolutionary constraints in the *MEPE* intronic region.

Selection signatures and conservation

Within mammals, *MEPE* in rodents is evolving faster, presenting a high amount of transitions and transversions. A similar trend is also observed in the tree shrew *T. belangeri*. However, since we only had one MEPE sequence from the order Scandentia we were unable to infer if this pattern is species-specific or if it is typical of this order. The increased number of substitutions in rodents was expected as previous studies have shown that rodents tend to accumulate more mutations in the coding regions (Wu and Li 1985; Li, Ellsworth et al. 1996). We hypothesize that the observed differences in these two orders have resulted from either a divergent functional role or simply a relaxation in Darwinian selection. It is not known if the function of the rodent *MEPE* is similar to that in humans (Liu, Wang et al. 2009), but all the functional motifs are conserved and the signatures of positive selection or the differences observed were only detected outside of these important motifs. It is clear that positive selection may have an important role in the functional divergence of homologous proteins during adaptation to different habitats (Levasseur, Gouret et al. 2006). Indeed, selection may be episodic as positive and negative selection shifts over time across different lineages, reinforcing the importance of comparing sequences that have diverged within appropriate time frames (Messier and Stewart 1997).The branch-site model, using rodents as foreground branches and allowing ω ratio variation not only between the branches but also among sites, identified 12 sites with strong signatures of positive selection. This suggests that the rodents and probably Scandentia may have lineage-specific selection differences in *MEPE*, not only in the magnitude of the selective pressure found in the branch, but also in the number of sites under selection. The acceleration of the substitution rates in rodents and the tree shrew potentially compromises the assessment of positive selection by increasing the number of synonymous mutations and because this heterogeneous site selection is observed in only two of the eight orders evaluated (i.e. Rodentia and Scandentia). The results may also be biased by the mixing of species with long and short generation times (Li, Ellsworth et al. 1996), as well as the related long-branch-attraction effect in phylogenetic reconstruction. Therefore, we did not include the rodents and the tree shrew in the site analysis.

The evolutionary analyses of mammalian *MEPE* codons (excluding the rodents and the tree shrew) found 32 sites under positive selection at codon level, and remarkably three were in functional regions of the protein, positions 6-Val and 11-Phe (Signal Peptide) and position 517-Gly (ASARM motif) (Figure 2-2).

Recent methods for investigating selection in protein coding genes have focused on evaluating the type of positive selection detected (directional or nondirectional, stabilizing

or destabilizing), determining the presence of purifying selection, and interpreting how selection affects overall protein structure and function. Amino acid substitutions have different effects on a protein depending on differences in physicochemical properties and their position in the protein structure (Antunes and Ramos 2007). Here, we performed multiple analyses to differentiate among the different types of selective pressures acting in MEPE at the amino acid level. The evaluation of the amino acid physiochemical properties changes in the mammalian MEPE identified 37 more sites (36 using TreeSAAP and one using CONTEST) with selection signatures compared with the results retrieved using codon models. This shows that total reliance on models based on d_N/d_S using codon models may not detect some important sites with signatures of selection, often because a single adaptive mutation may occur in a small number of species, resulting in an omega lower than one. By contrast, these could also primarily be amino acid stabilizing rather than destabilizing changes, and a $\omega>1$ may not always be indicative of adaptive evolution.

Combining all the selection analyses, we found 69 amino acids with evidence of positive selection (20 well-supported by both codon and amino acid level approaches) (Appendices II: Figure S4). Three clusters of positively selected sites revealed three new motifs that likely have a functional role, SEASEN (75-80), LNXEXS (96-101) and ENT (170-172) (Figure 2-10), using the human protein as site reference.

Selection analysis of *MEPE* in TreeSAAP using amino acid destabilizing properties revealed that the structural properties tend to be more affected by positive selection than the chemical properties. This suggests that the flexible and intrinsically unstructured nature of *MEPE* is linked to its multiple biological roles. The ASARM motif shows a “high tendency” to be a “disordered region and highly acidic”, although the conformation of ASARM should be dependent on the phosphorylation level (Martin, David et al. 2008). The ability to bind to hydroxyapatite is also correlated with phosphorylation state and PHEX cleavage of MEPE is dependent on the Serine phosphorylation status (Addison, Nakano et al. 2008). Moreover, our results shows that the protein tends to accumulate numerous residues with potential phosphorylation sites and this can be important to the folding/function of the protein. Proteins fold to minimize their free energy, although the structure also reflects an organization that can allow the recognition of a ligand or a transition state (Shoichet, Baase et al. 1995). In fact, there is a balance between protein function and stability, and most of functional sites are non-optimal in terms of stability. If a residue is replaced by another residue, the protein activity will be reduced but the stability will be increased (Shoichet, Baase et al. 1995). In MEPE we detected 75 sites with a Γ lower than -2kcal/mol, indicating that a large number of sites in MEPE are non-optimal and therefore possibly involved in protein function. Moreover, 13 of those sites showed signatures of selection in one analysis, and sites 55, 127 and 276 in both codon and amino acid level analyses. Proteins have different secondary-

structures and physicochemical properties and roles that help determine their evolutionary flexibility (Ridout, Dixon et al. 2010). Thus, amino acids that comprise disordered regions, such as random coils, are more likely to be under positive selection than expected from their proportion in the proteins, compared with the residues in helices and β-structures which are subjected to less positive selection (Ridout, Dixon et al. 2010). Indeed, when we compare the evidence of positive selection with the protein secondary structure in MEPE we observed that the number of sites under selection in the random coils and disordered regions are slightly higher than expected. This suggests that a high number of sites probably have a functional role or are at least relevant to an increase in MEPE protein flexibility.

Presently, most of the research on MEPE has centered on the biological role of the RGD and ASARM regions. However, our comparative study of mammalian *MEPE* orthologues revealed that the protein has lineage-specific properties (e.g. biochemical, evolutionary rate, intronic conservation), and that outside these two well-described motifs there are 69 sites (20 with high confidence level) under positive selection and of probable functional relevance. As positively selected sites might be either near catalytically important regions of the proteins (Morgan, Loughran et al. 2010) or be functionally relevant sites (Casasoli, Federici et al. 2009; Moury and Simon 2011), these sites are good candidates for mutagenesis and structural studies to determine the functionality of MEPE relative with the other SIBLING proteins.

2.6 Conclusions

MEPE is found in reptiles, birds and mammals (eutheria and metatheria), and to date has not been identified in monotremes. The description and study of MEPE in other taxonomic groups will be crucial to fully understanding the differences reported in avian and mammalian orthologues, and the adaptive significance of these differences. The absence of this gene in some vertebrate lineages suggests that SPP1 might partially cover the functions of MEPE in those groups. *MEPE* retains a strong phylogenetic signal at both coding and non-coding regions in mammals, probably due to in the functional relevance of these regions. Nevertheless, the gene is highly variable, particularly in the largest exon outside the functional motif, while other regions appear to be under strong positive selection. We found 20 sites with a significant signature of positive selection at both nucleotide and amino acid level complimentary analyses (in addition to other 69 sites with evidence of selection at either the nucleotide or the amino acid level). The analyses identified three motifs (LNXEXS, SEASEN and ENT) with selection signatures suggesting important adaptive functions. We also showed that Rodentia and Scandentia have an accelerated evolutionary rate with a unique evolutionary pattern. Finally, we showed that MEPE tends to accumulate

amino acids that promote “disorder” and that present potential phosphorylation targets, supporting the contention that other regions outside the dentonin and ASARM might have crucial functional roles and demonstrating the need for future studies to understand the importance of these regions.

2.7 Acknowledgements

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to JPM (SFRH/BD/65245/2009) and the project PTDC/BIA-BDE/69144/2006 (FCOMP-01-0124-FEDER-007065) and PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610). This work was further supported by a grant from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism. We thank Siby Philip from LEGE/CIIMAR for discussion and helpful suggestions. Comments made by the anonymous reviewers improved a previous version of this manuscript.

Chapter 3 - *Convergent selection in bone-associated genes modeled through the evolution of flight in birds and bats*

3.1 Abstract

Bones have been subjected to considerable selective pressure throughout vertebrate evolution, especially during events such as the adaptation of powered flight. We performed comparative genomic analyses of 89 bone-associated genes from 47 avian (including 45 of newly genomes) and 39 mammalian genomes, demonstrating that birds had almost twice the number of bone-associated genes under positive selection (55% or 49 of 89 genes) as mammals (~35% or 31 genes). Remarkably, after removing flightless birds and the bat, positive selection in these bone-associated genes increased within the remaining birds (~58.4%; 53 genes) and decreased in the remaining mammals (~26%; 22 genes). Twelve of the positively selected genes in birds have been linked with functional pathways that clearly would be relevant to powered flight, including bone metabolism, bone fusion, and reactive oxygen species (ROS) production, and each of these had different evolutionary rate in bats than terrestrial mammals. In birds, genes involved in bone resorption, such as *TPP1*, had a high number of sites under Darwinian selection, of which five were determined to be functionally-active. The high levels of positive selection observed in bird and bat ossification genes suggest that there was a period of intense selective pressure to improve flight efficiency that was probably closely linked with constraints on body size. Most of the positive-selected genes in birds have previously been linked with bone regulation and remodeling, while in mammals genes under positive selection were generally related with bone development.

3.2 Introduction

Powered flight evolved independently in birds and bats, but required similar tradeoffs and limitations, including strong constraints on traits such body size (Smith, Brown et al. 2004; Puttick, Thomas et al. 2014) and skeletal structure to minimize energy requirements (Dumont 2010). While body sizes have tended to increase through evolutionary time in many lineages (Hone, Dyke et al. 2008), the sizes of flying vertebrates have been more constrained (Alexander 1998). However, postcranial skeleton pneumatization (hollow air-filled bones) and bone modifications (such as. bone fusion) may have provided increased evolutionary flexibility among birds (Gutzwiller, Su et al. 2013) (Figure 3-1A).

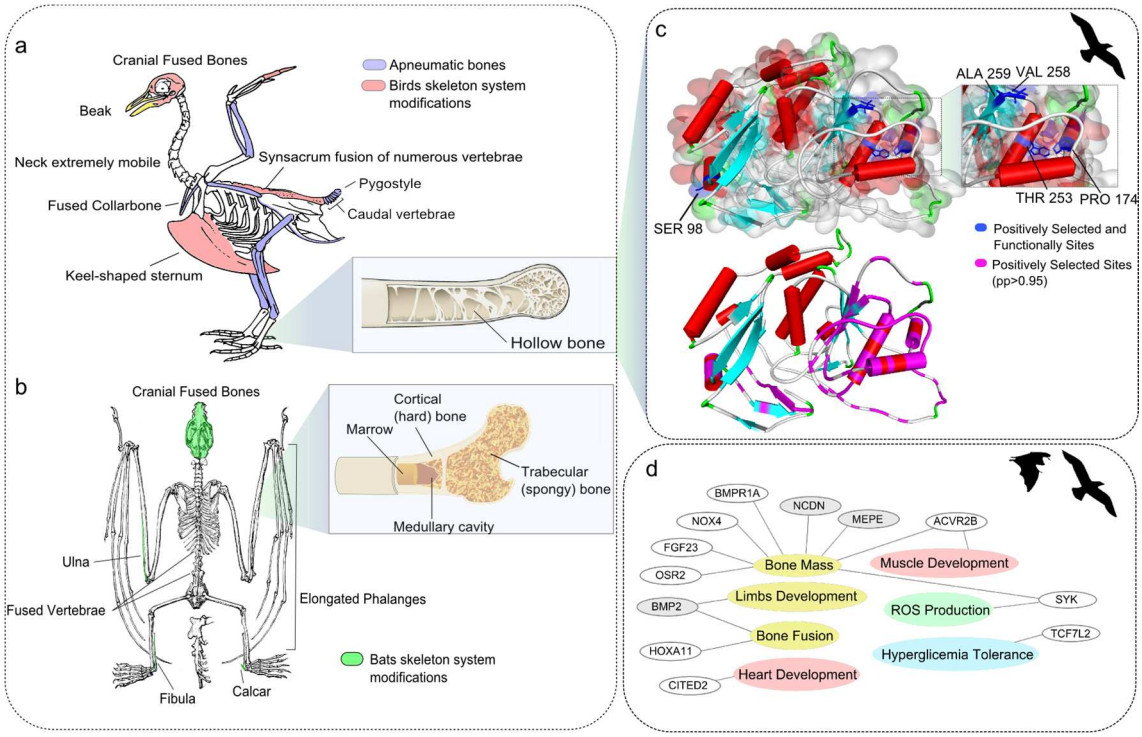

Figure 3-1. Skeleton adaptations in birds and mammals and adaptive selection in bone-associated genes. a) Rock pigeon skeleton (adapted from Wikimedia Commons licensed under a Creative Commons Attribution-Share Alike 3.0 Unported (CC BY-SA 3.0)) showing the key bone modifications observed in birds (red and yellow), and bones containing red-blood-cell-producing marrow (apneumatic bones, blue shaded). Most bones (except very small ones) are pneumatized (Schepelmann 1990). The structure of a pneumatic bone is highlighted in the light blue box (licensed by Rice University under a Creative Commons Attribution License (CC-BY 3.0)). b) Skeleton of Large flying fox (adapted from Wikimedia Commons licensed under a Creative Commons Attribution-Share Alike 3.0 Unported (CC BY-SA 3.0)) and the key features observed in bats skeleton system. The typical bone structure of long bones is highlighted in the light blue box (adapted from Wikimedia Commons licensed under a Creative Commons Attribution-Share Alike 3.0 Unported (CC BY-SA 3.0)). c) Example gene, *TPP1*, involved in bone resorption, and the positively selected and functionally relevant sites. Sites are labeled using the three letter amino acid code and numbered for their protein position in the zebra finch sequence. Positively selected sites in Model 2a with a posterior probability to be under positive selection above 0.95. d) Positively selected genes in birds and those genes showing a dissimilar evolutionary rate in bats when compared to other mammals (lower evolutionary rate - colored in grey; and higher evolutionary rate - colored in white). Representation of the link between gene and physiological/development systems (colored accordingly: skeleton system (yellow), muscular system (red), glucose (blue) and cellular metabolism (green)) plausible related with flight adaptation.

In birds hollow bones are formed with pneumatic foramina or openings in the wall of the bone that permit air sacs to perforate internal bone cavities (Currey 2003; Fastovsky and Weishampel 2005). While most flying birds have pneumatized bones, these are absent in flightless birds (Smith 2012), with a few notable exceptions such as gulls, which lack pneumatization or the extinct elephant bird Aepyornis, a running bird that had pneumatized femurs (Cubo and Casinos 2000; Fastovsky and Weishampel 2005). The development of

pneumatic bones in birds would have led to reductions in overall body mass and has also been associated with bone resorption (Smith, Rossie et al. 2005; Gutzwiller, Su et al. 2013). These pneumatic bones have often been assumed to have lightened the entire avian skeleton relative to mammals (Prange, Anderson et al. 1979) and to have reduced the metabolic cost of flight (Fedducia 1996; Podulka, Rohrbaugh et al. 2004; Gill 2007; Freeman, Sharp et al. 2008; Dumont 2010). However, some skeletal structures, such as the humerus, ulna-radius, tibio-tarsus and fibula, have more body mass in birds than mammals (Cubo and Casinos 1994), suggesting that modern bird skeletons have experienced diverse bone-specific selection patterns.

Recent research suggests that bird flight first evolved in cursorial birds with predominantly terrestrial lifestyles (Appendices III: Figure S1) (Dececchi and Larsson 2011). Some avian species (e.g., kiwi, penguins, and ratites) are thought to have lost, secondarily and independently, their ability to fly (Roots 2006) (Appendices III: Figure S1). Flight degeneration was probably a lengthy process and some extant species (e.g., chicken and turkey) retain transitional locomotive models that impose different selective constraints (or relaxation of the existing ones) on bone-associated genes (Shen, Shi et al. 2009). These differences among birds provide the opportunity to test correlations among species with different flight capabilities.

Bats, the only mammals capable of sustained flight, are unique from birds in several key ways that likely reflect different ecological adaptations and distinct evolutionary histories (Dececchi and Larsson 2011). Morphologically, bats do not have feathers and have elongated fingers instead of elongated forearms as seen in birds. In addition, bats, in contrast with birds, likely developed powered flight after first being able to glide (Shen, Liang et al. 2010) (Figure 3-1A; Figure 3-1B) and have bones with high levels of mineral density that increases the stiffness of the skeleton (Dumont 2010). On the other hand, as with birds, bats have relatively small bodies (Maurer, Brown et al. 2004), fused bones (Dumont 2010) and lightweight skeletons. In addition to skeleton structure, many of the other shared traits among birds and bats are probably also associated with the challenges imposed by the evolution of powered flight. These include improved respiratory systems (Thomas, Follette et al. 1995), high metabolic output (Ward, Bishop et al. 2002), hyperglycemia tolerance (Braun and Sweazea 2008; Kelm, Simon et al. 2011), diminished production of reactive oxidative species (Barja 1998; Brunet-Rossinni 2004) and smaller intestines (Caviedes-Vidal, McWhorter et al. 2007).

Here, we tested the evolutionary rate of change in 89 bone-associated genes in 47 avian and 39 mammalian genomes and evaluated genetic distinctions among flying versus non-flying species to assess patterns of selection in genes involved in bone development. Birds displayed almost a two-fold increase in the number of the bone-associated genes

under positive selection, and the majority of which were associated with regulatory process of bone remodeling. In contrast, most mammal bone-associated genes under positive selection are linked with bone development, while bats shows different evolutionary rates in bone remodeling genes.

3.3 Methods

Sequences and alignment

A list of bone-associated genes was retrieved from the GO database by querying the term “bone” in QuickGO (Binns, Dimmer et al. 2009). The present avian dataset encompass 89 bone-associated genes (Appendices III: Table S1), performing 3,388 sequences, ~38 sequences per alignment, derived from 47 bird genomes provided by the Avian Genome Consortium (Zhang). Sequences for each gene were translated to amino acids, aligned using MUSCLE (Edgar 2004) and back-translated to nucleotides. Aberrant sequences, sequences containing frame-shifts (e.g. stop codons) and duplicated sequences were removed from the multiple sequence alignment (MSA), resulting in an average of 38 species per gene. The mammalian dataset was derived from 39 genomes (2,903 sequences, ~32 per gene) that were manually retrieved from ENSEMBL (Flicek, Ahmed et al. 2013; Flicek, Amode et al. 2014). The MSA of each gene (composed on average by 33 sequences) was built using the same strategy as with the avian genes. The 89 genes were concatenated using SequenceMatrix v 1.7.8 (Vaidya, Lohman et al. 2011) to a one MSA containing all the avian data, and second MSA containing the 89 mammalian genes. A phylogenetic tree was built separately for birds and mammals using the 89 concatenated genes with PhyML v3.0 (Guindon, Delsuc et al. 2009) under the Generalized Time-Reversible (GTR+Γ+I) model and the branch-support was provided by aLRT (Anisimova and Gascuel 2006).

To control random associations between bats and birds, we tested a second dataset encompassing 50 genes associated with brain development. The sequences were retrieved and aligned using the same procedure used for the bone-associated genes.

Selection tests

Site Models

CODEML, as implemented in PAML v4.7 (Yang 1997; Yang 2007), was used to test for selection signatures in the avian and mammalian bone genes using three models (Models 0, 1 and 2). Model 0 was used to test the global selective pattern and the likelihood of

each model was estimated from a nested comparison of models M1 vs M2. Sites with significant signatures of selection were retrieved after a post-hoc analysis using Bayesian Empirical Bayes (BEB), which accounts for errors and is therefore more suitable than Naive Empirical Bayes (a less reliable analysis particularly for smaller datasets) (Yang, Wong et al. 2005). To test for potential confounding effects of including flightless birds or flying mammals, the same analyses were repeated using site models to test for signatures of selection after removing these species from the dataset using Phyutility (Smith and Dunn 2008).

Branch models

We tested for selection using branch models with two-ratio models that permitted variation in the omega (ω) ratio between the background and foreground branches. The two-ratio models were compared against a one-ratio model where no variation within the tree is allowed. In the bird and mammal datasets the "exceptions" (flightless birds and flying mammals) were compared against the flying birds and flightless mammals, providing an understanding of which genes were under differential selection patterns in the two clades. Spearman's correlations were performed in SPSS v20 (SPSS Released 2011).

Three-Dimensional Structure Modeling

To determine the physical location of the positive-selected amino acids in the 3D *TPP1* zebra finch protein structure, we ran I-TASSER (Zhang 2008). The model had a TM Score of 0.97± 0.05 and C-Score = 1.81 (from a possible range of -5 to 2). An accurately-inferred topology should have a C-score above -1.5 (Zhang 2008) and a TM score above 0.5 suggests that the obtained topology was not random.

Gene set enrichment

Functional annotation enrichment analyses were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID v6.7) (Huang da, Sherman et al. 2007; Huang da, Sherman et al. 2009). Each derived gene list was processed in DAVID for annotation of functional terms. Venn diagrams were generated using VENNY (Oliveros 2007).

Correlation model between body mass and bone-associated genes

CoEvol 1.3c (Lartillot and Poujol 2011) implements a phylogenetic model that correlates the evolution of substitution rates (e.g. d_s, ω) with continuous phenotypic characters (e.g. body mass, longevity). The MSA of the 89 bone-associated genes were divided into

two different datasets, one including all birds, and the other restricted to only the flying bird species. The same analyses were also run for all mammals and flightless mammals. To ensure convergence, we ran two different chains to at least an effective number of 50. Calibration of the tree was done using the divergence-time-based option in TimeTree (Hedges, Dudley et al. 2006) (Appendices III: Table S2). Body mass estimates were retrieved from ADW (Myers, Espinosa et al. 2014) (Appendices III: Table S3).

CoEvol models evolutionary rates of substitution and phenotypic characters as a multivariate Brownian diffusion process along the branches, correcting for the uncertainty about branch lengths and substitution history in the phylogenetic tree. Correlations among rates of substitution and phenotypic characters were calculated with posterior probabilities varying from 0 to 1 using a Bayesian Markov chain Monte Carlo and correcting for phylogenetic *inertia* using the independent contrast method. Posterior probabilities (pp) close to 0 indicate a negative correlation while values close to 1 indicate a positive correlation. Cutoffs of $pp < 0.05$ and $pp > 0.95$ suggest negative or positive covariance between the substitution rates and the phenotypic trait, respectively.

3.4 Results

Gene length and location in the genome

The 89 bone-related genes (Appendices III: Table S1) represent a subset of the genes associated with bone development (Bassett, Gogakos et al. 2012). These bone-associated genes were distributed widely across the genomes of mammals and birds (Figure 3-2). The mean coding-sequence length was longer in mammals than in birds (1599.8 vs 1385.1 bp; Mann-Whitney U, $p=0.005$), while slightly longer in flightless birds than flying birds (1445.3 vs 1374.4 bp; Mann-Whitney U, $p=0.63$), was slightly longer (1.3%) in flying mammals (1644.7 vs 1622.4 bp for flying versus flightless).

Figure 3-2. Genomic location of bone-associated genes. The circular ideogram represents the genomic location of bone-associated genes in four of the studied species. Triangles represents the gene position of the bone-associated genes. Blue indicates human chromosomes (mammal representative). Dark orange the zebra finch (flying species) and green and yellow the chicken and turkey (flightless species), respectively.

Selection on coding sequences

Sites models show higher positive selection in bird bones genes

We tested for signatures of selection in 89 bone-associated genes from 39 mammalian and 47 avian genomes using "site models" and "branch models". In site models, a site-class is considered to have evolved under a positive selection when the Likelihood-ratio test (LRT) of the nested models (Model 2a vs 1a) is statistically significant when compared with a χ^2 distribution (p-value < 0.05). Of the 89 analyzed mammalian genes, 31 (~25%) had favored the alternate model (evolved under positive selection) (Figure 3-3; Appendices III: Table S4). In birds, 55% (49 of 89) of the bone genes were positively selected (Figure 3-3; Appendices III: Table S5). The difference between the number of positively selected genes

among birds is significantly higher in birds than in mammals (z-score=2.68, two-tailed p=0.0073). Similarities in genes under positive selection observed in the mammalian and avian dataset is restricted to 19 out of 89 bone-associated genes (~21%), while 30 (57%) were positively selected in mammals; of the bird genes, 12 (38%) were exclusively under positive selection in birds (Figure 3-4A).

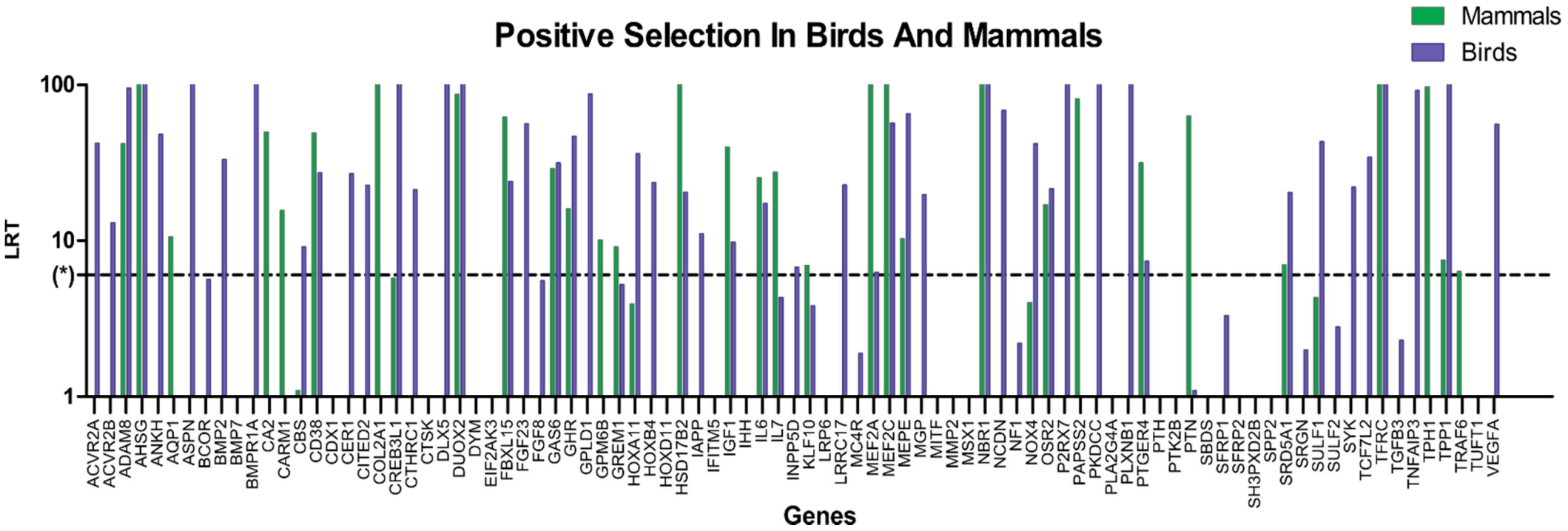

Figure 3-3. Positive selection in bird and mammal bone-associated genes. Blue bars represent the Likelihood ratio-rest between the model 2 vs 1 in log-scale for each gene. Blue bars for birds and green bars for mammals. Values below 1 (p-value ≥ 0.6) and above 100 (p-value << 0.01) are not represented on the graph. (*) represent the critical value in the χ^2 distribution for a p-value=0.05 (~5.99).

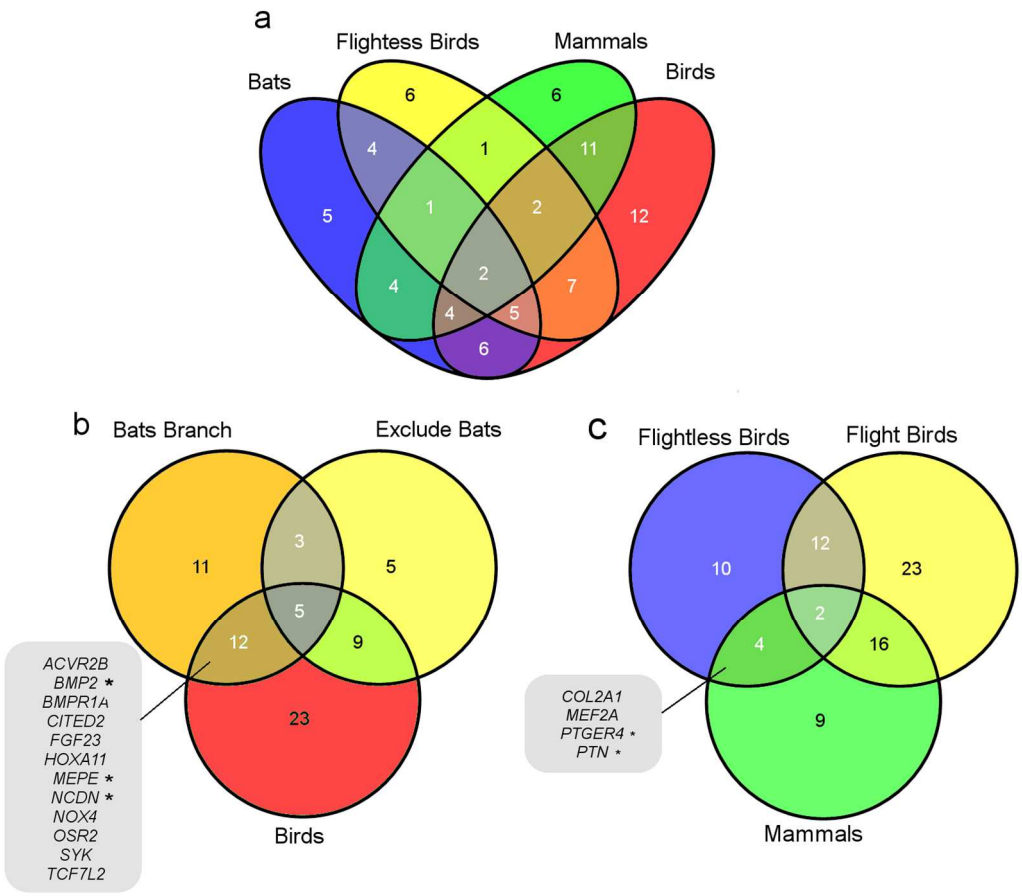

Figure 3-4. Venn diagram of the positively-selected bone-associated genes. a) Interception between the positively selected genes in mammals, birds and the same dataset including only terrestrial mammals and birds with flight ability. b) Interception between positively selected genes in mammals, birds and those genes showing a different evolutionary rate in mammals (bats) and in birds (flightless species). c) Positively selected genes in birds and terrestrial mammals (excluding bats) and those showing a different evolutionary rate in bats (bat branch). d) Interception between positively-selected genes in mammals, flying, and flightless. Asterisks (*) represent genes where the foreground branch was slower than background.

In birds the highest global omega values (0.53 and 0.71) were observed for *AHSG* (Alpha-2-HS-glycoprotein) and *P2RX7* (P2X purinoceptor 7), respectively (Appendices III: Table S5). Both of these genes are associated with bone mineral density and bone remodeling (Yang, Wang et al. 2007; Jorgensen, Husted et al. 2012). However, considering only the number of sites with omega > 1.0 and Posterior Probability (PP) ≥ 0.95, two genes involved in bone reabsorption, *TPP1* (Tripeptidyl peptidase I) and *TFRC* (Transferrin Receptor), had the highest number of positively selected sites, 95 and 33, respectively, corresponding to 19.8% and 4.2% of the alignment length (Appendices III: Table S5). In the zebra finch, five of the positively-selected *TPP1* residues occurred at functional/active sites (98S, 174P, 253T, 258V and 259A; Figure 3-1C), which correspond to human homologous positions (127R, 202P, 278Q, 286N and 287I (information retrieved from Uniprot) (Wu, Apweiler

et al. 2006). Interestingly, the positively selected residues with a PP ≥ 0.95 are mainly in *Peptidase S53* (from residue 171 to 368 by homology inference of the human sequence) (59 out of 95 residues, ~62%) (Figure 3-1C). Since *TPP1* is secreted by osteoclasts and *Peptidase S53* is involved in bone collagen proteolysis (Page, Fuller et al. 1993), the positive selection may be related with the optimization of this proteolytic process during bone resorption.

Impact of flight evolution in the evolutionary rate of bone-associated genes

When bats are removed from the mammalian dataset only 22 of 86 genes (three MSA's were absent any Chiroptera representative) showed signatures of positive selection (~26%), 10 of the 31 genes were no longer positively-selected and one new gene was added (Appendices III: Table S6). This reduction is suggestive of a different evolutionary rate in Chiroptera relative to the other mammals, likely due to the impact of flight in bats bones. In contrast, in birds the removal of species with reduced ability or inability to fly, [*Aptenodytes forsteri* (Emperor Penguin), *Gallus gallus* (Chicken), *Meleagris gallopavo* (Turkey), *Pygoscelis adeliae* (Adelie Penguin), *Struthio camelus* (Ostrich) and *Tinamus major* (Great Tinamou) (Figure 3-5)] resulted in a strong bias in the positively selected genes. Since flightless birds removal from the MSA resulted in eleven additional genes, and for seven genes were no longer significant, total of 53 genes (~58.4%) (Appendices III: Table S7).

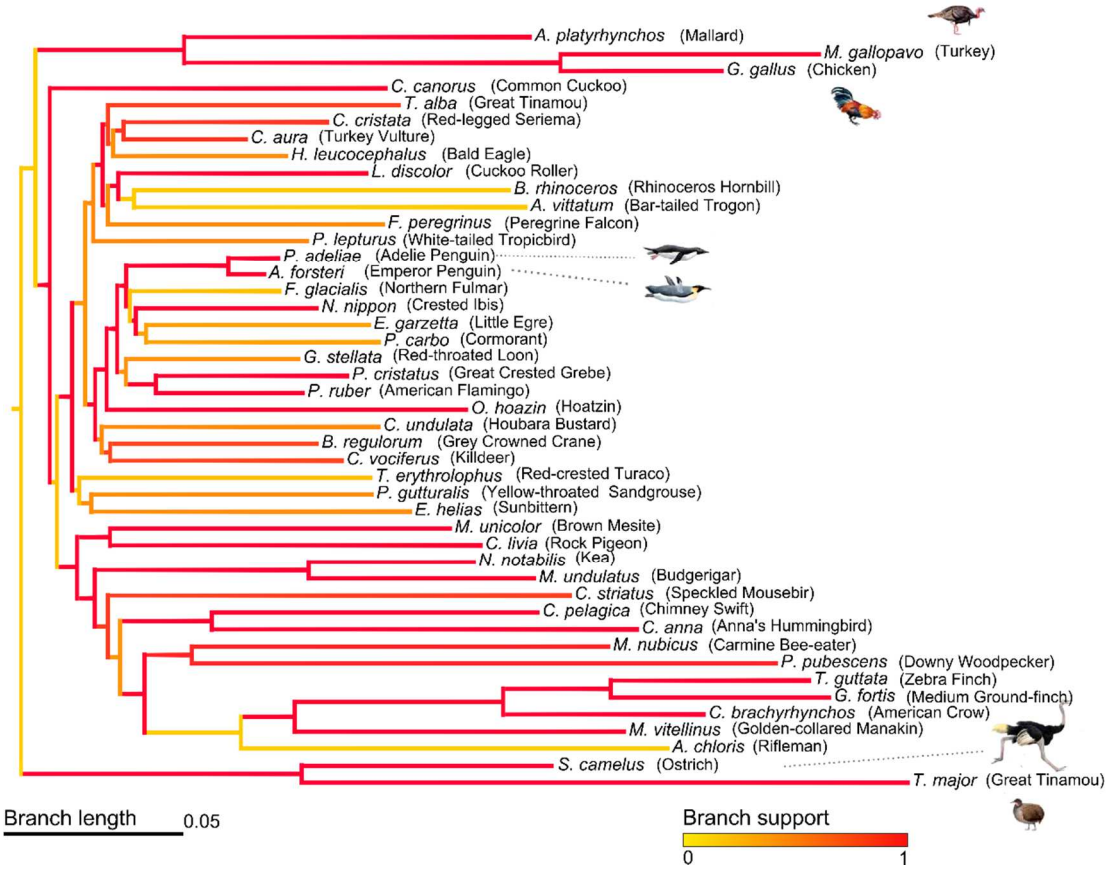

Figure 3-5. The gene-tree-based phylogeny from concatenation analysis of 89 genes in 45 bird genomes using maximum likelihood. The species with images are flightless birds. Each branch is colored according to the branch support value obtained. The species *Haliaeetus leucocephalus* (Bald Eagle) and *Pelecanus crispus* (Dalmatian Pelican) have been excluded from the analysis given the low number of retrieved sequences (n<=5).

Branch models shows increased selection in bone genes of flying species

For the branch model analyses, the datasets were labeled accordingly to their life-habit. This approach permitted the identification of genes evolving under a different evolutionary rates in the different lineages flightless and flying species. Flightless birds (Figure 3-5) were defined as those unable to sustain flight for long distances (such as Turkey or Chicken). All mammals all were considered to be flightless except for the little brown bat (*Myotis lucifugus*) and the flying fox (*Pteropus vampyrus*) (Figure 3-6).

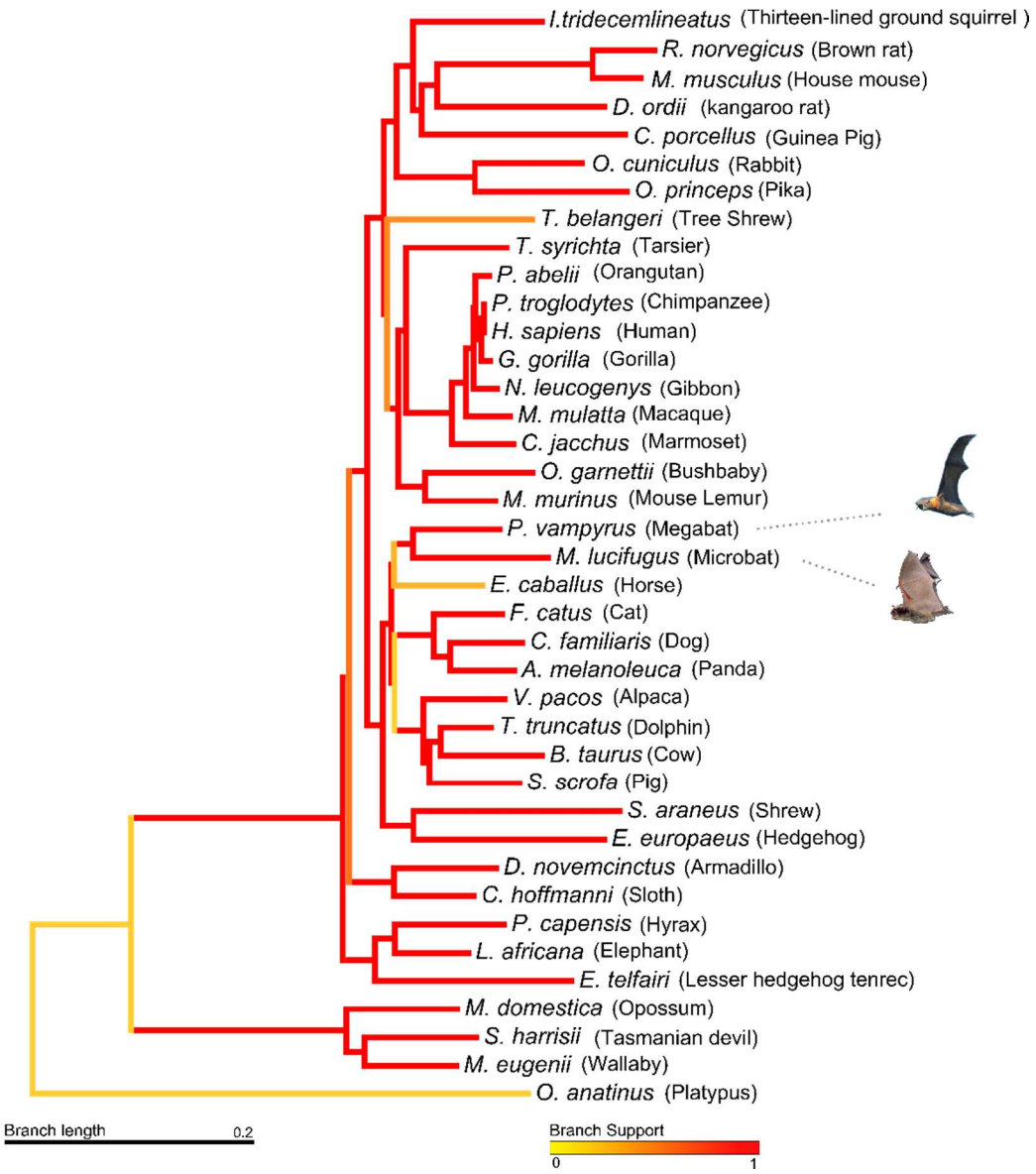

Figure 3-6. The gene-tree-based reconstruction phylogeny recovered from concatenation analysis of 89 genes in 39 mammalian genomes using maximum likelihood. The species with images represent the mammalian species with powered flight. Each branch is colored according to their support value.

The correlation between mammals and birds had the lowest rho (ρ) value for flightless birds and flying mammals (Spearman's ρ=0.579; p-value<0.01) (Table 3-1). The highest similarities in the observed ω values were obtained within each taxonomic clade; for bats and other mammals ρ=0.833 (p-value <0.01) and for flightless and flying birds ρ=0.883 (p-value <0.01). These patterns suggest that although a relatively small number of sites were affected, they were sufficient to be detected as evolving under positive selection, yet were insufficient to result in a significant different evolutionary rate between flying and flightless species.

Table 3-1. Spearman correlations between the estimated ω for branches: Flight vs Non-Flight Birds and Other Mammals vs Bats.

Spearman Correlations[b]				
	Flying Birds	Flightless Birds	Bats	Others Mammals
Flying Birds	-	.883**	.605**	.717**
Flightless Birds		-	.579**	.668**
Bats			-	.833**
Others Mammals				-

**. Correlation is significant at the p<0.01 (2-tailed). b. Listwise N = 85

Flying species shows prevalence of positive selection in bone regulatory genes

In birds the gene set enrichment analysis showed that, in contrast with the results from entire dataset, that the genes under positive selection are mainly involved in processes such as bone selection regulate ossification, bone mineralization, and biomineral formation (Table 3-2). Remarkably, when positively-selected genes in all birds are crossed with those evolving under a different evolutionary rate in bats and after excluding terrestrial mammals, a similar pattern is revealed, with a prevalence of positive selection in regulatory and remodeling genes. Notably, 12 genes that were positively selected in birds also had different evolutionary rates in bats compared with the other mammals. Since these genes showed a different evolutionary rate in bats relatively to terrestrial mammals, while showed no evidence of positive selection in mammals broadly (Figure 3-4B).

Table 3-2. Gene set enrichment of gene lists.

Gene List	Term	Count	p-value	Bonferroni Correction
Entire Gene List	GO:0001501:skeletal system development	33	2.2E-31	3.5E-28
	GO:0030278:regulation of ossification	19	2.1E-24	3.3E-21
	GO:0060348:bone development	21	8.7E-24	1.4E-20
	GO:0030500:regulation of bone mineralization	14	5.1E-22	8.2E-19
	GO:0070167:regulation of biomineral formation	14	1.5E-21	2.4E-18
	GO:0001503:ossification	17	5.2E-18	8.5E-15
	GO:0051094:positive regulation of developmental process	20	2.3E-15	3.6E-12
	GO:0045124:regulation of bone resorption	9	7.2E-15	1.2E-11
	GO:0046850:regulation of bone remodeling	9	7.2E-15	1.2E-11
	GO:0045597:positive regulation of cell differentiation	18	2.1E-14	3.4E-11
Birds (All)	GO:0030278:regulation of ossification	13	1.0E-17	1.2E-14
	GO:0030500:regulation of bone mineralization	10	2.3E-16	2.7E-13
	GO:0070167:regulation of biomineral formation	10	4.5E-16	5.3E-13
	GO:0001501:skeletal system development	15	7.5E-13	9.0E-10
	GO:0045124:regulation of bone resorption	6	8.4E-10	1.0E-06
	GO:0046850:regulation of bone remodeling	6	8.4E-10	1.0E-06
	GO:0034103:regulation of tissue remodeling	6	3.2E-09	3.9E-06

	GO:0045667:regulation of osteoblast differentiation	7	4.6E-09	5.6E-06
	GO:0045779:negative regulation of bone resorption	5	6.8E-09	8.1E-06
	GO:0046851:negative regulation of bone remodeling	5	6.8E-09	8.1E-06
Mammals (All)	GO:0001501:skeletal system development	10	9.4E-09	7.5E-06
	GO:0060348:bone development	7	1.6E-07	1.3E-04
	GO:0001503:ossification	6	3.4E-06	2.7E-03
	GO:0005615:extracellular space	9	1.6E-05	1.1E-03
	GO:0044421:extracellular region part	10	2.3E-05	1.7E-03
	GO:0010033:response to organic substance	9	7.5E-05	5.8E-02
	GO:0051240:positive regulation of multicellular organismal process	6	1.3E-04	9.7E-02
	GO:0042981:regulation of apoptosis	9	1.6E-04	1.2E-01
	GO:0043067:regulation of programmed cell death	9	1.7E-04	1.3E-01
	GO:0010941:regulation of cell death	9	1.8E-04	1.3E-01
(Flightless Birds Branches ∩ Mammals) \ Flight Birds	GO:0001503:ossification	2	2.5E-02	8.9E-01
	GO:0060348:bone development	2	2.7E-02	9.1E-01
	GO:0001501:skeletal system development	2	6.9E-02	1.0E+00
(Bats branches ∩ Birds All) \ Terrestrial Mammals	GO:0030500:regulation of bone mineralization	5	6.4E-09	3.4E-06
	GO:0070167:regulation of biomineral formation	5	8.4E-09	4.5E-06
	GO:0030278:regulation of ossification	5	3.3E-07	1.7E-04
	GO:0001501:skeletal system development	6	2.9E-06	1.5E-03
	GO:0045667:regulation of osteoblast differentiation	4	4.9E-06	2.6E-03
	GO:0045597:positive regulation of cell differentiation	5	2.4E-05	1.3E-02
	GO:0051094:positive regulation of developmental process	5	5.1E-05	2.7E-02
	GO:0009967:positive regulation of signal transduction	5	6.5E-05	3.4E-02
	GO:0048598:embryonic morphogenesis	5	7.6E-05	3.9E-02
	GO:0030501:positive regulation of bone mineralization	3	9.1E-05	4.7E-02

In mammals the positively selected bone genes are mainly involved with the skeletal system, bone development and ossification (Table 3-2). Interestingly, the group of genes under different evolutionary rates in flightless birds than mammals and excluding those positively selected in flying birds (Table 3-2), resulted in a list of genes that were similar to those observed in mammals, i.e. mainly developmental genes. This included four genes that had different rates in flightless birds and were positively selected in all mammals, but not positively selected in birds with powered flight (Figure 3-4C) and included genes that have been linked with ossification, bone development, and skeletal system development.

Testing gene randomness effects: bone related genes, brain developmental genes and random dataset

To control the possibility of that the observed patterns in our bone related genes occurred at random we evaluated an additional dataset, consisting of 50 genes involved in brain development in birds and mammals. The number of positively selected brain genes in birds was 22 (~44%) and excluding the flightless species the number decreased to 18 (~36%). Similarly, in mammals, when bats were removed, the number of positively selected genes decreased from 16 (~32%) to 13 (~26%). The overlap of positively selected genes in birds evolving under a different evolutionary rate in bats and after excluding terrestrial mammals, included 5 (10%) positively selected genes, similar to the 12 (~13.5%) observed in bone-associated genes (Appendices III: Table S8).

Additionally was performed the same analyses for a dataset of 50 randomly chosen genes. The results show a lower number of positively selected genes in birds 21 (~42%), and after excluding flightless species the number of positively selected genes is reduced to 19 (~38%). Similarly, in mammals, when bats were removed, the number of positively selected genes decreased from 18 (~36%) to 12 (~24%). Using the same strategy as above, positively selected genes in birds and evolving under a different evolutionary rate in bats, the list is of only 2 (10%) (Appendices III: Table S9).

Correlation between substitution rates and body mass

To determine relationships between evolution rates in flying species and body mass, we used the Bayesian method CoEvol which provided comparisons between rates of change in phenotypic traits and rates of molecular evolution (Lartillot and Poujol 2011). In CoEvol, a high posterior-probability association covariance between the rate of change in d_S, ω, GC, and the change of a phenotypic trait is evidence of a link between molecular and phenotypic processes. The separate estimation of covariance for d_S and ω distinguishes mutational effects of d_S from selective effects of ω. We generated two bird datasets, one including all the available sequences from flight and flightless species, and a second with only species capable of a powered flight. The reduced sample size of flightless species (n=6) is insufficient to an accurate analysis of this dataset. Therefore, comparison between all birds vs. only flying birds was used to help understand the effect in the model estimation when flightless birds were included. A similar approach was employed for mammals, using a dataset including all mammals compared with other sets using only terrestrial mammals. When only flight-enabled species were tested, a nearly-negative covariance was found between d_S and average body mass (R= -0.332, pp (posterior probability) =0.087;Table 3-3). When flightless species were included in the covarion model, in addition to the d_S correlation

with body mass (R= -0.5065, pp=0.0245), there was also a negative covariance between GC content and body mass (R= -0.5855, pp=0.026), and a positive correlation between ω and the body mass (R= 0.546, pp=0.985) (Appendices III: Figure S2).

Table 3-3. Covariance between d_S, ω, gc content, and the three body mass measures (minimum, maximum and average) in 45 bird genomes. The upper triangular shows the posterior probability, (** - <= 0.025 or >=0.975; * - <=0.05 or >=0.95). The covariance between the estimates are on the lower triangular. Bold values are from analyses off all birds and values between brackets are from analyses excluding flightless species.

		Posterior Probability (PP)					
		d_S	ω	gc	Minimum Weight	Maximum Weight	Average Weight
Covariance	d_S	-	**0.17** (0.0104)	**0.855** (0.955*)	**0.0225*** (0.1075)	**0.0235**** (0.085)	**0.0245**** (0.087)
	ω	**-0.3355** (-0.538)	-	**0*** (0.00605)	**0.975**** (0.725)	**0.98**** (0.745)	**0.985**** (0.74)
	gc	**0.387** (0.506)	**-0.872** (-0.6615)	-	**0.029*** (0.29)	**0.0275*** (0.25)	**0.026*** (0.255)
	Minimum Weight	**-0.511** (-0.308)	**0.5355** (0.18)	**-0.571** (-0.192)	-	**1**** (1**)	**1**** (1**)
	Maximum Weight	**-0.5045** (-.3385)	**0.5395** (0.201)	**-0.5815** (-0.2375)	**0.985** (0.97)	-	**1**** (1**)
	Average Weight	**-0.5065** (-0.332)	**0.546** (0.199)	**-0.5855** (-0.227)	**0.993** (0.9865)	**0.9985** (0.9965)	-

Table 3-4. Covariance between d_S, ω, gc content, and the three weight measures (minimum, maximum and average) in 39 mammal genomes. . The upper triangular shows the posterior probability (** - <= 0.025 or >=0.975; * - <=0.05 or >=0.95). The covariance between the estimates are on the lower triangular. Bold represent values for all mammals while between brackets values after excluding bats.

		Posterior Probability (PP)					
		d_S	ω	gc	Minimum Weight	Maximum Weight	Average Weight
Covariance	d_S	-	**0.0245**** (0.0925)	**0.940** (0.800)	**0.005**** (0.014**)	**0.0164**** (0.03*)	**0.011**** (0.023**)
	ω	**-0.5145** (-0.4015)	-	**0.0176**** (0.001**)	**0.980*** (0.99**)	**0.970*** (0.99**)	**0.975**** (0.99**)
	gc	**0.3075** (0.2065)	**-0.473** (-0.5795)	-	**0.050*** (0.042*)	**0.0345*** (0.029*)	**0.0365*** (0.028*)
	Minimum Weight	**-0.5435** (-0.523)	**0.4795** (0.5305)	**-0.413** (-0.3935)	-	**1**** (1**)	**1**** (1**)
	Maximum Weight	**-0.478** (-0.454)	**0.467** (0.527)	**-0.455** (-0.44)	**0.956** (0.95)	-	**1**** (1**)
	Average Weight	**-0.5005** (-0.477)	**0.4815** (0.543)	**-0.4535** (-0.439)	**0.970** (0.9655)	**0.9975** (0.997)	-

Mammals exhibited a different trend. When bats were included, the data suggested that there was a negative correlation between body mass and d_S (R=-0.542, pp=0.001**) and body mass with GC content (R=-0.408, pp=0.048*) and that there was a positive correlation with body mass and ω (R=0.578, pp=0.001**) (Table 3-4). In contrast, when bats were excluded, only ω (R= 0.543, pp=0.99) and d_S (R= -0.477 pp=0.011**) were significantly correlated with average body mass. In contrast, when bats were included, the data suggested there was a negative correlation between body mass and d_S (R=-0.542, pp=0.001**)

and body mass with GC content (R=-0.408, pp=0.048*) and that there was a positive correlation with body mass and ω (R=0.578, pp=0.001**) (Table 3-4). These findings suggest that including or excluding bats has a significant effect on the results, lending greater support for the selection of bone genes involved in flight. This can be partially explained by the relatively small number of bats in the dataset (~5% of the total amount of sequences) compared with the larger percentage of flightless species (~13%) in the avian comparison. Additionally, the large flying fox is often reported as the largest bat, and therefore potentially introduces a slight bias in the analyses given his large body mass.

3.5 Discussion

We assessed the evolutionary patterns of 89 bone-related genes in 47 avian and 39 mammalian genomes, and demonstrate that there has been significantly higher positive selective pressure on several of the bone-associated genes of birds, particularly in genes involved in regulatory processes. In contrast, mammals had higher positive selection in bone-development genes. Moreover, just as in birds, flying mammals (bats) had several genes with evolutionary rates that contrasted with the patterns observed in other mammals. These results highlight convergent changes in bone genes in the evolution of flight.

Body Mass and Bone-Associated Genes

The different evolutionary trajectories of developing the capacity to fly in birds and bats led to some distinct mechanical and biochemical solutions to the adaptive challenges. Nevertheless, both birds and bats have bones with high mineral content (Dumont 2010) and their body sizes approach the predicted theoretical limit, i.e, the tradeoff between the mechanical power and the capacity for metabolic output essential for flight (Maina 2000). Among different avian orders the skeletal measurements and body mass are correlated and as they are limited by ecological and biomechanical constraints on bone dimensions (Field, Lynner et al. 2013). The different life habits among birds partially explains the higher correlation between body mass and ω that was observed when assessing the dataset including all the bird species. Since this covariance suggests a relaxation on the selective pressure on bone-associated genes in non-flying species, the findings are consistent with the hypothesis that the skeleton of flightless birds can be larger than in flying birds. The absence of this correlation within flying species may reflect the lower variation in the body mass of flying species, and differences on the foraging habits irrespective to their body size, since bone structure is often associated with the life habit of the species (Smith 2012). In contrast, extant mammals display a wider range of body mass than extant birds (Bouzat 2000), sup-

porting the observed correlation between ω and average body mass. When the representative species of Chiroptera (n=2) are included, the estimation between the phenotypic characters and bone-associated genes is similar. Therefore, additional sequences could provide a deeper understand of the evolution of bone-associated genes and the body mass correlation in bats.

Evolutionary rate in flying versus non-flying species

Genomic sequences are expected to uncover selective processes that species have undergone through time, including the adaptation to new ecological niches. Although vertebrate powered flight is not restricted to birds (e.g. Chiroptera in mammals, Appendices III: Figure S1), flight in most ubiquitous in birds. Despite powered flight has been linked with low body mass (Schmidt-Wellenburg, Engel et al. 2008), high metabolic rate (Munshi-South and Wilkinson 2010), metabolic efficiency (Morris, Nelson et al. 2010), and specialized mechanical systems supported by skeletal adaptations, many aspects of flight remain unclear, including how bone-related genes evolved in birds and other taxonomic groups such as bats. The high rates of selection that we found for several bone-related genes suggest that the observed variation among avian species is higher than would be expected under models of neutrality. Therefore, signatures of adaptive and positive selection in these genes are likely a fundamental feature of trait modeling in the evolution of the skeleton. Moreover, the number of positively selected bone-associated genes in birds is higher and even more pronounced when flightless birds are removed from the analyses.

Although purifying selection is unable to “produce” good genes, radical and extensive changes in genes are not very common. Instead, only a few sites are generally involved with detectable signatures of positive selection. Consistent with this expectation, no genes were observed with strong evidence of selection (ω>1), and the differences among flying and flightless species were within 0.34 (the difference between foreground and background branches). This similarity in the molecular data between flightless vs flying species is supported by our findings that: 1) differences in the selection pressure among flight and flightless birds were insufficient to produce general or radical differences: 2) the acquisition of flight only strongly affected a few bone-associated genes in bats relative to other mammals: and 3) only a few sites were affected and there were no extensive changes within these genes.

Extensive selection on bone-associated genes compared to other datasets

The analysis of two additional datasets, random and brain developmental genes shows a prevalence of positive selection in bone-associated genes, since they have higher

proportion of positively selected genes and the impact of the removal of flying species is higher in bone-associated gene analyses. This is suggestive of a direct impact of flight in bone-associated genes, and therefore the evolution of those genes is likely intimately associated with the arousal of vertebrate's flight.

Flight extended impact in bone-associated genes

Our results strongly suggest that a relatively small number of genes involved in bone structures may have independently evolved in birds and bats in similarly ways that permitted the transition from terrestrial to aerial life styles (see Appendices III: Figure S1). Of the 89 bone-associated genes, only 12 showed signatures of selection in both birds (site models selection) and bats (branch model exhibiting acceleration/deceleration relatively to terrestrial mammals with significant statistical support). These genes, summarized below, therefore probably reflect key genetic pathways and adaptations enabling flight. And since several bone-associated genes are involved in other processes, therefore the comparison between flying and non-flying species provides evidence of the impact of flight on their evolution on bone structure but may also be associated with other processes where those genes are involved (Figure 3-1D).

BMP2 (Bone morphogenetic protein 2) has been implicated in the stimulation of cartilage proliferation and differentiation and in the increase in digit length in bat embryonic forelimbs (Sears, Behringer et al. 2006). The lengthening of forelimbs was an essential step in the evolution of flight in vertebrates (Padian and Chiappe 1998; Dececchi and Larsson 2013). In addition, birds share several other features, including a fused cranial bone, which might be linked with *BMP2* (Chen, Zhao et al. 2004). Importantly, several other examples of bone fusion (e.g. vertebrae fusion) have also been cited as being crucial for the evolution of flight (Cubo and Casinos 2000).

OSR2 (odd-skipped related 2) has been associated with forelimb, hindlimb and craniofacial development (Lan, Kingsley et al. 2001) and is a likely candidate gene for many of the fundamental changes in the limbs of birds and bats. At the beginning of avian evolution, the allometric coupling of forelimb and hindlimb with body size was disrupted, and as wings began to significantly elongate, they maintained a positive allometric relationship with body size, but their legs significantly shortened (Dececchi and Larsson 2013). This would have facilitated the diversification of forelimb and hindlimb shapes and sizes that are currently observed in extant birds (Dececchi and Larsson 2013) and which are closely linked with foraging habits in birds and bats (Dececchi and Larsson 2013).

HOXA11 (homeobox A11) may also be related with bone fusion, as this gene has been reported to influence radio-ulnar fusion (Thompson and Nguyen 2000) and bats may

also display partial fusion of those bones (see Figure 3-1C). Although birds presented no signs fusion of the radio and ulna, these bones are typically apneumatic in birds and therefore contain bone marrow; and *HOXA11* has been associated with bone marrow failure syndrome (Thompson and Nguyen 2000).

FGF23 (fibroblast growth factor 23), *MEPE* (matrix extracellular phosphoglycoprotein), *NCDN* (neurochondrin), *NOX4* (NADPH oxidase 4) are involved in bone metabolism (Dateki, Horii et al. 2005; Fukumoto 2008; Rowe 2012; Goettsch, Babelova et al. 2013). Bone metabolism genes are often associated with alterations of Bone Mineral Density (BMD) (Alexopoulou, Jamart et al. 2006), and BMD alterations in birds and bats have previously been linked with flight adaptations.

BMPR1A (bone morphogenetic protein type IA gene) is involved in bone remodeling, and the ablation of this receptor in osteoblasts increases bone mass (Baud'huin, Solban et al. 2012). This makes *BMPR1A* a prime candidate for the maintenance of bone strength, which is essential for a stiff, but lightweight skeleton system in flying species (Dumont 2010). Similarly, *ACVR2B* (activin receptor type-2B) is involved in the control of bone mass, but interestingly is mediated by *GDF-8* (myostatin) which is also involved in improving muscle strength (Hamrick 2010).

CITED2 (Cbp/P300-interacting transactivator, with Glu/Asp-rich carboxy-terminal) is involved in bone formation (Lee, Taub et al. 2009) but also plays a pivotal role in muscle mass regulation since it also counteracts glucocorticoid-induced muscle atrophy (Tobimatsu, Noguchi et al. 2009). This makes it a prime candidate gene, as flight in vertebrates requires powerful muscles, particularly those connected to sternum bones (Olson and Feduccia 1979). *CITED2* is also involved in some heart diseases (Li, Pan et al. 2012), which may be of relevance since birds (Grubb 1983) and small bats (Canals, Atala et al. 2005) possess larger hearts compared with vertebrates of similar size.

SYK (spleen tyrosine kinase) encodes a multifunctional protein involved in several processes, including osteoclastogenesis, and is therefore closely linked with bone destruction (Liao, Hsu et al. 2013). Interestingly this gene is also involved in the production of reactive oxygen species (ROS) and the immune system (Bae, Lee et al. 2009). Both birds and bats have developed multiple mechanisms for resisting oxidative damage, offering a plausible explanation to their longer life expectancy when compared with non-flying mammals of similar body size (Munshi-South and Wilkinson 2010). Although the association between lifespan and ROS production is an ongoing debate (Montgomery, Hulbert et al. 2012), birds and bats tend to produce smaller amounts of ROS (Barja 1998; Brunet-Rossinni 2004) and recently signs of selection were found in multiple genes involved in repairing DNA damage that is triggered in bats when flying (Zhang, Cowled et al. 2013).

TCF7L2 (Transcription factor 7-like 2) is associated with bone mineralization (Friedman, Oyserman et al. 2009), but is also considered the most significant genetic marker linked with Diabetes mellitus Type 2 risk and is a key regulator of glucose metabolism (Vaquero, Ferreira et al. 2012). The signatures of selection observed in birds and bats in *TCF7L* are remarkable given the high blood glucose levels observed in birds (Szwergold and Miller 2013), fruit and nectar-feeding bats (Kelm, Simon et al. 2011; Shen, Han et al. 2012). The tolerance of birds and bats to blood-hyperglycemia may therefore be reflected in the evidence for positive selection observed in our analyses, as flight requires efficient glucose metabolism and efficient transportation to the energy-demanding organs (e.g. flight muscles) that is essential to powered flight (Shen, Liang et al. 2010; Shen, Han et al. 2012).

Despite the similarities between bats and birds (Figure 3-1D), extensive positive selection is observed in some genes in birds that are not apparent in bats. Notably, *P2RX7* and *TPP1,* which are mainly involved in bone resorption (Page, Fuller et al. 1993; Armstrong, Pereverzev et al. 2009) showing a high prevalence of positive selection only in birds. In birds, the pneumatic epithelium that forms the diverticula is capable of extensive resorption of bone material given it's close association with osteoclasts (Witmer 1997). Bone remodeling, particularly the resorption, may be crucial in the formation of the bone trabeculae and by extension, the formation of the pneumatic bones. Recently, polymorphisms described in *P2RX7* have been associated with osteoporosis in humans (Wesselius, Bours et al. 2013), which is typically linked with increased bone resorption and a decrease in bone mineral density (BMD) (Riggs 2000). Here we demonstrated that genes involved in bone remodeling (particularly evident in the sub-process bone resorption) had multiple signals of strong positive selection in birds, but contrary to osteoporosis, bird bones attain a high value of BMD (Dumont 2010).

Gene set enrichment, bone remodeling and their implication in life-habits

Pneumatization preceded the origin of avian flight evolved independently in several groups of bird-line archosaurs (ornithodirans) (Benson, Butler et al. 2012), and therefore cannot be the result of adaptation for flight (Benson, Butler et al. 2012). It has been suggested that skeletal pneumaticity, in early evolutionary stages, provided no selective advantage (Wedel and Taylor 2013) and also did not significantly affect the skeleton through the lightening or remodeling of individual bones (Wedel and Taylor 2013). Although skeletal density modulation would have resulted in energetic savings as part of a multi-system response to increased metabolic demands and the acquisition of an extensive postcranial skeleton, pneumaticity may have favored a high-performance endothermy (Benson, Butler et al. 2012). Although bone pneumaticity may have facilitated the transition to flight in birds,

it may not have been a necessary step, since bats evolved the ability to fly without postcranial skeletal pneumaticity.

Genes involved in bone remodeling have been subjected to a higher prevalence of positive selection in birds. This finding is interesting since: 1) development of postcranial skeletal pneumaticity occurs after hatching (Hogg 1984); 2) the skeleton is a metabolically active tissue that undergoes continuous remodeling throughout life (Hadjidakis and Androulakis 2006); and 3) bone remodeling may lead to a more porous bone structure (Taylor, Horvat-Gordon et al. 2013). Bone remodeling involves the removal of mineralized bone by osteoclasts followed by the formation of a bone matrix through the osteoblasts that is subsequently mineralized (Hadjidakis and Androulakis 2006). It is generally assumed that bone remodeling is essential for maintaining skeletal mechanical properties and mineral homeostasis (Parfitt 2002). Therefore the higher prevalence of positive selection in bone-remodeling genes suggests that bones with higher mineral density were attained as a response to the selective contingencies imposed by flying, including bone remodeling and bone resorption. The similarities in bats and flying birds, showing bones with high mineral content, the genes involved in bone remodeling probably play a pivotal part of avian diversification and adaptation to a wide variety of ecological and behavioral niches.

The evolution of flight in birds and bats was a pivotal event for their successful adaptation into new ecological niches. However, the transition to flight imposed new challenges on their bone structure. The high rate of positive selection in bone-associated genes in birds suggests that there was a strong link among changes in these genes and the adaptations necessary for flight. Limitations imposed on body size were probably also a key factor in bird evolution, as we have shown here that body mass covaried significantly with the omega value only when flightless birds were included. Similarly, the adaptation of bats to flight was associated with acceleration/deceleration of the evolutionary rate in several bone-associated genes relative to others mammals. Evidence of adaptive selection in birds and bats also were apparent in genes plausibly linked with bone-remodeling, bone fusion, lengthening of forelimbs, as well as with functions outside the skeleton system, including ROS production and glucose tolerance that also would have had a major influence on the capacity for powered flight. However, the examples of positive selection that were only observed in birds, such as the evolution of a more-diversified and richer-variety of protein-encoding genes involved in bone resorption (e.g. *TPP1* and *P2RX7)* and the formation of bone trabeculae that are likely critical to the evolution of hollow or pneumatic bones*,* suggest that these might be crucial steps in the evolution of avian flight that are unique to them.

3.6 Conclusions

The evolution of flight in birds and bats was a pivotal event for their successful adaptation into new ecological niches. However, the transition to flight imposed new challenges on their bone structure. The high rate of positive selection in bone-associated genes in birds suggests that there was a strong link among changes in these genes and the adaptations necessary for flight. Limitations imposed on body size were probably also a key factor in bird evolution, as we have shown here that body mass covaried significantly with the omega value only when flightless birds were included. Similarly, the adaptation of bats to flight was associated with acceleration/deceleration of the evolutionary rate in several bone-associated genes relative to others mammals. Evidence of adaptive selection in birds and bats also were apparent in genes plausibly linked with bone-remodeling, bone fusion, lengthening of forelimbs, as well as with functions outside the skeleton system, including ROS production and glucose tolerance that also would have had a major influence on the capacity for powered flight. However, the examples of positive selection that were only observed in birds, such as the evolution of a more-diversified and richer-variety of protein-encoding genes involved in bone resorption (e.g. *TPP1* and *P2RX7)* and the formation of bone trabeculae that are likely critical to the evolution of hollow or pneumatic bones, suggest that these might be crucial steps in the evolution of avian flight that are unique to them.

3.7 Acknowledgements

JPM was funded by the PhD grant SFRH/BD/65245/2009 from the Portuguese “Fundação para a Ciência e a Tecnologia” (FCT) and by a grant from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism. AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013 and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490).

Chapter 4 – *Role of positive selection and recent gene duplication on generation of novelty in Mammalian dentition patterns*

4.1 Abstract

A wide number of genes are involved in tooth development in vertebrates. Several studies, focused mainly in mice and rats, have provided an in depth depiction of the processes coordinating tooth formation and shape. Here we surveyed 236 tooth-associated genes in 39 mammalian genomes, testing for signatures of selection signatures to assess patterns of molecular adaptation in genes regulating mammalian dentition. Of the 236 genes, 36 showed strong signatures of positive selection that may be responsible for the phenotypic diversity observed in the mammalian dentition. Mammal-specific tooth-associated genes had accelerated mutation rates compared with older genes found across vertebrates, suggesting that these relatively new-evolved genes might be involved in some of the more-recent differences in dental patterns observed among mammals. Our results showed that more recent genes have fewer interactions (genetic and physical), are involved in fewer Gene Ontology terms and have relatively faster evolutionary rates. Concordantly, the introns of these positively-selected genes also exhibited accelerated mutation rates, which may reflect additional adaptive pressure in the intronic regions associated with regulatory processes influencing tooth-gene networks. Mammalian dentition is coordinated by at least 236 genes, with around ~15% of those genes showing strong signatures of positive selection and being involved in process like mineralization and structural organization of tooth specific tissues such as enamel and dentin. Moreover, 12 mammalian-specific genes (younger genes) provide insights on the diversification of mammalian teeth as they have a higher evolutionary rates and different expression profiles compared with those found across vertebrates.

4.2 Introduction

As a major determinant of vertebrate ecology, teeth have a crucial role in species survival. Tooth development has been subjected to strong selective constraints since they first appeared in the oral cavity over 460 million years ago (Mya) during the Ordovician (Smith and Coates 1998). While mammalian teeth share basic components, they exhibit great diversity in number, size and shape. However, in spite of their importance for animal survival, teeth have been lost independently in multiple lineages of tetrapods (Davit-Beal, Tucker et al. 2009), including mammals (e.g. pangolins). And some mammals have teeth without enamel (e.g. sloths), or both tooth and enamel reduction (e.g. platypus).

Mammals differ from other living vertebrates by having very complex teeth and a restricted capacity for tooth renewal (Jernvall and Thesleff 2012). Moreover, mammals show a strong correlation between their feeding habits, patterns of tooth formation (e.g.,

cardiform, villiform, incisor, canine, molariform) (Koussoulakou, Margaritis et al. 2009) and their number of teeth (Koussoulakou, Margaritis et al. 2009). While some non-mammals have multi-rowed dentition and replace their teeth regularly throughout their lifetime, mammals have only one row of teeth and either renew their teeth only once or without any replacement, as observed in some rodents (Jarvinen, Tummers et al. 2009; Koussoulakou, Margaritis et al. 2009; Mikkola 2009). Thus, vertebrate evolution is characterized by a reduction in the tooth number (from polyodonty to oligodonty), by a shift in timing of tooth development (from polyphyodonty to di- and/or monophyodonty) and by an increase in morphological complexity (from homodonty to heterodonty) (Salazar-Ciudad and Jernvall 2004). Furthermore, these mammalian features, including increased shape complexity, multi-cusp teeth, and stable tooth number facilitated the maintenance of the high metabolic rates of mammals by ensuring efficient processing of food (Armfield, Zheng et al. 2013).

Modern mammalian dentition develops through a series of well-defined morphological stages that require sequential and reciprocal interactions between the epithelium and mesenchyme (Mitsiadis and Graf 2009). In mice, the first sign of tooth development, the thickening of the oral epithelium, is observed at embryonic day 10.5 (E10.5) (Zhang, Chen et al. 2005; Mitsiadis and Graf 2009), when tooth sites and types are established (Zhang, Chen et al. 2005). Between embryonic days 12.5-13.5 (E12.5–E13.5) the tooth bud is progressively formed following the epithelium invagination of the underlying mesenchyme (Mina and Kollar 1987; Mitsiadis and Graf 2009). During days 14.5-15.5 (E14.5–E15.5) the growth of the epithelium leads to the formation of the cap structure (Mitsiadis and Graf 2009) and to its configuration during days 16.5-18.5(E16.5–E18.5) (Mitsiadis and Graf 2009). During the late bell stage, embryonic day (E18.5), mesenchyme cells form the dental follicle and dental pulp (Mitsiadis and Graf 2009). In spite of the wide phenotypic diversity among mammal dentition patterns, previous studies have demonstrated only slight differences in gene expression patterns, with human and mice teeth sharing considerable homology in ontogenesis and underlying molecular networks (Lin, Huang et al. 2007). The marked similarity between odontogenesis (in lamina, bud, cap, and bell stages) and gene expression profiles (Zhang, Chen et al. 2005) in mice and humans suggests that there are strong functional constraints in mammalian teeth development. The genetic control of tooth development encompasses, to-date, more than 300 genes (Thesleff 2006). However, this is probably an underestimate, since analyses of large datasets and new approaches using microarrays profile search functions have identified additional genes associated with odontogenesis (Kim, Lim et al. 2012; Landin, Shabestari et al. 2012).

The search for genes with evidence of positive selection is therefore likely to be an efficient way to identify nucleotide substitutions that are prime candidates for being associated with phenotypic divergence among species (Clark, Glanowski et al. 2003; Nielsen,

Bustamante et al. 2005; Kosiol, Vinar et al. 2008). In spite of some recent concerns about the use of low-coverage eutherian genomes in phylogenetic studies (Milinkovitch, Helaers et al. 2010) (Prosdocimi, Linard et al. 2012), recent studies focusing on 2x mammalian genomes have successfully identified amino acid residues that have undergone positive selection and that overlap with disease-associated variants of high relevance to understanding key processes in human biology, health and disease (Lindblad-Toh, Garber et al. 2011).

Genes involved in adaptation and functional innovation often show the footprints of positive selection through elevated ratios of non-synonymous to synonymous nucleotide substitutions (Yang and Bielawski 2000; Nielsen, Bustamante et al. 2005; Philip, Machado et al. 2012). However, measures of protein contribution to fitness may not always correlate well with evolutionary rate (Wang and Zhang 2009) and several measures of correlation have been proposed, including the number of mRNA molecules per cell (Green, Lipman et al. 1993), protein dispensability (Hirsh and Fraser 2001), the codon adaptation index (Wall, Hirsh et al. 2005), sequence length (Lipman, Souvorov et al. 2002), the number of interactions and estimates of solvent accessibility (Franzosa and Xia 2009). However, the best correlation between proteins of critical importance and evolutionary rate is expression level, since highly expressed proteins tend to evolve slowly (Krylov, Wolf et al. 2003; Subramanian and Kumar 2004). Although they are still limited by weak statistical power to discriminate between positive selection and neutral evolution, searches for selective signatures in genome-wide studies have provided important insights (Montoya-Burgos 2011). When positive selection acts only on a subset of codons, the best approach is to use “site models” implemented in the PAML package (Yang and Swanson 2002; Yang 2007) to identify functional units under differential selective pressures (Montoya-Burgos 2011) since positively-selected sites tend to cluster in the coding sequence (Clark, Eisen et al. 2007).

Here we performed comparative evolutionary analyses of tooth-related genes to identify signatures of selection that may have shaped tooth phenotypic diversity among mammals. Of the 236 tooth-associated genes analyzed in 39 mammalian genomes, we detected strong selection signatures in 36 genes using both gene and species trees. Moreover, younger genes (mammalian-specific) had accelerated evolutionary rates and differential expression profiles or expression in early stages of development compared with older genes (vertebrate-specific).

4.3 Methods

Sequences and annotation

Genes associated with tooth development, tooth disease and mammalian tooth phenotypes were retrieved from the Gene Ontology (GO) database (Ashburner, Ball et al. 2000; Harris, Clark et al. 2004) and the RGD database (Shimoyama, Smith et al. 2011; Laulederkind, Hayman et al. 2013). To restrict the gene dataset we only used the associated processes listed in the GO and Mammalian Phenotype (MP) (see Appendices IV: Table S1). The final dataset included 247 genes, from which 11 genes were excluded (because there were less than 20 available sequences). The 7,893 coding sequences used in this survey were retrieved from *ENSEMBL* v64 or v65 (Flicek, Amode et al. 2012) using the PyCOGENT 1.5.3 (Knight, Maxwell et al. 2007) implemented in EASER (Maldonado, Khan et al. 2013) querying *ENSEMBL compara* database. All the retrieved results were manually inspected and when the sequences could not be retrieved using the script, they were manually downloaded. The corresponding gene coordinates were obtained using Biomart in *ENSEMBL* to format the annotation file needed to build the ideogram in Idiographica (Kin and Ono 2007).

Alignment and data filtering

For each gene an alignment was built using the retrieved coding sequences translated to amino acids and further back-translated to nucleotides using MUSCLE (Edgar 2004) implemented in SEAVIEW (Gouy, Guindon et al. 2010). The multiple sequence alignment was refined in GBLOCKS (Castresana 2000) using the relaxed parameters (Talavera and Castresana 2007) to eliminate poorly aligned positions, reducing the false positives resulting from improper aligned positions. The filtered alignment was used to inspect possible evolutionary models using MrAIC (Nylander 2004) with the AICc correction (models were restricted to Bayes models to save calculation time). Phylogenetic tree gene-based reconstructions were obtained with PhyML under the corresponding evolutionary model estimated for each gene while branch supports were obtained using the aLRT test (Anisimova and Gascuel 2006). The tree topology was further used as the gene tree in evolutionary analyses after the removal of branches length, allowing CODEML to calculate each branch length during the likelihood estimation of each model. The final dataset incorporated 236 filtered alignments, obtaining an average of 33.44 sequences and a length around 2112.37 base pairs (bp) per multiple sequence alignment. The species tree topology was obtained

from *ENSEMBL* (Appendices IV: Figure S1) to determine if it was concordant with the accepted mammalian phylogeny (Meredith, Janecka et al. 2011). Trees were pruned, as necessary due to missing taxa, using Phyutility (Smith and Dunn 2008).

Evolutionary rate and protein age

For each gene the number of nonsynonymous substitutions per nonsynonymous site (d_N) and the number of synonymous substitutions per synonymous site (d_S) were calculated using a maximum-likelihood method (CODEML) implemented in PAML software package v4.6 (Yang 2007). Estimations of d_N, d_S and d_N/d_S, were obtained using six different models (Model 0, 1a, 2a, 7, 8 and 8a). Equilibrium codon frequencies of the model were used as free parameters (CodonFreq = 2). Model 0 (M0, one-ratio) was used to estimate global d_N/d_S, d_N and d_S, and for further correlation estimation. Model 1a (M1a, nearly neutral) distributes the sites in two site-classes varying between 0 and 1, assuming that all sites have $d_N/d_S \leq 1$. Model 2a (M2a, positive selection), unlike M1a, estimates the proportion of sites under positive selection, $d_N/d_S > 1$. Models 7 (M7, beta) and 8 (M8, beta+ω>1), approximate the d_N/d_S variation over sites through a beta distribution, estimating the proportion and the d_N/d_S ratio of the positively-selected sites, while M8 only includes site-classes above neutrality. The models allowing positive selection along the alignment (M2a and M8) were compared pair-wise against stricter models, M1a and M7 respectively, using likelihood ratio tests (LRT). Comparisons between models M8 and M8a were used to identify deviations from neutrality. This pair-wise comparison focuses on testing whether sites belonging to a site-class with a $d_N/d_S > 1$ are evolving differently from nearly neutrality ($d_N/d_S \approx 1$). The results of LRT = 2*[*lnL* (alternate model) – *lnL* (*null* model)] (or LRT = ΔlnL), for each pairwise comparison, M1a vs M2a, M7 vs M8, M8vs M8a, was compared against a χ^2 distribution. The degrees of freedom, used to obtain the χ^2 critical values, were calculated by comparing the parameters difference between the *null* and alternate model, for each pairwise test. The positively-selected sites positions where mapped to the human sequences using an *in-house* script available upon request.

Because, positive-selection analyses tend be less reliable in regions of poor alignment. For quality control, all multiple sequence alignments used for testing for positive selection were submitted to GUIDANCE (Penn, Privman et al. 2010) to obtain an alignment confidence score. The correlation and the confidence estimates of each alignment were plotted in a scatter plot (Appendices IV: Figure S2).

Exons and Introns Evolutionary Rate

Gene coordinates obtained from *Ensembl Biomart* were used to retrieve the phyloP (Pollard, Hubisz et al. 2010) site scores for introns and exons using the USCS browser (Kent, Sugnet et al. 2002). The pre-calculated values available in USCS tables only include placental mammals (Goldman, Craft et al. 2013). The empirical cumulative distribution (ECDF) from introns and exons and the Mann-Whitney U were obtained using MATLAB and Statistics Toolbox Release 2012b. Given that the number of analyzed positions (intronic and exonic) from negatively selected genes were greater than the number of positions in positively-selected genes, we built a script for sampling (allowing repetitions) the values from each intronic and exonic regions. For comparison and to reduce calculation times, booth pools of values were restricted to 300,000 sampling points in introns and to 100,000 in exonic positions. To validate the procedure, for each scenario, three random samples of introns and exons from positively and negatively selected genes were generated from each pool of values and were tested for homogeneity using the Mann-Whitney U and Kernel Density tests.

Protein age, characteristics and functional clustering

Protein ages were estimated with PPODv4_OrthoMCL_families and Dollo parsimony and grouped into three age classes defined as: ≤ 220.20 Myr (Mammalian specific), >220.20 Myr and ≤ 454.60 Myr (Vertebrate specific), >454.60 Myr (Older proteins). Using the *ProteinHistorian* (Capra, Williams et al. 2012) (Appendices IV: Table S2). The isoelectric points (pI) for human proteins were obtained from pepstats as implemented in EMBOSS (Rice, Longden et al. 2000). For positively-selected genes, the disorder status was calculated for each protein with SPINE-D (Zhang, Faraggi et al. 2012) using human sequences as the query. Positively-selected genes were grouped into functional clusters based on DAVID (Huang da, Sherman et al. 2009). The protein interactions were retrieved from BioGRID (Stark, Breitkreutz et al. 2006; Chatr-Aryamontri, Breitkreutz et al. 2013) and all proteins with more than 100 interactions were excluded. Statistical analysis was performed in SPSS (Corp. Released 2011; SPSS Released 2011).

Expression of tooth associated genes during development

Expression profiles were obtained from NCBI GEO (Barrett, Suzek et al. 2005; Barrett, Wilhite et al. 2013). We used two mouse experiments that corresponded with tooth

germ tissue at embryonic day 13.5 (Lachke, Ho et al. 2012) [GEO:GDS4453] and to the post-natal stage (Pemberton, Li et al. 2007) [GEO:GSE7164] and one human experiment corresponding to embryonic stages from 4-9 weeks after fertilization (Yi, Xue et al. 2010) [GEO:GSE15744]. For each dataset were used the GPL-associated files to filter the tooth-associated genes and later was log2 normalized to reduce noise. For different probes associated with the same gene, their value was averaged. The average expression value was used for probes associated with genes with more than one time point. Cluster analysis using k-means was performed in MATLAB (MATLAB. 2012) and the statistical analysis was performed using SPSS (Corp. Released 2011; SPSS Released 2011).

4.4 Results

Gene localization and functions

Genes associated with tooth development were plotted on an ideogram (Figure 4-1) showing their location in the human genome. Of the 247 tooth-associated genes, 10 are located on Chr X and one on Chr Y, while the remaining 236 are autosomal. *MECP2* was the only X-linked gene with evidence of positive selection, compared with 35 positively-selected autosomal genes (~15.4%).

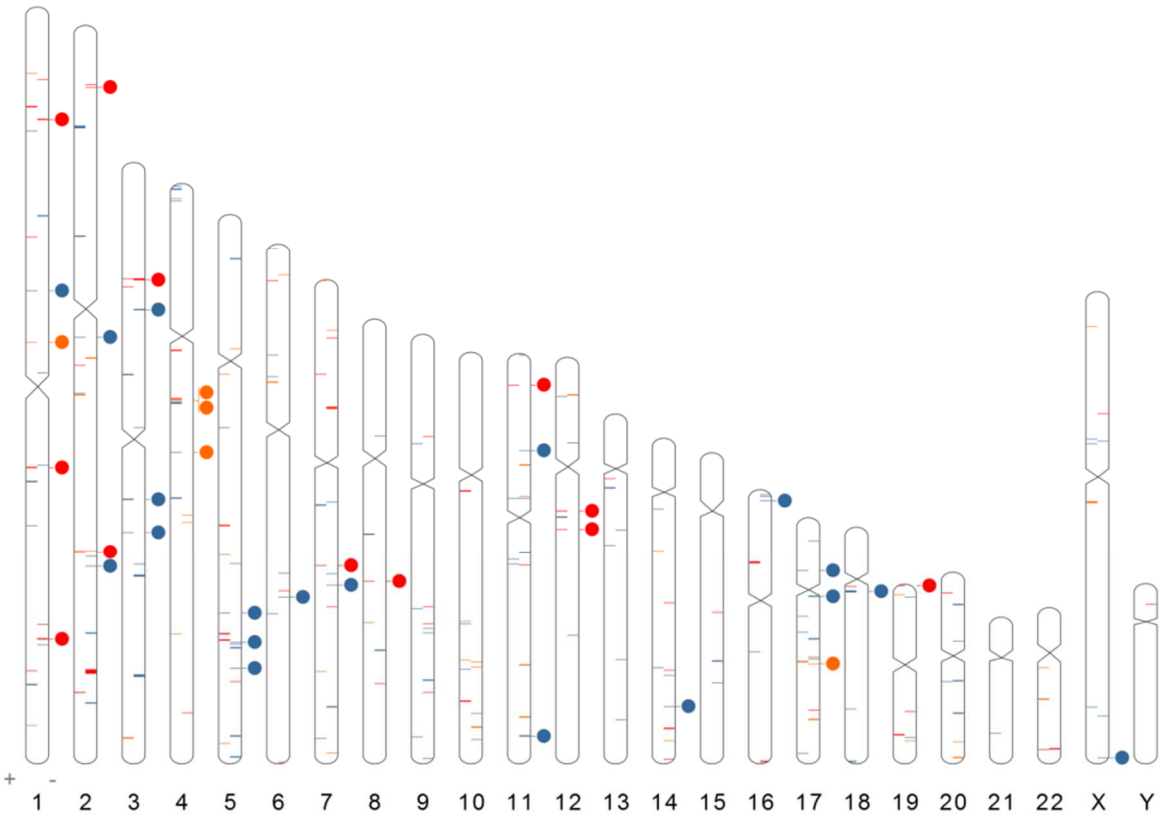

Figure 4-1. Ideogram of the human genome. Human chromosomal location of the tooth-associated genes. Blue identifies genes associations based on Mammalian Phenotype (MP), red represent genes associations based on Gene Ontology (GO) and orange denotes genes associations from both MP and GO. Solid circles identify positively selected genes. Each chromosome is labeled with their respective number (autosomal chromosomes) or letter (sexual chromosomes) code. The symbols (+) and (-) represent the DNA strand orientation.

The majority of the tooth-associated genes identified in this study is not restricted to only tooth-associated processes, but are also involved in other processes (Appendices IV: Table S1). This pleiotropic effect has been reported for genes such as *BMP4*, e.g. primarily associated with colorectal cancer (Houlston, Webb et al. 2008) and later in Parkinson's disease (Simon-Sanchez, Schulte et al. 2009).

To understand the genomic positions of genes in other mammalian species we plotted the gene locations in mouse and dog, which represent Rodents and Carnivores and with Primates, make up 20 of the 39 species included in this study. Gene location in these other species is also dispersed, and positively-selected genes are not closely linked in either mouse or dog (Figure 4-2).

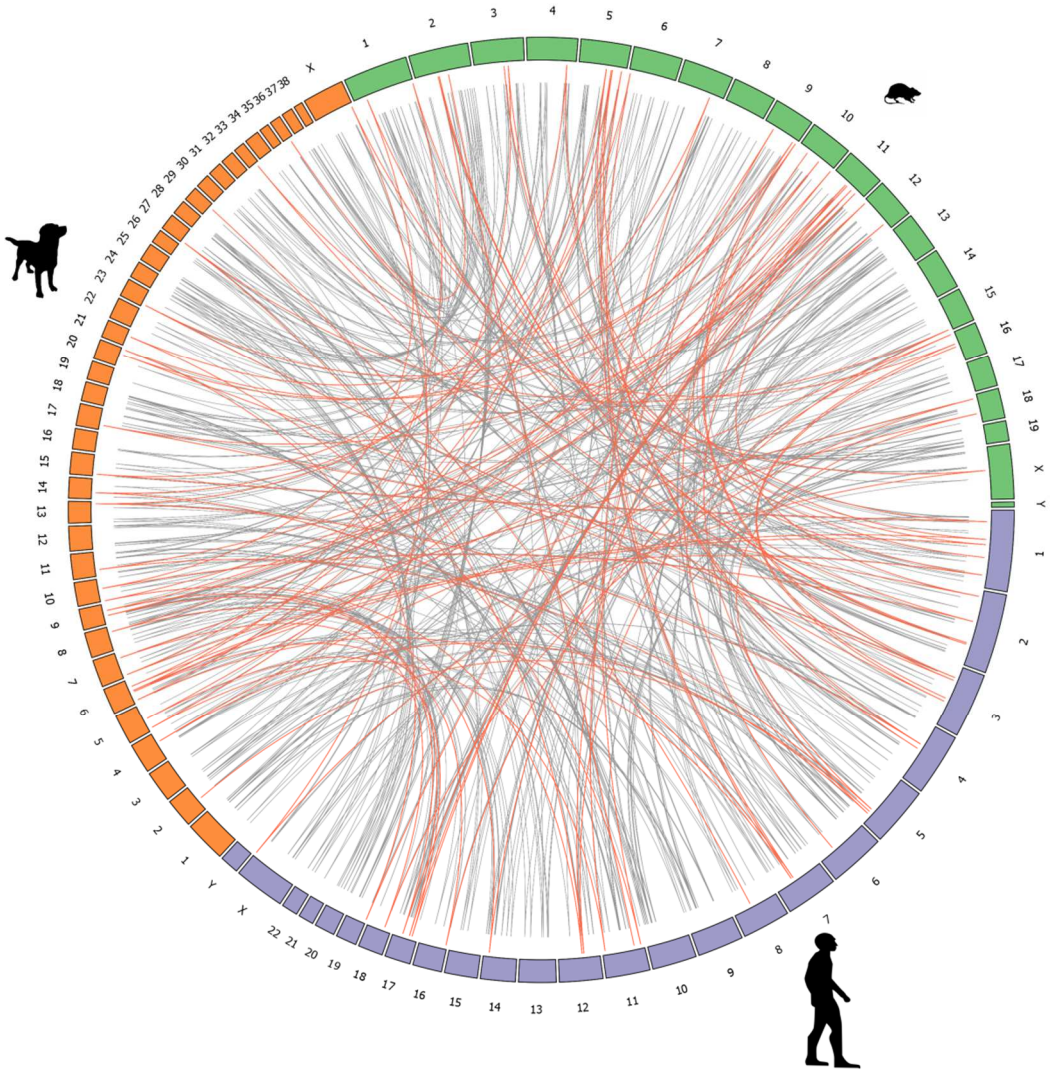

Figure 4-2. Genomic location of tooth-associated genes in the dog, human and mouse genome. The dog chromosomes are represented in orange, mouse chromosomes in green and human chromosomes in purple. The red connecting lines identify tooth-associated genes significantly influenced by positive selection in the three genomes and the gray lines correspond to tooth-associated genes non-positively selected.

Selective regime in tooth-associated genes

When the gene tree was used as input for CODEML analysis, M8 was preferred in 158 genes relatively to M7, although when using the strict pairwise comparison M8 vs M8a, only 40 genes showed the site class (w>1) significantly above neutrality. When using the species tree, selection analysis revealed that the alternate model (M8) was accepted in 163 genes and was restricted to 51 in the M8 vs M8a comparison. The interception of both analyses (gene tree vs species tree) allowed us to infer that 36 genes (~15.3%) showed strong signatures of selection. The thorough comparison between the gene tree and species tree allowed us to reduce the possible bias that an incorrect phylogenetic topology may

introduce into positive-selection analyses. This consistency under different evolutionary assumptions strongly supports the presence of positively-selected sites in 36 genes. The pairwise comparison of M7 vs M8 has been shown previously to be less robust (but more powerful) than the M1a vs M2a comparison (Nielsen and Yang 1998). Under model M2a and using the gene-based tree, 35 genes showed signatures of positive selection while 41 genes favored the alternate model when the species tree was used. Although it is significantly faster, the comparison between M2a vs M1a retrieved 23 of the genes that were also identified as being under positively-selected with M8 vs M7 and M8 vs M8a. Therefore, ~9.7% of the genes showed signatures of selection, independent of the model and the phylogenetic assumption. Despite being faster, M2a was the most sensitive to the phylogenetic assumptions since the results obtained from the species tree and gene tree were less similar when compared with the more parameter-rich pairwise analysis. The Spearman's correlation between the model M2a vs M1a and M8 vs M7, show that the primer model comparison is more sensitive to the input tree used in the detection of positive selection (Appendices IV: Figure S3).

The majority of the proteins involved in tooth formation showed evidence of being under strong negative selective pressure, as 130 genes (around 55% of the analyzed genes) had an omega ratio bellow 0.1. However, was absent a relationship between the global omega ratio and the presence of a selection signature since the genes with strong selection signatures had omega ratios from 0.077 to 0.650. From our results, the positively-selected sites under M7 vs M8 using the gene tree retrieved 286 sites with evidence of positive selection. Using the same approach (i.e. concordance between species and gene tree), we were able to retrieve 231 sites under positive selection (posterior probability above 0.95). Positively-selected sites positions were annotated using the human protein as reference (Appendices IV: Table S3). The posterior probabilities were calculated for each site using the human sequences as references for M8 and assuming the gene trees are not located on the extremity of the gene (Appendices IV: Figure S4). Remarkably, ~73.3% of the positively-selected sites were located in disordered regions, therefore matching regions commonly characterized by the lack of a stable tertiary structure (Figure 4-3).

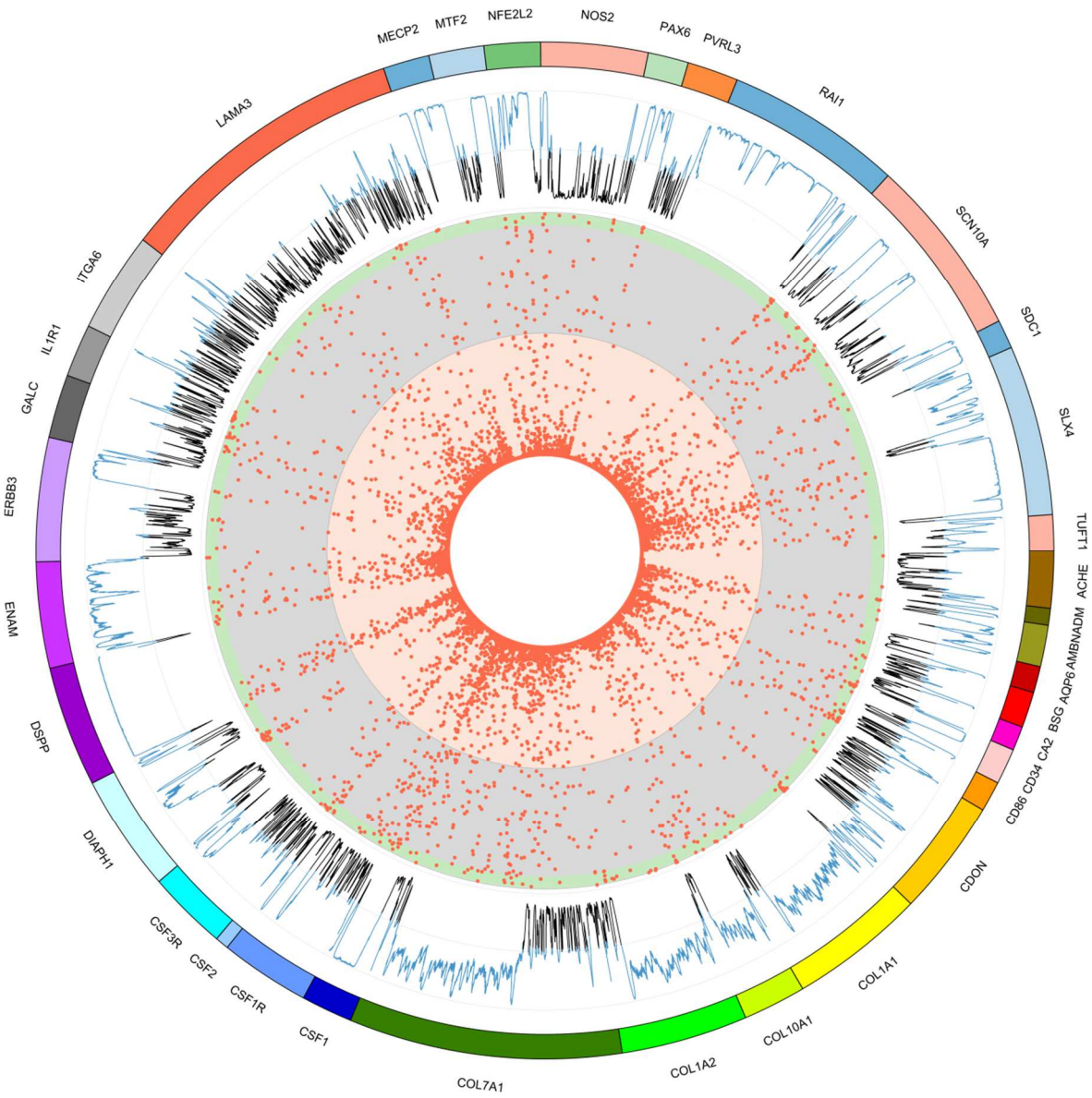

Figure 4-3.Tooth-associated genes under positive selection. The BEB posterior probability under M8 obtained using the gene tree is plotted in red dots in the center of the figure. The green region corresponds to a PP≥0.95, the grey region to 0.5≤PP<0.95, while the red region corresponds to PP<0.5. The graphic line corresponds to the calculated disorder probability, with the blue lines identifying the disordered regions.

Alignment uncertainty and phylogenetic resolution

The MSA from the positively-selected genes were submitted to GUIDANCE to confirm that the majority of the alignments were robust and therefore most of the positive selection was not due to improper alignment or due to uncertainty in some regions. In the 36 positively-selected genes, no associations were observed between the proportion of sites under selection and any detected alignment uncertainty (Appendices IV: Figure S5). Because the terminal portions of the alignments tend to be more difficult to align, it has been

reported that these regions may have a tendency to have high false-positive ratios. However, in our dataset, the positive-selected sites were dispersed relatively evenly from tail to core, decreasing the probability that poor alignment quality may have led to some false-positive or false-negative results. Moreover, the TREE-PUZZLE results showed that there is no association between evolutionary rate and the uncertainty in the phylogenetic signal, as in the majority of the positively-selected genes, fewer than 10% of quartets were unresolved with only a few exceptions (*ADM*, *AMBN*, *AQP6*, *CA2*, *CSF2*, *MTF2*, and *PVRL3*) (Appendices IV: Table S4).

Intronic acceleration in positively selected genes

Empirical Distribution Function (ECDF) showed that there was an intimate association between accelerated mutation rates in exons and the intronic regions of the corresponding genes (Figure 4-4a) as the positively-selected genes showed acceleration in both exonic and intronic regions when compared with the negatively selected genes. There was a significantly higher departure from neutrality in positively selected genes for introns and exons based on a Man-Whitney U test , $p<<0.01$ (Figure 4-4A). Also evident since the 50% more accelerated sites, are within lower value of phyloP scores (Figure 4-4A). These phyloP score values were obtained from USCS computed values, in their calculations the non-placental mammals were excluded but there is no expectation that this would significantly alter the outcome from phyloP analysis. The first intron of positively and negatively selected genes also were significantly different, although at a lower level (p-value=0.0149). Using a confidence level set at 0.05, are denoted differences between the first intron of negatively and positively selected genes. Yet using a stricter cut-off value these differences fail to be statistically significant at a critical value of 0.01. Therefore this show a more constrained evolution in the first introns, irrespective to the presence of positive selection in coding regions, since the differences between these introns of positively and negatively selected genes are less supported when compared to the analysis considering all the introns.

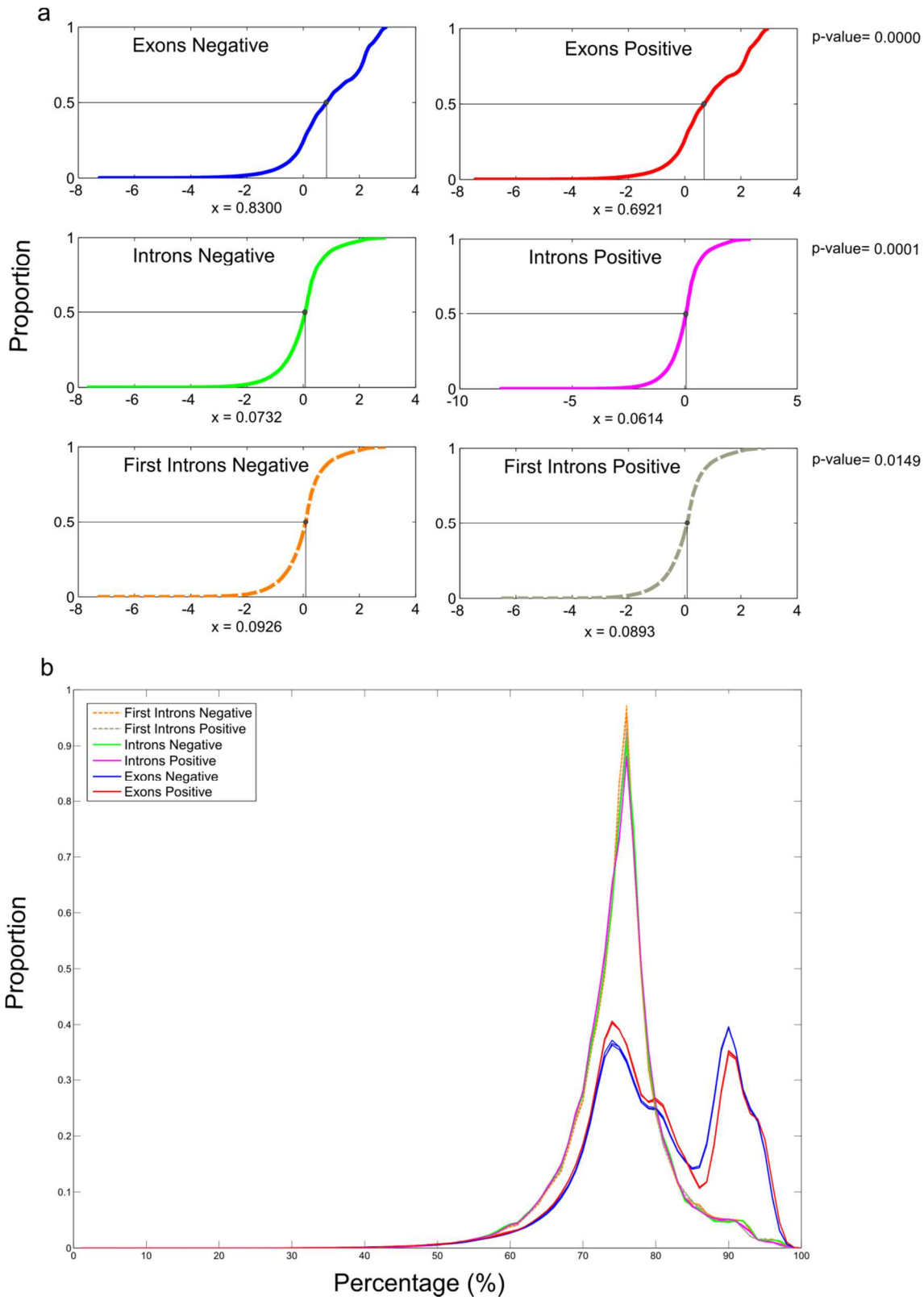

Figure 4-4. Comparison between phyloP scores of positively and negatively selected genes.a) ECDF obtained for tooth associated genes, introns and exons. P-values represents the Man-Whitney U test result from the 3 pairwise comparisons. B) Kernel density analysis.

Positively selected genes implicated in diseases

From the list of 36 genes under positive selection, 20 are associated with genetic diseases in the OMIM database. However, only two of these, *ENAM* and *DSPP,* have diseases that are specifically associated with teeth (Amelogenesis imperfecta type IB and Amelogenesis imperfecta type IC with *ENAM* and Deafness, autosomal dominant 36 with dentinogenesis, Dentin dysplasia type II, Dentinogenesis imperfecta Shields type II and Dentinogenesis imperfecta Shields type III with *DSPP*). The functional clustering analysis, using a classification stringency of *"high"*, revealed 17 clusters from the 36 positively-selected genes (Appendices IV: Table S5). Two of these clusters were intimately associated with biomineralization and/or structural constituents of tooth enamel (*ACHE*, *AMBN*, *COL1A1*, *DSPP*, *ENAM* and *TUFT1*).

Acceleration of recent proteins

The proteins were classified into three distinct phylogenetic groups according to their predicted gene age: Mammalian (mammalian-specific), Vertebrate (vertebrate-specific) and Old (older protein). For each protein clusters we calculated average omega, number of positively selected sites, GC content and GO processes for each category. Despite high variability, d_N/d_S estimates from M0 in CODEML supported the hypothesis that more-recently evolved proteins had accelerated evolutionary rates (Figure 4-5), as the average omega from mammalian-specific proteins was slightly higher than proteins that arose before the mammalian divergence. The younger proteins, i.e. mammalian specific, were shorter, were involved in fewer GO processes, had protein coding sequences with slightly-lower GC content, and had fewer interactions (Figure 4-5). Moreover, the positively selected genes encoded proteins that were slightly more acidic and had a lower average age (Figure 4-6).

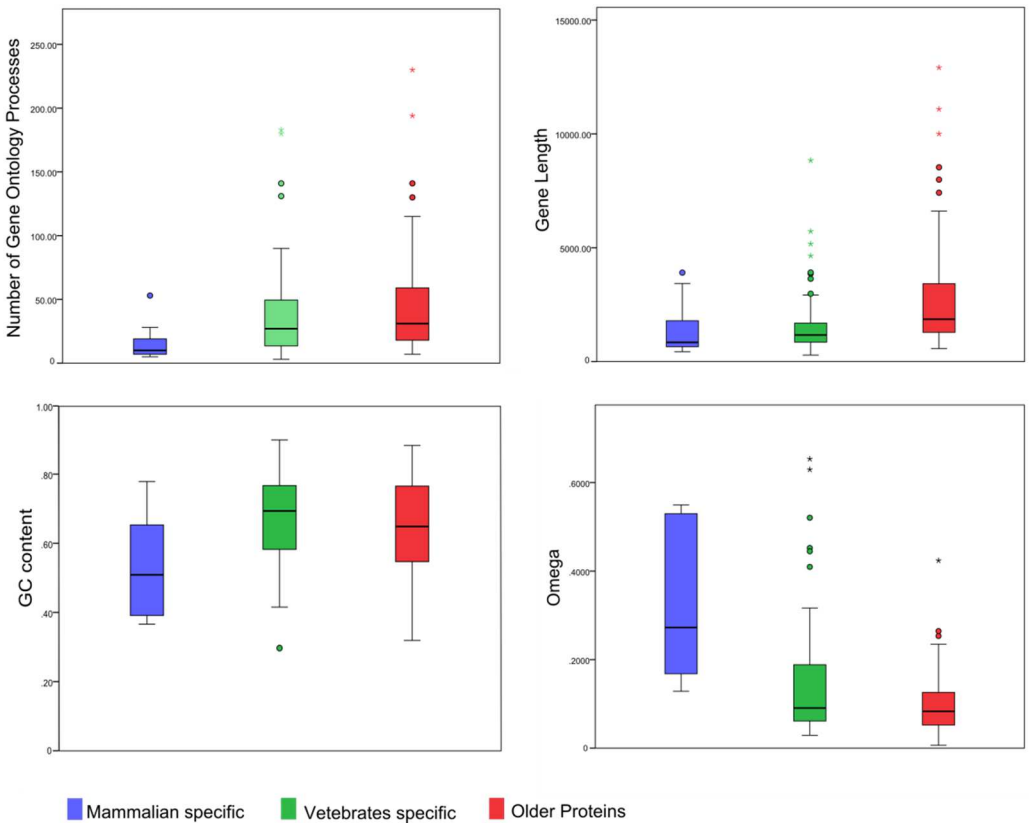

Figure 4-5. Age class clusters of the tooth associated genes. Average length of the protein, number of interactions and evolutionary rate (Omega) for the age class are represented for the GO (Gene Ontology) processes.

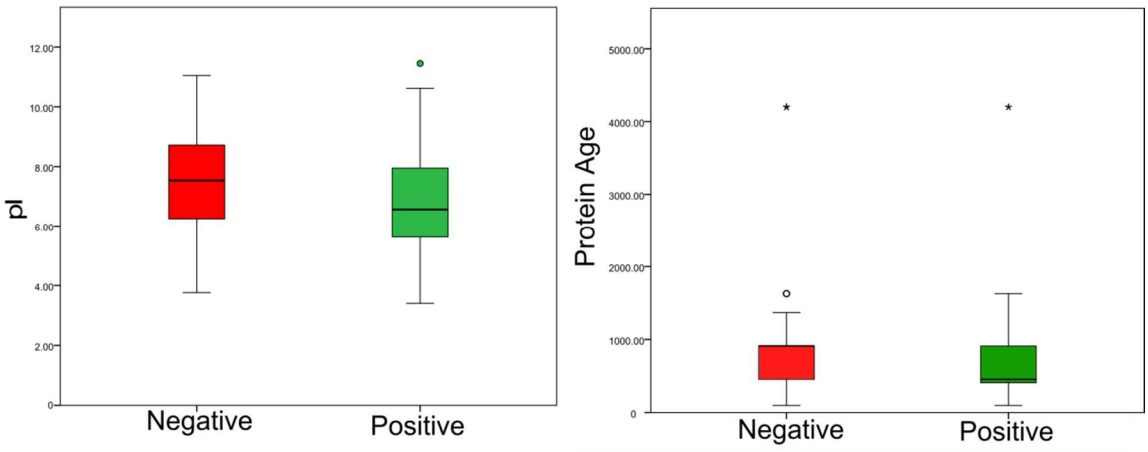

Figure 4-6. Tooth associated genes under positive and negative selection. Positively selected genes (represented in green) and non-positively selected genes (in red) given the average length and pI of the protein-coding genes.

Expression pattern of tooth-associated genes

Expression data supported the hypothesis that the younger genes are less expressed in early stages of tooth development. The GDS4453 experiment, which corresponds to an early stage of tooth development in mice E13.5, showed that at this stage there is a slightly lower expression of “young” proteins. Moreover, results from GSE7164 (Figure 4-7), which corresponds to a post-natal stage, showed that the there is a more-similar expression pattern of the younger proteins relative compared with either vertebrates or older proteins.

There were no significant difference between positively-selected genes and negatively-selected genes in GDS4453 and GSE7164. The expression data from GDS4453, corresponding to weeks 4 to 9 of human embryonic development revealed that the expression of younger proteins was lowest from 4^{th} to 6^{th} week, similar to patterns observed in other stages (GDS4453 and GSE7164). Interestingly, the 36 positively-selected genes had different expression patterns during these stages. From the 16 k-clusters examined, only clusters 1, 5 and 6 did not have any gene under positive selection (Appendices IV: Figure S6). We did not find any correlation between GC content (at CDS) and GC3 (GC content in the 3^{th} position) and expression level (data not shown).

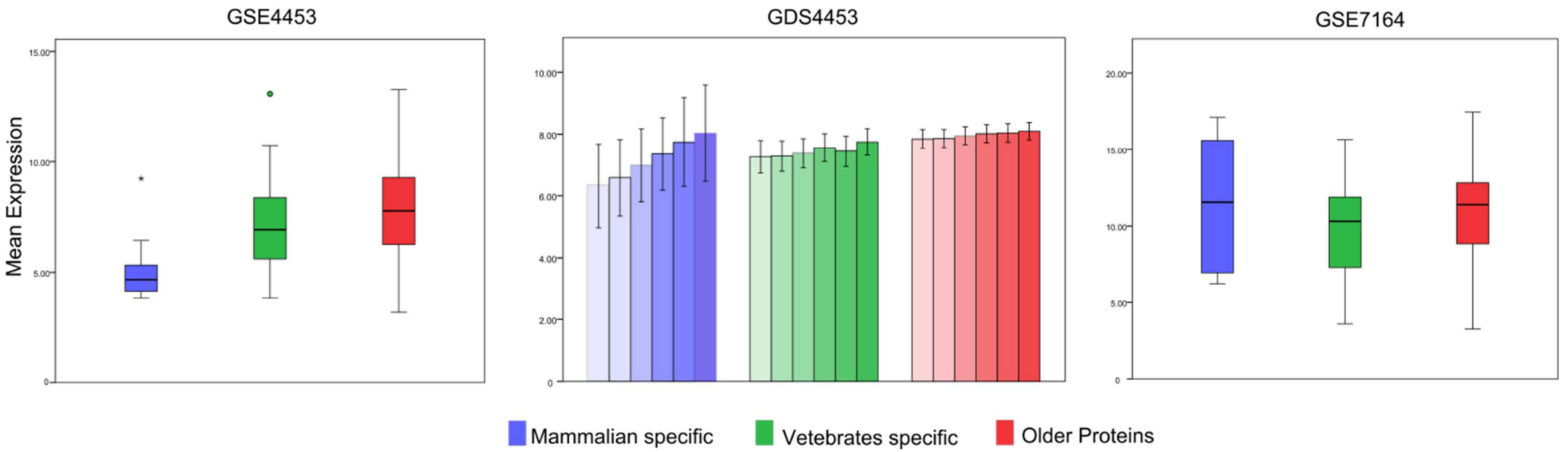

Figure 4-7. Expression profile of the tooth-associated genes. Results from experiments GDS4453, GDS4453 and GSE7164 are represented from left to right, respectively. In GDS4453 the color gradient corresponds to the different stages, from 4th (clearer) to 9th week (opaque). The mammalian-specific tooth associated protein-coding genes are down-regulated in early developmental stages.

4.5 Discussion

Most vertebrates possess teeth in jaws. The few exceptions, such as birds, lost teeth through evolution. Therefore, since teeth first appeared in jawed vertebrates around 460 Mya (Smith and Coates 1998), dentition has been subjected to purifying selection. The appearance of teeth involved an intricate coordination of multiple genes that likely shared functions required for coordination of tooth development. However, most of these genes were not novelties and were involved in other functions previously. Genes that are physically located close to each other are more likely to be co-expressed and to share a common ancestral function than more-dispersed genes (Cohen, Mitra et al. 2000; Woo, Walker et al. 2010). However, tooth-associated genes in the mammalian genome are widely dispersed (Figure 4-1; Figure 4-2). This suggests that tooth development depends on the coordination of multiple genes that previously were involved in other functions. The earliest tooth-like structures of the vertebrate oral cavity were first located outside the mouth and served diverse functions including protection, sensation and hydrodynamic advantage (Koussoulakou, Margaritis et al. 2009).

While the majority of the genes were subjected to purifying selection, some have evolved under positive selection, at least at some sites. Previous studies showed that positively-selected sites are functionally relevant (Morgan, Shakya et al. 2012; Dasmeh, Serohijos et al. 2013) and therefore sites with a significant high omega value are expected to have a determinant fitness role. Since natural selection has shaped the current diversity of tooth dentition in mammals, sites with evidence of positive selection signatures should be linked with differential selective advantages in each species. However, distinguishing neutral selection from a positive selection regime acting on genes is often complicated. Here we overcome this uncertainty by comparing the results of two robust methods to detect selection by using both gene trees and species trees, with the premise that this approach is more reliable and less subject to statistical noise when estimating the degree of selective pressures acting on the genes.

While the majority of the sites were evolving under negative selection, the presence of sites with an omega greater than one supports their role in determining protein functionality, and therefore demonstrating their role in the development of mammalian phenotypic differentiation. The present study demonstrates that tooth-associated genes have different selection signatures and therefore affirms their important role in mammalian adaptations. We identified 36 genes that are most-likely to be responsible for the tooth diversification among mammals. Within these 36 genes, we found 286 sites under positive selection that

were mostly located within intrinsically-disorder protein regions. This confirms previous findings that there is an over-representation of positively selected sites encoding intrinsically disordered regions of proteins (Nilsson, Grahn et al. 2011). Furthermore, there was no evidences of under-representation of functional amino acids in intrinsically disordered regions of proteins (Nilsson, Grahn et al. 2011). Here, we found that positive selection in tooth-associated genes was more persistent in disordered regions, which is important since disordered regions in proteins allow the kinases, phosphatases, and phosphorylation-dependent binding to obtain access to target sequences and therefore to regulate local protein conformation and activity (Collins, Yu et al. 2008). Moreover, there is a strong correlation between biomineralization and structural disorder of proteins (Kalmar, Homola et al. 2012). Therefore, these sites, particularly those corresponding with disordered regions, are potentially of prime relevance to the function of these proteins, and thus are potential sites for site-directed mutagenesis.

Within the group of positively-selected genes we found two clusters of genes that were involved in tooth-specific, biomineralization. As these two clusters were composed by genes with a crucial relevance to the tooth formation, they are therefore potential candidates for future study to determine their specific roles in the phenotypic diversification of the dentition in mammals. One of these positively-selected genes, *ENAM*, was previously demonstrated to have signatures of positive selection in human populations (Kelley, Madeoy et al. 2006) and in Kalmar dogs (Kalmar, Homola et al. 2012). In addition, ENAM has been linked with tooth enamel thickness and dietary changes in primates (Kelley and Swanson 2008). From our analyses, we suggest that *ACHE*, *AMBN*, *COL1A1, DSPP* and *TUFT1* are also genes that have been involved in mammalian dentition adaptations. It was previously suggested that the *AMBN* and *ENAM* are multifunctional proteins, essential in early stages of tooth development (Landin, Shabestari et al. 2012). However, here we re-analyzed three different microarrays (Pemberton, Li et al. 2007; Yi, Xue et al. 2010; Lachke, Ho et al. 2012), and the results suggests a higher expression of those genes during tooth development in later stages.

Previous studies have demonstrated an unexpectedly high degree of sequence conservation in introns (Hare and Palumbi 2003) and among intron position in orthologous genes (Henricson, Forslund et al. 2010), as well as the presence of mutational cold spots corresponding to regions that are under negative selection higher than protein coding regions (Katzman, Kern et al. 2007). Here we reported that introns in negatively-selected genes are also under a higher selective regime than in positively selected genes. Given the functional importance of the intronic regions, it is expected that this asymmetrical evolutionary rate may have functional relevance. Several studies have demonstrated the presence of regulatory elements in mammalian introns, particularly in the first introns (Oshima,

Abrams et al. 1990; Jonsson, Foresman et al. 1992). Here we also observed an asymmetrical evolutionary rate between the first introns of the positively selected genes and of the negatively selected genes, supporting previous observations that divergence rates are higher in the first introns than in the non-first introns (Gazave, Marques-Bonet et al. 2007). It is probable that this asymmetrical evolutionary rate between positively and negatively selected genes is not restricted to the tooth-associated genes, but instead is a common trend observed in other genes. Previously it was demonstrated that purifying selection had a strong effect on shaping intron sequence evolution between human and chimpanzee and it was hypothesized that introns can evolve under positive selection (Gazave, Marques-Bonet et al. 2007).

A recent study showed that the core dental gene network is a common feature found in species with the first pharyngeal tooth and all of its jawed descendants (Fraser, Hulsey et al. 2009). An ancient dental regulatory network (*BARX1*, *EVE1*, *LHX7*, *LHX8* and seven HOX's genes) and dental circuit (*BMP2*, *BMP4*, *DLX2*, *EDA*, EDAR, *PAX9*, *PITX2*, *RUNX2* and *SHH*) was reported in cichlids (Fraser, Hulsey et al. 2009). Here we found that all of these genes, except *EVE1,* were under purifying selection. It has been suggested that this core dental network is evolutionarily essential since there is no corrected patterning of the dentition without the involvement of those genes (Fraser, Hulsey et al. 2009) and the appearance of those genes predates the vertebrates' emergence (Fraser, Hulsey et al. 2009).

Although previous studies have reported correlations between evolutionary rate, structural properties and age class (Toll-Riera, Bostick et al. 2012) our results support the hypothesis that younger proteins, i.e. mammalian-specific proteins, are involved in fewer GO processes, are involved in fewer interactions, are shorter and have higher evolutionary rates. Likewise, GC content in these younger proteins is slightly lower than in older protein-coding sequences. Although some of these observations have been previously reported, the importance of these patterns are still being debated. In our dataset, the higher evolutionary rates were observed in the younger proteins, suggesting that most of the phenotypic diversity observed in the mammalian dentition may rely on "new proteins", while "older" proteins are more-likely to be under strong purifying selection. In addition, our analyses of expression data revealed that these younger proteins are expressed less in early stages of tooth development compared with later stages.

4.6 Conclusions

The top-down analysis of 236 tooth-associated genes revealed 36 genes with evidence of significant positive selection based on at least one the methods. Positively selected sites tended to be located in disordered regions of the protein, and therefore are more likely to be functionally relevant. Clustering analysis revealed six genes (*ACHE*, *AMBN*, *COL1A1*, *DSPP*, *ENAM* and *TUFT1*) that were positively selected that had clear links with odontogenesis. However, their full role in mammalian phenotypic diversity is still unknown. The asymmetrical evolutionary rate among introns of positively-selected genes and negatively-selected genes suggests that intronic regions may also have a role in the mammalian diversification. Age-class analysis revealed that the more-recently evolved proteins are expressed in later developmental stages and because given their higher evolutionary rate, are probably linked with the diversification of the mammalian dentition.

Our results suggest that the evolution of mammalian dental patterns arose through strong positive selection of genes previously involved principally in other functions and are evidence of innovations arising from the adaption of genes previously involved in other networks.

4.7 Acknowledgements

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to JPM (SFRH/BD/65245/2009) and SP (SFRH/BD/47938/2008). This work was further supported by a grant from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism. AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013 and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490).

Chapter 5 *- Adaptive Functional Divergence of the Warm Temperature Acclimation-Related Protein (WAP65) in Fishes and the Ortholog Hemopexin (HPX) in Mammals*

5.1 Abstract

Gene duplication is an important mechanism leading to genetic novelty. Different, non-exclusive processes are likely involved, and many adaptive and non-adaptive events may contribute to the maintenance of duplicated genes. In some teleosts, a duplicate copy of the mammalian ortholog Hemopexin (HPX) is present, known as the warm temperature acclimation-related protein (WAP65). Both WAP65 and HPX have been associated with iron homeostasis due to the affinity to bind the toxic free heme circulating in the blood stream. We have assessed the evolutionary dynamics of *WAP65* and *HPX* genes to understand the adaptive role of positive selection at both nucleotide and amino acid level. Our results showed an asymmetrical evolution between the paralogs WAP65-1 and *WAP65-2* after duplication with a slight acceleration of the evolutionary rate in *WAP65-1*, but not in *WAP65-2*, and few sites contributing to the functional distinction between the paralogs while the majority of the protein remained under negative selection or relaxed negative selection. WAP65-1 is functionally more distinct from the ancestral protein function than *WAP65-2*. HPX is phylogenetically closer to *WAP65-2* but even so functional divergence was detected between both proteins. Also, HPX showed a fast rate of evolution when compared to both *WAP65-1* and *WAP65-2* genes. The assessed three-dimensional structure of WAP65-1 and WAP65-2 suggests that the functional differences detected are not causing noticeable structural changes in these proteins. However, such subtle changes between WAP65 paralogs may be important to understand the differential gene retention of both copies in 20 out of 30 teleosts species studied.

5.2 Introduction

The hemopexin (HPX) is a plasma glycoprotein consisting of a single polypeptide chain with 60-kDa that binds to heme in an equimolar ratio with high affinity, higher than albumin, competing for free heme in the plasma (Muller-Eberhard 1988). Free heme is toxic to the cells and it is a potential source of iron for pathogens (Paoli, Anderson et al. 1999). The evolution of such a high affinity is a functional/structural response to the necessity of sequestering free heme within the blood stream with great efficiency (Paoli, Marles-Wright et al. 2002). In mammals, the HPX–heme complex is transported to specific receptors in the liver and up-taken by the liver parenchymal cells through receptor-mediated endocytosis mediated by the scavenger receptor LRP1 (also known as CD91) (Hvidberg, Maniecki et al. 2005). The heme group is released from *HPX* upon an environment with low pH (<5.0), such as the endosome, most likely by the protonation of one or both histidines residues

coordinating the heme binding and the movement of the HPX domains and/or linker leading to the disruption of the heme pocket (Paoli, Anderson et al. 1999). After the heme release, the iron is stored into the hepatic ferritin while the apohemopexin returns to the circulation (Morgan, Liem et al. 1976). Several studies suggest that hemopexin is not only a plasma transporter of the heme but also act as a multifunctional agent in important health-related processes, such as iron homeostasis, antioxidant protection, bacteriostatic defense (limiting the access by pathogens to heme), nerve regeneration, and gene expression to promote cell survival (Delanghe and Langlois 2001). Analysis of the internal homology in amino acid sequence indicates that HPX comprises two homologous domains of about 200 residues each, joined by a 20-residue linker (Paoli, Marles-Wright et al. 2002). The Fe (III) of the heme is coordinated by two histidine residues and further stabilized by a host of non-covalent interactions provided by a large number of invariant aromatic and basic residues (Paoli, Anderson et al. 1999).

In fishes, there is an ortholog gene of the mammalian HPX, the warm-temperature-acclimation-associated 65-kDa protein (*WAP65*) that was initially identified in goldfish (*Carassius auratus*) and later in several other fishes, namely *Acanthopagrus schlegeli, Cyprinus carpio, Ictalurus punctatus, Oryzias latipes, Takifugu rubripes* and *Xiphophorus helleri* (Kinoshita, Itoi et al. 2001; Hirayama, Nakaniwa et al. 2003; Nakaniwa, Hirayama et al. 2005; Aliza, Ismail et al. 2008; Choi, An et al. 2008; Takano, Sha et al. 2008). It has been observed in *I. punctatus*, *O. latipes* and *T. rubripes* the presence of two paralogs, the *WAP65-1* and the *WAP65-2*, both structurally similar to the mammalian HPX, although exhibiting highly differential patterns of spatial expression (Sha, Xu et al. 2008). WAP65-1 is expressed in a wide range of tissues while *WAP65-2* is only expressed in the liver (Sha, Xu et al. 2008). The regulation with warm temperature and bacterial infections is also highly different: WAP65-1 is constitutively expressed, whereas *WAP65-2* is highly regulated both by warm temperature and bacterial infections, these two stimuli acting synergistically to induce the expression of *WAP65-2* (Sha, Xu et al. 2008). The water temperature is one of the most notable factors that bear a spatial and temporal influence on aquatic organisms (Kinoshita, Itoi et al. 2001). While seasonal temperature changes take place over weeks or months, physiological reorganization compensating for such changes is often referred as temperature acclimation (Hazel and Prosser 1974).

As iron is one of the pivotal elements during bacterial infections (Cherayil 2011), several studies have explored the potential involvement of WAP65 in immune responses given its structural similarity to *HPX* (Sha, Xu et al. 2008). In goldfish, WAP65-2 respond to bacterial infection but notably it might also function as an immune response protein (Kikuchi, Watabe et al. 1997). Both WAP65-1 and WAP65-2 act as a multifunctional agent in several biological processes like immune response (Peatman, Baoprasertkul et al. 2007; Peatman,

Terhune et al. 2008; Shi, Chen et al. 2010), iron homeostasis, heavy metal exposure (Aliza, Ismail et al. 2008), temperature acclimation (Kikuchi, Watabe et al. 1998; Sha, Xu et al. 2008) and development (Hirayama, Kobiyama et al. 2004; Nakaniwa, Hirayama et al. 2005). The multifunctional aspects of WAP65 proteins suggested that both genes underwent neofunctionalization and have diversified their functions (Sha, Xu et al. 2008). Therefore, WAP65-1 have evolved to encompass new functions, whereas WAP65-2 have retained its initial functionality as a major role, namely in the acclimation to warm temperature and in the immune response (Sarropoulou, Fernandes et al. 2010). The differential expression pattern suggests a functional distinction between WAP65 proteins although the residues contributing to the functional distinction of the paralogs has not been characterized and neither the mechanism involved in the fixation of both copies in teleosts. Moreover, the lack of comparative analyses between the paralogs in teleosts and the mammalian HPX precluded the identification of the different evolutionary forces that influenced the duplicated copies in fishes and the counterpart singleton gene in mammals.

To understand the evolution and divergence of the *WAP65* paralogs and the mammalian *HPX*, we have characterized in detail signatures of positive selection acting on these protein-coding genes. We have evaluated the molecular evolution of HPX and *WAP65* in 66 vertebrates, analyzing selection signatures that may have been responsible for the functional divergence between *WAP65* in fishes and *HPX* in mammals, particularly by testing the branch immediately after duplication. Our analyses showed that positive selection has significantly influenced the evolution of these proteins in fishes following the duplication event that originated the *WAP65-1* and *WAP65-2* paralogs, with few sites contributing to the functional distinction. Moreover, adaptive evolution is likely responsible not only for the functional divergence between *WAP65-1* and *WAP65-2* paralogs, but also for the functionally distinctiveness of these proteins relatively to the mammalian HPX, as suggested by the evolutionary acceleration of *HPX* relatively to both *WAP65-1* and *WAP65-2* genes. The modeled three-dimensional (3D) structure of WAP65-1 and WAP65-2 shows that the functional distinction of the paralogs is not associated with the ability to bind free heme.

5.3 Methods

Sequence analyses

The nucleotide sequences and protein sequences of WAP65-1, WAP65-2 and the mammalian ortholog HPX were retrieved from GenBank and ENSEMBL. Several TBLASTN searches were done in order to retrieve non-annotated sequences from EST databases. Multiple EST alignments were performed using ClustalW in Bioedit (Hall 1999). The open reading frame of each EST was manually inspected and corrected in order to perform an amino acid alignment using translated nucleotide to avoid the insertion of incorrect bases. A consensus sequence was build when multiple ESTs were found for the same species. All the sequences retrieved were represented at least by two different ESTs, the bases with ambiguity were manually correct, or when not possible we used the IUPAC recommendations (Cornish-Bowden 1985). The multiple sequence alignments (MSA) were built using MUSCLE (Edgar 2004) in SEAVIEW (Gouy, Guindon et al. 2010) and to avoid the improper alignment of non-homologous evolutionary positions, the alignments were performed using the translated nucleotides and back-translated to nucleotides. The MSA were therefore organized in three different datasets: i) *WAP65-1*, ii) *WAP65-2*, iii) *HPX*, used for further analyses of positive selection at both the nucleotide and the amino acid level, reducing the bias of base saturation presented in the dataset if combining both fishes and mammalian sequences. Two additional MSA were built; one using the 66 species studied in this work and other considering all the WAP65 proteins (WAP65-1, WAP65-2 and the WAP65 of cartilaginous fishes, named here as WAP65c). In the fish species where only one *WAP65* copy was detected, we performed TBLASTN searches to inspect ESTs that may indicate the presence of an additional gene copy. We detected additional ESTs in nine different species but we have not included such sequences in further detailed analyses given its short length. However, we built a phylogenetic tree with those sequences suggesting that the retrieved ESTs were phylogenetically similar to WAP65-1 in two species and to *WAP65-2* in seven species (Appendices V: Figure S1).

Phylogenetic analyses

Bayesian phylogenetic inferences were performed with MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003). The best-fit model of nucleotide substitution used was selected with jModeltest (Posada 2008). The reconstructions were obtained after adjusting the parameters accordingly to the best-fit model in agreement to the likelihood obtained for each evolutionary model after the Akaike Information Criterion correction (AICc)

(Appendices V : Table S1). Bayesian inference (BAY) methods with Markov Chain Monte Carlo (MCMC) sampling were performed in MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003b) and the phylogenetic trees reconstruction were built starting with a random tree using four Markov chains (three heated and one cold) running for 10 000 000 generations, sampling every 1000 generations and burning 25% of the sampled trees, with prior probability distributions of the individual model parameters according to the model specified (Appendices V: Table S1). A maximum likelihood (ML) phylogenetic tree was also constructed in PhyML (Guindon and Gascuel 2003) applying the corresponding evolutionary model, bootstrapping 1000 for the clade support and using the NNI tree searches. Additionally, we built trees in MEGA 4.0 (Tamura, Dudley et al. 2007) using the Neighbor-Joining (NJ) algorithm and the Maximum Composite Likelihood model (Tamura, Nei et al. 2004). We performed the two-sided Kishino-Hasegawa test (KH), the Shimodaira-Hasegawa test (SH), and Expected Likelihood Weights (ELW) in TREE-PUZZLE (Schmidt, Strimmer et al. 2002) to determine the best-fitting tree for each MSA (Appendices V: Table S2).

Detection of positive selection

Nucleotide level

WAP65-1, WAP65-2 and *HPX* coding sequences (CDS) were analyzed separately to minimize nucleotide saturation and base compositional bias. Tests for positive selection were performed with the likelihood method implemented in PAML v4.3 (Yang 2007) using a gene-level approach based on the ratio (ω) of non-synonymous (dN) to synonymous (dS) substitutions rate (i.e., ω = dN/dS). The Likelihood Ratio Tests (LRT) was used to compare the nested pair of models that allows variation in ω among codons but assuming the same distribution in all lineages: the null models, M1a and M7 (β) against the alternative (positive selection models), M2a and M8 (β + ω > 1). The likelihood of the two nested models, a model that does not allow a site class with ω > 1 and a model that allow (null vs. positive selection, respectively) is compared using a LRT test. The LRT=−2ΔlnL (ΔlnL = the difference in log likelihoods of the two models) follows χ2 distribution with degrees of freedom (df) equal to the difference in number of parameters between models. Additionally, we also used the M8a, with the omega value fixed to 1, checking if the site class above one was statistically different from the neutrality. For all the obtained LRTs, the transition-transversion ratio was calculated from the data and the equilibrium codon frequencies were obtained using the average base composition at the three codon positions (CodonFreq=2). The ambiguous sites were removed from these models (cleandata=1) since they

correspond to indels or ambiguous characters. A significant LRT only demonstrates that the selection model is preferred to the neutral model; it does not provide any kind of indication of the sites under selection (Osorio, Antunes et al. 2007). A posterior analysis is needed and thus we used a Bayesian Empirical Bayes (BEB) approach to calculate the posterior probability (PP) for each site to be within a specific site-class. A specific site is considered to be under strong selection if having a high probability (PP>0.95) to belong to the class with ω>1 (Yang, Wong et al. 2005). The Bayes Empirical Bayes (BEB) is a robust method, reliable for both small and large datasets (Yang, Wong et al. 2005). We also tested a branch model, using the simplest one ratio model, versus a two-ratio model labeling the post-duplication (PD) branch in teleosts, here referring to the branch immediately after the duplication event (Figure 5-1). This analysis provided information if those labeled branches would indeed had an altered evolutionary rate. Given the absence of a significant alteration in the selection pressure when considering the entire protein, we performed an alternative test, using the branch-site models. These models allow ω ratios to vary simultaneous among lineages of interest and along with the codons sites. Here, we used the branch-site analysis model A test 2, also referred to as the branch-site test of positive selection (Zhang, Nielsen et al. 2005) to understand which sites after the duplication showed signatures of selection in the PD branch, immediately after duplication.

Amino acid level

Recent methods for investigating selection in proteins (or coding genes) have focused on evaluating the type of positive selection detected (directional or non-directional, stabilizing or destabilizing), purifying selection, and how the identified selection affects the overall structure and function of the protein (Porter, Cronin et al. 2007). Amino acid substitutions may induce various effects on a protein depending of the physicochemical properties changed and also in the position of the substitution in the protein structure (Porter, Cronin et al. 2007). We performed an analysis to differentiate between types of selective pressures acting in *WAP65-1*, *WAP65-2* and *HPX*, including (i) positive selection, (ii) stabilizing selection (maintaining the overall biochemistry of the protein) and (iii) destabilizing selection (causing radical structural or functional shifts in local regions of the protein), which provided insight into the structural and functional consequences of the identified residues under selection (McClellan, Palfreyman et al. 2005). In TreeSAAP v3.2 (Woolley, Johnson et al. 2003), we analyzed 31 different amino acid properties in search for positive destabilizing selection and considering the properties with significantly greater amino acid replacements (relatively to the neutral expectations) with magnitude categories +7 and +8 (i.e., the two most radical property change categories). We defined an empirical

threshold of three properties as evidence that a site is under positive selection and excluded the sites with ambiguous characters in more than five sequences (e.g. indels).

Functional divergence

A likelihood ratio test based method (Gu and Vander Velden 2002; Gu 2003), was used to inspect type-I and II functional divergence implemented in Detecting Variability in Evolutionary Rates among Genes (DIVERGE v.2.0) (Gu and Vander Velden 2002). This method attributes a coefficient of functional divergence to each amino acid residues θ, which is in fact the change of evolutionary rate at the amino acid site between two clades. Moreover, the advantage of this method is that it uses amino acid sequences and, thereby, is not sensitive to saturation of synonymous sites (Li, Liu et al. 2009). Type I functional divergence is defined by a change in the selective constraint in a specific site after duplication, i.e. amino acid configurations that are very conserved in one gene but highly variable in the other gene, either by relaxation of existing purifying selection or by gaining functional importance at a previously unimportant site (Gu 1999; Gu 2001). In contrast, Type II represents amino acid configurations that are very conserved in both genes but whose biochemical properties are very different, e.g., charge positive versus negative, implying that these residues may be responsible for functional specification (Gu 2001). Type I and/or Type II divergence can occur as the result of either neofunctionalization or subfunctionalization. Here we tested the functional divergence for the pairs WAP65-1/WAP65-2, WAP65-1/HPX and WAP65-2/HPX, comparing mammals and the teleosts paralogs. The same analysis was not performed in the cartilaginous fishes (WAP65c) that were represented by only three sequences, not satisfying the recommend condition of the cluster to have four or more sequences (Gu and Vander Velden 2002). The sites contributing to the type-I functional divergence where pointed out using two different strategies; I) defining the cut-off value for the posterior probability after consecutively removal of the highest scoring residues from the MSA until the LRT of the coefficient of functional divergence becomes non-significant $p>0.05$, ii) defining an empirical cut-off of 0.8 in the posterior probability for each site, since the lowest value obtained in the previous criteria was 0.8. The estimated θ_I values for the pairs of cluster can be used to construct a matrix of functional distance (d_F) values. Given this matrix, a standard least squares method can be implemented based on the formula $d_F(A,B) = b_F(A) + b_F(B)$ to estimate the b_F for each gene cluster, where $b_F(x)$ is the functional branch length of a given gene cluster *x* (Gu 2003).

WAP65-1 and WAP65-2 Three-Dimensional Structure Modeling and Functional Analysis

The three-dimensional (3D) structure of WAP65-1 and WAP65-2 was predicted using the I-TASSER server (Zhang 2008) to obtain the 3D model of both paralogs in fish. The models were obtained using the *Dicentrarchus labrax* sequences of both the paralogs WAP65-1 [NCBI: ABL75414] and WAP65-2 [NCBI: DAA12504]. The model with the correct topology should have a C-score above -1.5, varying from [2;-5]. A higher value than 0.5 in the TM score means that the obtained topology is not random (Zhang 2008). MultiProt (Shatsky, Nussinov et al. 2004) was used to calculate the root mean square deviation (RMSD) after the superimposition of the obtained structures for WAP65-1 and WAP65-2. These two structures were also superimposed with the sequence [PDB: 1QJS] of the rabbit (*Oryctolagus cuniculus*). The protein structure around the heme group was obtained using the Accelrys Discovery Studio 3.1 software (AccelrysSoftwareInc. 2012). In order to access the functional/active sites we submitted the WAP65-1 and WAP65-2 models obtained from I-TASSER to the Partial Order Optimum Likelihood (POOL) server (Somarowthu and Ondrechen 2012). This is a reliable method applicable to proteins with novel folds, particularly when the obtained models have enough quality (Somarowthu and Ondrechen 2012). The architecture of the protein domains was characterized using the Simple Modular Architecture Research Tool (SMART) (Schultz, Copley et al. 2000) and we considered only E-values bellow 1.0 to increase the accuracy of the estimates.

5.4 Results

Sequences and annotation

We retrieved 20 sequences of the WAP65-1 gene from teleost fishes, 15 previously annotated from GenBank and ENSEMBL (12 and 3, respectively), and five new CDS that we manually annotated from EST databases (Appendices V: Table S3 and Figure S2). For the paralog WAP65-2, we retrieved 21 sequences from teleost fishes, 14 previously annotated from GenBank and ENSEMBL (10 and 4, respectively), and seven coding sequences were manually annotated from EST databases (Appendices V: Table S3 and Figure S2). Additionally, we retrieved three WAP65c sequences from cartilaginous fishes (two from databases and one additional sequence from EST databases). The gene WAP65 has been detected in the arctic lamprey (*Lethenteron camtschaticum*) (Appendices V: Figure S3), but due to the low coverage of the CDS, this sequence has not been used in further analyses. In total, 66 sequences were retrieved, three WAP65 from cartilaginous

fishes, 41 WAP65 coding sequences from teleosts (Appendices V: Table S3) and 22 representatives of the mammalian HPX (Appendices V: Table S4).

Phylogenetic Analyses

Phylogenetically, HPX has been identified in mammals (placentals, marsupials and monotreme species), birds, and amphibians (Dooley, Buckingham et al. 2010), while fishes present the ortholog *WAP65* and in some teleosts two copies of this gene are present. Here, we have studied the *WAP65* in fishes (cartilaginous and teleosts) and the ortholog HPX in mammals. The final MSA comprehending 66 sequences had a total length of 1,629 bp and was used to reconstruct the WAP65/HPX gene tree, with the best-fit evolutionary model selected by hierarchical likelihood ratio tests being the TIM2+I+G. The obtained tree topology with the ML, NJ and BAY analyses showed a clear distinction of the *WAP65-1* and the *WAP65-2* present in teleost fish (Figure 5-1). The *WAP65c*, *WAP65-1*, *WAP65-2* and *HPX* clades were well supported in the three phylogenetic reconstructions (ML, NJ, and BAY). However, the results from the KH, SH and ELW tests performed in TREE-PUZZLE suggest that the BAY and ML are significantly better than the NJ gene tree based reconstruction (Appendices V: Table S2). The obtained topology using BAY and ML was similar and well supported for the interior branches, although with some minor topologic differences at the terminal branches (Figure 5-1).

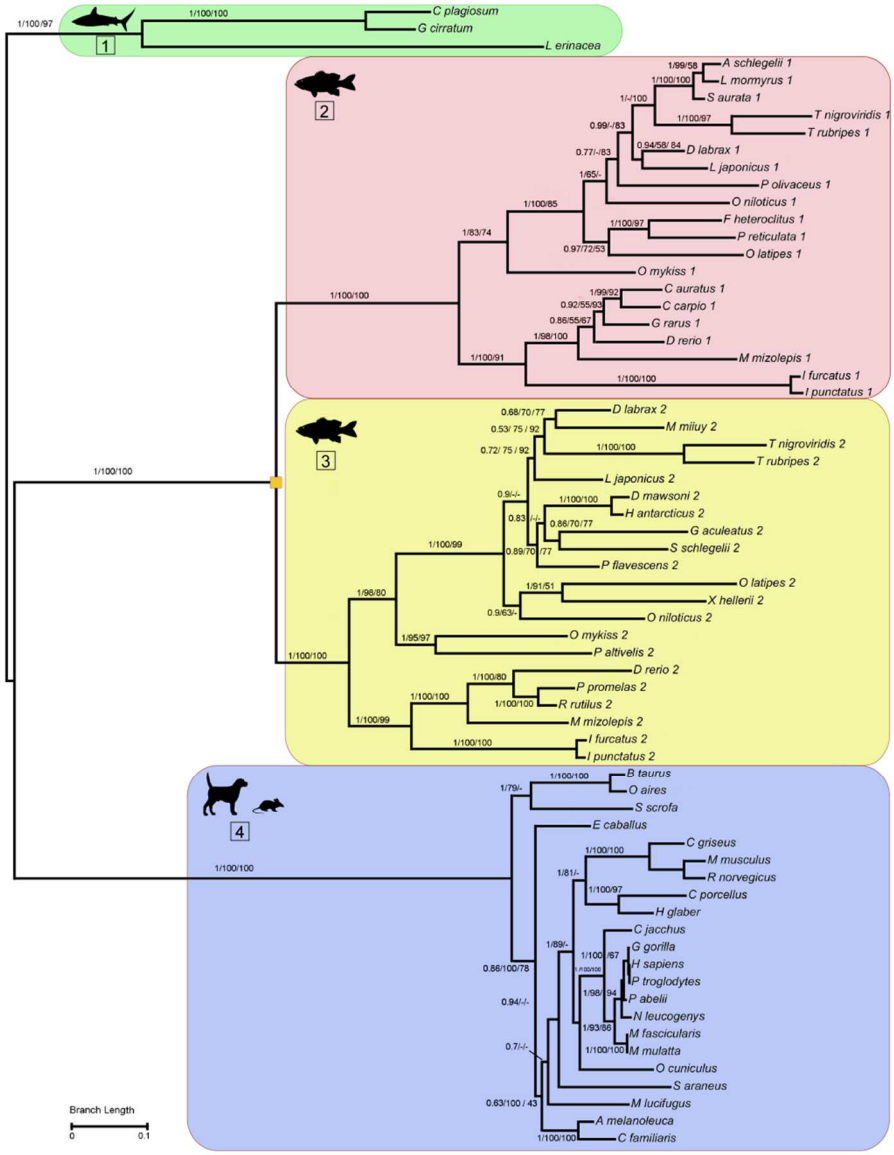

Figure 5-1. Phylogenetic tree of WAP65/HPX. The tree was obtained based on the nucleotide alignment of 66 WAP65/HPX sequences encompassing 1,629 bp. The numbers near the nodes indicate the branch support for the three different analyses (BAY/ML/NJ). The bootstrap values for ML and NJ analyses below 50 are represented by a "–". Each clade represents the different taxonomic groups: 1—cartilaginous fishes WAP65c; 2—teleosts WAP65-1; and s3—teleosts WAP65-2; 4 – mammalian HPX.

The obtained tree using *WAP65*/*HPX* retained a phylogenetic topology similar to the species tree, grouping together as expected the closest fish and mammalian Orders with a few exceptions, e.g. *Perca flavescens* did not grouped with the other representatives of the Perciformes, *Harpagifer antarcticus*, *Dicentrarchus labrax* and *Dissostichus mawsoni*. The Perciformes phylogeny and it relation with the other fish clades is not yet fully resolved (Near, Eytan et al. 2012), and often the gene based tree differs from the accepted species tree (Louis, Muffato et al. 2013). In the mammals clade, the Laurisatherians did not grouped within an independent clade as would be expectable, but that may have been influenced by the absence of other mammalian species that have diverged earlier (e.g Afrotherians,

Monotremes) in the final MSA. The more divergent mammalian species were removed given the high saturation introduced in the MSA (Appendices V: Table S4), and final analyses were performed including a mammalian dataset having no nucleotide saturation. The MSA of WAP65-1 and WAP65-2 also did not showed the presence of saturation (Appendices V: Figure S4) satisfying the criteria to access the selection signatures at the gene-level.

The full MSA presenting the 66 species used in this work was therefore divided in four different MSAs: (i) *WAP65-1*, (ii) *WAP65-2*, (iii) *WAP65-1*, *WAP65-2* and *WAP65c*, (iv) *HPX*. The three test performed in TREE-PUZZLE revealed that the BAY and ML fits well the data in all MSAs, but the same was not applicable to the NJ reconstruction, as the MSA (iii) *WAP65-1*, *WAP65-2* and *WAP65c* retrieved a significant lower likelihood when compared with the other two methods. Despite the differences between the ML and BAY trees were non-significant, the BAY tree obtained the best likelihood in all the three tests performed in TREE-PUZZLE and therefore it was used in further analyses (e.g. positive selection and functional divergence).

Selection in the post-duplication branch

The gene tree reconstruction based on the fish *WAP65* genes, (iii) *WAP65-1*, *WAP65-2* and *WAP65c*, was used to access selection signatures in the branches. The d_N/d_S ratios were estimated in a likelihood framework at a lineage-specific level, labeling the PD branch (the branch immediately after the duplication event) in WAP65-1 and WAP65-2. The obtained likelihood was compared with a model allowing only one value of d_N/d_S along the tree, and the LRT of the obtained value was compared with a chi-square table, with one-degree of freedom. In both PD branches the LRT showed no statistical significant difference between the one ratio and the two-ratio model, with a LRT=3.27 (p=0.07) in the WAP65-1 PD branch and a LRT=0.80 (p=0.37) in the WAP65-2 PD branch (Appendices V: Table S5). Given the absence of statistical significance between the two-ratio model and the one-ratio model in both branches we tested a branch-site model in the same branches, i.e. the branch immediately after duplication. We used the MSA that contain WAP65 and HPX genes in the branch-site analysis to include HPX in the background sequences. The branch-site analysis when labeling the WAP65-1 PD branch showed eighth sites with a PP>0.95, while the WAP65-2 PD branch fail to detect any site under selection (Table 5-1). The alternate model is significantly preferred relatively to the null model with a significance bellow 0.01. The results suggest that WAP65-1 and WAP65-2 did not undergo strong positive selection after the duplication, but instead only a few sites have had significant selection signatures in the PD branch, particularly in WAP65-1.

Table 5-1. Positive selection in branch-site model using WAP65-1 or WAP65-2 post-duplication branch. The sites under selection are referred to WAP65-1 and WAP65-2 in *Dicentrarchus labrax*, respectively, the sites highlighted in bold represents the sites with a Posterior Probability above 0.95.

Gene	Model	Parame-ters	*Lnl*	Parameters estimated	LRT	P-value	Positively selected sites
WAP65-1		135	-25554.0	p_0 =0.724; p_1 =0.173; p(2a+2b) =0.103; ω_0 = 0.16; ω_1 = 1.0; background: ω_{2a} = 0.16; ω_{2b} = 1; foreground: ω_{2a} = ω_{2b} = 55.90	20.6	<0.01	41V(0.85),47P(0.555), **56F(0.970)**,62K(0.574), 71A(0.713),**125S(0.998)**, 127V(0.603),152I(0.624), **174F(0.970)**,176S(0.552), **186F(0.953)**,187M(0.569), 197M(0.535),204K(0.935), **265K(0.955)**,**271T(0.987)**, 275D(0.949),276T(0.531), **281F(0.97)**,**297H(0.957)**, 313E(0.709),317H(0.54), 335I(0.583),365R(0.557), 377K(0.678)
	Null Model (fix omega)	134	-25564.3	p_0 =0.669; p_1 =0.161; p(2a+2b) =0.170; ω_0 = 0.143; ω_1 = 1.0; background: ω_{2a} = 0.159; ω_{2b} = 1; foreground: ω_{2a} = ω_{2b} = 1			
WAP65-2	Model A	135	-25567.0	p_0 =0.763; p_1 =0.182; p(2a+2b) =0.056; ω_0 = 0.161; ω_1 = 1.0; background ω_{2a} =0.161; ω_{2b} =1; foreground, ω_{2a} =ω_{2b} =998.99	7.2	<0.01	140T(0.806),183H(0.879), 184L(0.598),212S(0.873), 269M(0.815),281F(0.577), 285R(0.729),290V(0.602), 291T(0.734),318A(0.585), 357Q(0.838),384I(0.68)
	Null model (fix omega)	134	-25570.6	p_0 =0.773; p_1 =0.187; p(2a+2b) =0.040; ω_0 =0.161; ω_1 =1.0; background ω_{2a} =0.161; ω_{2b} =1; foreground, ω_{2a} =ω_{2b} =1			

Selection at the gene level

The selective pressure acting in the HPX and WAP65 is distinct revealing that the WAP65-1 gene presents the lower ratio of dN/dS, ω=0.18, WAP65-2 had a ω=0.21 and the HPX showed the highest value of ω=0.39 (Table 5-2). Although the overall ratio ω <1 suggests that most of the sequence is under purifying selection, we found evidence for positive selection for some sites exhibiting a high probability to be in the site class ω >1 using the likelihood phylogeny-based methods in PAML. In the CDS of the genes WAP65-1 and HPX, the M2a were statistically better than the null model (M1) (Table 5-2), although this model fails against the null model in the WAP65-2.

Table 5-2. Maximum likelihood analysis using CODEML models in WAP65-1, WAP65-2 and HPX. The six models (M0, M1a, M2a, M7, M8 and M8a) and the corresponding calculated likelihood. The obtained values for significance level are also shown, NS when the LRT is not significant for each pairwise comparison.

Gene	Model	*Lnl*	Parameters	Test	LRT	p-value
WAP65-1	Model 0: one-ratio	-11516.13	ω =0.18			
	Model 1: NearlyNeutral	-11269.58	p =0.83 0.17 ω =0.13 1.00			
	Model 2: PositiveSelection	-11265.28	p =0.82 0.17 0.01 ω =0.13 1.00 3.30	M1 vs. M2	8.6	0.01
	Model 7: beta	-11212.84	p =0.56 q= 1.82			
	Model 8: beta&w>1	-11194.86	p_0 =0.96 p =0.716 q =2.96 (p_1=0.04) ω =1.53	M7 vs. M8	36.0	<0.01
	Model 8a: beta&w>1	-11197.45	p0 =0.92 p =0.83 q =4.10 (p_1 =0.08) ω =1.00	M8 vs. M8a	5.2	0.02
WAP65-2	Model 0: one-ratio	-12495.58	ω =0.21			
	Model 1: NearlyNeutral	-12054.96	p =0.72 0.28			
	Model 2: PositiveSelection	-12054.96	p =0.72 0.24 0.04 ω =0.10 1.00 1.00	M1 vs. M2	0	NS
	Model 7: beta	-11970.71	p =0.34 q=0.97			
	Model 8: beta&w>1	-11964.58	p_0 =0.87 p= 0.48 q =2.40 (p_1=0.13) ω =1.01	M7 vs. M8	12.3	<0.01
	Model 8a: beta&w>1	-11964.53	p_0 =0.87 p =0.48 q=2.50 (p_1 =0.13) ω =1.00	M8 vs. M8a	0.1	NS
Hemopexin	Model 0: one-ratio	-9486.90	ω =0.39			
	Model 1: NearlyNeutral	-9297.12	p =0.68 0.32 ω =0.14 1.00			
	Model 2: PositiveSelection	-9292.60	p =0.68 0.30 0.02 ω =0.14 1.00 2.37	M1 vs. M2	9.0	0.01
	Model 7: beta	-9287.85	p =0.35 q=0.56			
	Model 8: beta&w>1	-9276.25	p_0 =0.91 p =0.48 q =1.04 (p_1=0.09) ω =1.65	M7 vs. M8	23.2	<0.01
	Model 8: beta&w>1	-9283.70	p_0 =0.77 p =0.65 q =2.56 (p_1=0.23) ω =1.00	M8 vs. M8a	14.9	<0.01

In the more parameters rich nested pair of models, the results showed that the M8 (the model allowing a site-class above 1) fits better the data than the null model, in all the three genes. For WAP65-1, WAP65-2 and HPX, the LRT with M8 vs. M7 had a p-value lower than 0.01. Under the M8 and the BEB post-hoc analysis, WAP65-1 had three sites (24A, 25A and 405L, using the *D. labrax* as reference sequence) showing a PP higher than 0.95, and one site supported with a PP above 0.95 also under M2a, the site 25A. Under M8 the WAP65-1 revealed also two additional sites showing a PP above 0.90, the sites 20Q and 22Q, which despite being bellow the strict criteria of a PP>0.95 may suggest that those sites could be under positive selection. Additionally, we compared the M8 vs. M8a to inspect if the site-class showing a ω>1 was significantly above the neutrality, and the LRT showed a p-value of 0.023. In WAP65-2 gene the codon-based models revealed only one site (residue 286A) in M8 under selection with a PP>0.95. Nevertheless, the comparison between the M8 vs. M8a is not statistically significant suggesting that this gene may be

evolving more under neutrality rather than under positive selection. In the mammalian HPX, the number of sites under positive selection in M8 is the highest among the three analyzed genes, showing a PP>0.95 in eight sites: 9V, 80K, 82V, 175G, 177M, 358D, 365I and 447A. However, none of these sites had a PP>0.95 under the M2a. The strict test of the two models M8 vs. M8a showed a p-value below 0.01, suggesting that the site-class above one is statistically higher than the neutrality and therefore these sites likely evolved under strong positive selection.

Selection at the protein level

The use of d_N/d_S ratios based models for detecting selection is generally not sensitive enough to detect subtle molecular adaptations since it not account the possibility that adaptation might result from few amino acids changes (McClellan, Palfreyman et al. 2005). In order to complement our previous results based on the codon-based models we performed the TreeSAAP analysis (Woolley, Johnson et al. 2003) using the same three MSA evaluated in the CODEML site analysis. This complementary method using the physiochemical amino acids changes may be necessary to properly access natural selection among sites within generally conservative protein-coding genes. We used an empirical threshold of three properties under selection to signalize a specific amino acid carrying signatures of selection. The results retrieved 20 sites with more than three properties under selection for WAP65-1 and two of those sites (24A and 405L, position referred to the *D. labrax* WAP65-1 sequence) were also reported under the M8 in the PAML analysis (Appendices V: Table S6). For WAP65-2, the result was different from those obtained in the codon-based analysis, with 16 new sites detected to be under selection at the amino acid level, less than those observed in the paralog gene WAP65-1. In the mammalian HPX 16 sites showed at least three amino acid properties under positive selection. In the two paralogs, WAP65-1 and WAP65-2, all the positively selected sites retrieved had at least one property under selection, supporting the codon-based analyses. Nevertheless, in the HPX three sites were supported by the codon analyses (two of each showed a PP above 0.95) but not by the TreeSAAP analyses.

Functional divergence of WAP65-1, WAP65-2 and HPX

Gene duplication-specific changes in the substitution rates referred to as type-I functional divergence, might reflect differences in the evolutionary rate at amino acid sites after gene duplication. Type II sites are those that are highly conserved in both clusters but are fixed for amino acids with different biochemical properties between sisters clusters,

implying that these residues are responsible for the functional differences between these groups. The functional divergence analysis was performed in DIVERGE v2.0 (Gu and Vander Velden 2002) showing the existence of functional divergence between WAP65-1 and WAP65-2 but also among this two paralogs and the HPX singleton. The higher type-I and type-II divergence was observed between WAP65-1 and HPX (Figure 5-2A).

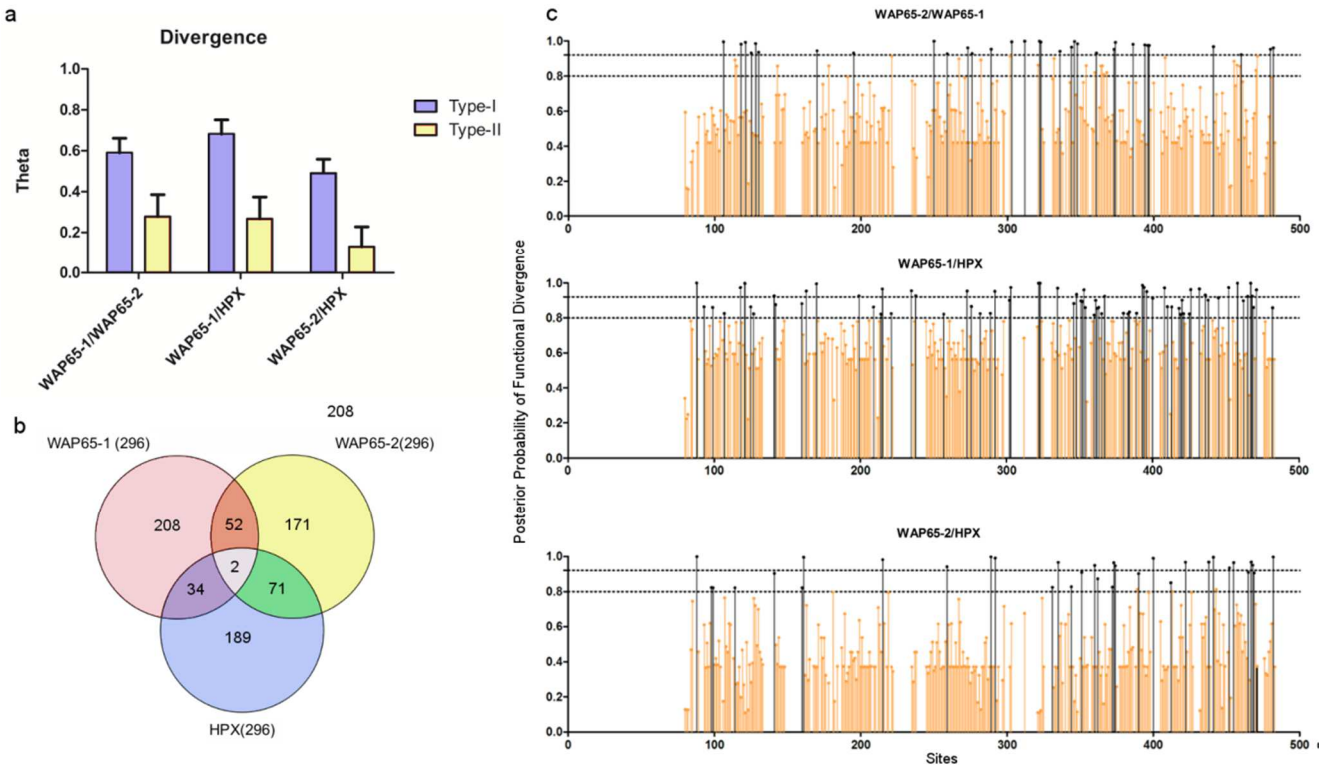

Figure 5-2. Functional divergence type-I and type-II. a) The Type-I and Type-II functional divergence of WAP65-1, WAP65-2, and HPX. b) Venn diagram of the number of sites above the posterior probability of 0.8 contributing to the functional divergence between the compared pairs. c) Site-specific profile of the pairwise comparison of the functional divergence obtained from DIVERGE 2.0, Type-I functional divergence in WAP65-1, WAP65-2, and HPX. The cut-off values for each pair comparison were defined after the sequential removal of the highest values till the acceptance of the null hypothesis.

The type-I divergence can be detected after a change in the site-specific rate following gene duplication, where either relaxation of existing purifying selection or the gain of functional importance at a previously unimportant site may occur (Gu and Vander Velden 2002), though a high number of sites changing the evolutionary rate (up or down) after the duplication imply functional divergence between the proteins (Zheng, Xu et al. 2007). The pairwise comparisons of WAP65-1, WAP65-2 and HPX for functional divergence type-I were significant with a p-value bellow 0.01. After removing the sites with the highest values of functional divergence until the p-value of the LRT in the divergence type-I is below 0.01, we defined the cut-off value for each pair of comparisons. For the HPX/WAP65-1 the critical

value was 0.8 showing 73 sites above this cut-off, for HPX/WAP65-2 was 0.82 with 33 sites above this value of PP, while for WAP65-1/WAP65-2 the calculated cut-off was 0.92 having 32 sites above this criterion (Figure 5-2B). Even when the cut-off is reduced to 0.8 in the other two pairwise comparisons the θ_I site-specific profile revealed the highest number above this limit in the WAP65-1/WAP65-2, showing 54 sites above this criterion relatively the 36 obtained in the WAP65-2/HPX comparison (Figure 5-2C). The functional divergence analysis for type II (θ_{II}) revealed a similar trend, HPX/WAP65-1 (0.268 ± 0.106), HPX/WAP65-2 (0.131 ± 0.097) and WAP65-1/WAP65-2 (0.280 ± 0.106), corresponding to a p-value of 0.006, 0.088 and 0.004, respectively (Figure 5-2C). Applying an empirical cut-off of 2 (representing sites with more than 66% of probability to be functionally divergent), the higher number of functional divergent sites where obtained for the WAP65-1/WAP65-2 comparison with a total of 78 sites, 73 for HPX/WAP65-1 and 28 for WAP65-2/HPX, although if using a cut-off of 8 (representing a probability above 88%), the same pair of comparisons would show 10, 1 and 2, respectively (Appendices V: Figure S5).

The WAP65-2 seems to have retained the ancestral function of WAP65 before the WGD in teleosts, while WAP65-1 became more functionally distinct as suggested by the bF functional distance values obtained in DIVERGE v.2.0 (Figure 5-3). Surprisingly, HPX also showed a high bF value suggestive of neofunctionalization and accumulation of significant differences relative to the putative ancestral function of WAP65.

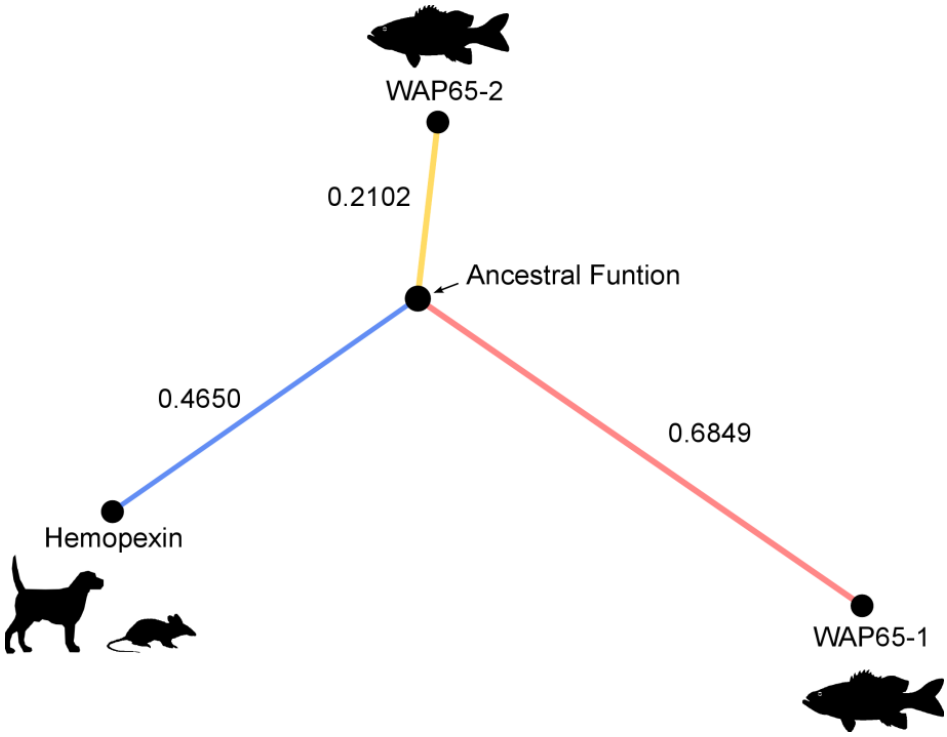

Figure 5-3. Schematic representation of functional distance between WAP65-1, WAP65-2 and HPX. The functional branch length (bF) was estimated using DIVERGE v.2.0 (Gu, 2001).

Modeling the tertiary structures of WAP65-1 and WAP65-2

The obtained protein 3D structure model of the WAP65-1 from *D. labrax* retrieved a TM-Score 0.79±0.09 and a C-score= 0.55, while the WAP65-2 of the same species obtained a TM-score 0.77±0.10 and 0.42 in the C-score. When the estimated models from I-TASSER are reliable, the C-score should be above -1.5, varying from [2; -5] and a higher value than 0.5 in the TM score means that the obtained topology is not random (Zhang 2008). We superimposed both structures based on the alignment of the two estimated 3D models, which showed a high structural similarity, also observed in the secondary structure with a high overlapping of both the structures (Figure 5-4).

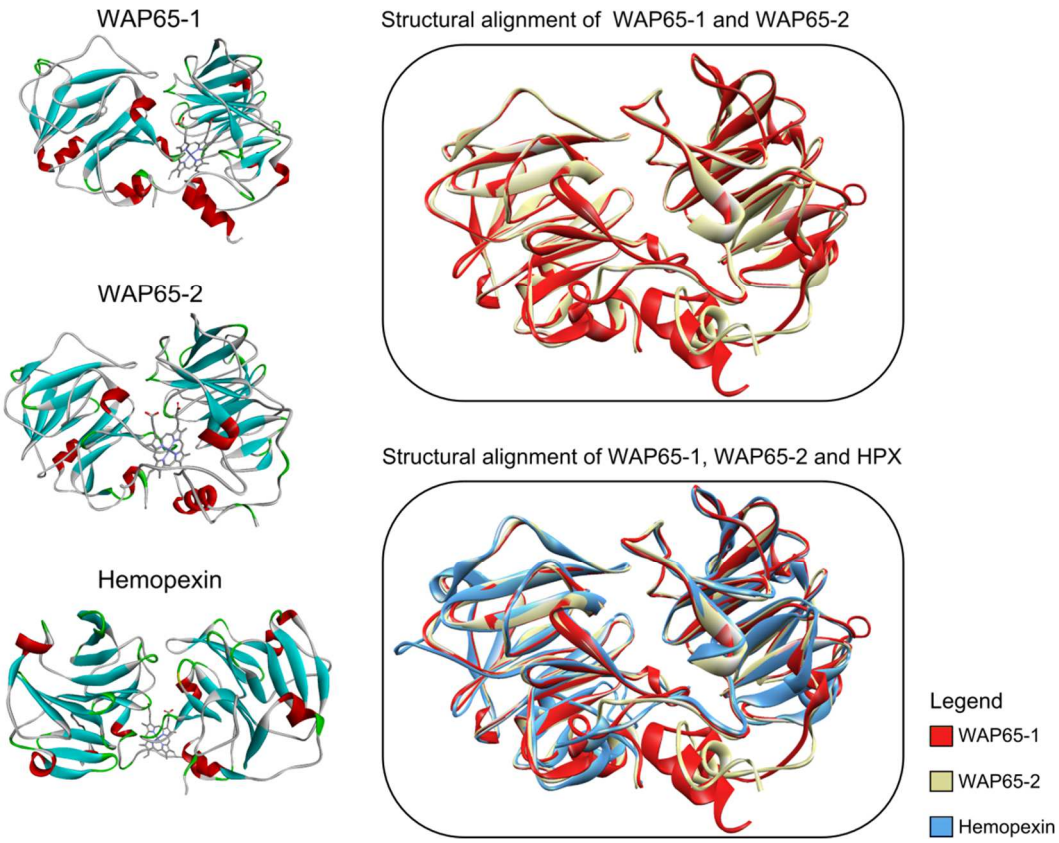

Figure 5-4. Structural similarity of WAP65-1, WAP65-2, and HPX. The two calculated 3D structures for D. labrax WAP65-1 [NCBI: ABL75414] and WAP65-2 [NCBI: DAA12504] were modeled in I-TASSER and superimposed with the mammalian HPX [PDB: 1QJS] using the MultiProt server (Shatsky, et al., 2004).

MultiProt evaluated the similarity of the sequences; the value obtained for the similarity between the two structures was 0.74 RMSD based on an alignment of 399 amino acids. Moreover, the superimposition of these two structures with the mammalian HPX also revealed a high structural similarly. The result suggests that the functional divergence is not due to an alteration of the structures despite the accumulation of different mutations in the paralogs. The amino acids responsible for the heme pocket in the mammalian HPX are already known, and we used a homology modeling strategy to infer the heme pocket in WAP65-1 and WAP65-2 in *D. labrax*. The amino acids coordinating the heme binding and the amino acids near the heme-binding site (<3.5Å) are also shown in the WAP65-1 and in for the WAP65-2 (Figure 5-5).

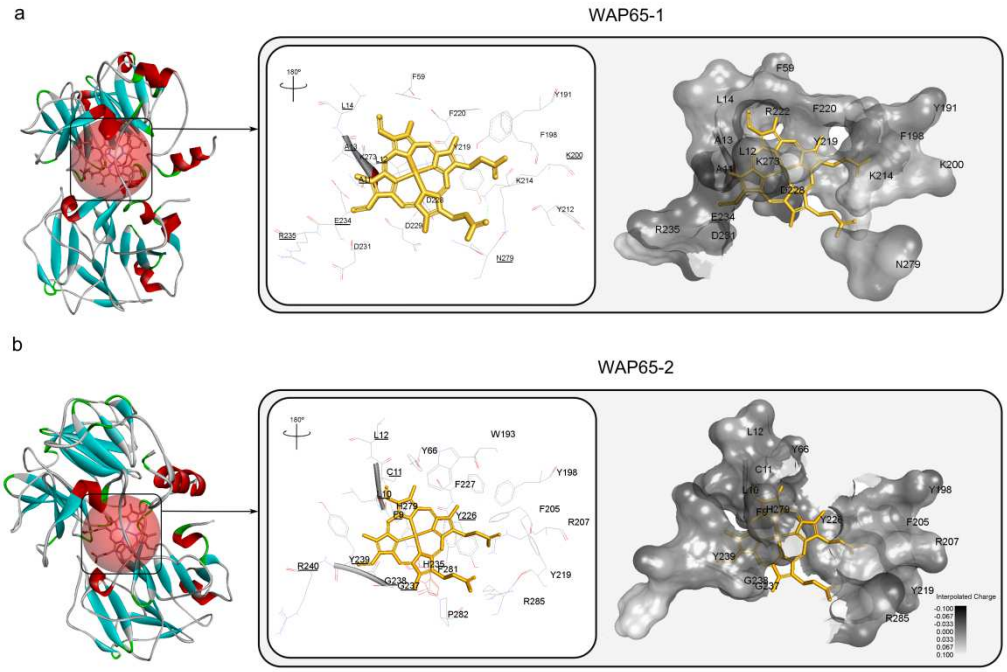

Figure 5-5. Heme-binding pocket structure in WAP65-1 and WAP65-2 in for D. labrax. Heme pocket: A-WAP65-1 and B-WAP65-2. In the left figure, it is represented as the heme-binding location for each paralog. In the center, the amino acids coordinating based on I-TASSER prediction and the amino acids on the proximity to the heme group (<3.5Å). For WAP65-1, the amino-acids in the right figure shows the protein structure around the heme group using Accelrys Discovery Studio 3.1 software (AccelrysSoftwareInc., 2012), the colors follow the interpolated charge scale.

Selection in the WAP65-1 and WAP65-2 heme-pocket and in other functionally relevant sites

We inspected the selection pressure and the functional divergence in the sites that are probably coordinating the heme binding of the proteins WAP65-1 and WAP65-2 and the residues near the heme group in those proteins. In the heme-binding pocket of WAP65-1 the sites binding to the free heme are under strong purifying selection, as well as the sites near the heme group, although the use of the amino acid level approach revealed three sites in the heme pocket to have one amino acid property under positive selection (219Y, 228D and 273K) and in the nearby site 234E (Appendices V: Table S6). In WAP65-2, we have not found positive selection in the 12 sites coordinating the heme binding, although site 279H showed a probability of 0.84 to belong to the site-class above one, suggesting that this site might be at least under relaxed purifying selection (no properties under selection have been nevertheless detected in these 12 sites). In the nearby sites, two sites (239Y and 240R) exhibited relaxed purifying selection, and five sites showed at least one amino acid property under positive selection (9F, 11C, 238G, 239Y and 240R) (Table 4).

The WAP65-1 and WAP65-2 from *D. Labrax* revealed 4 Hemopexin-like repeats (annotated as SM00120 in SMART), the four internal repetitions of these domains are located from the residues 81-133, 180-223, 243-286 and 288-333 for WAP65-1 and from the residues 89-140, 187-230, 249-292 and 294-339 for WAP65-2. The results from the POOL server revealed that three sites under positive selection in WAP65-1 were within the first 50 more functionally relevant sites (Appendices V: Table S7), and all the three residues (186F, 264R, 313E) located within the Hemopexin-like repeats (annotated as SM00120), and the sites 186F and 264R exhibited functional divergence type-I above 0.5. In WAP65-2 no sites was predicted to be within the first 50 more functional relevant, the nearest to fall in the category was the site 278L and showing functional distinction only relatively to WAP65-1 and HPX, was also similarly located within the Hemopexin-like repeats (Appendices V: Table S8).

5.5 Discussion

WAP65 duplication and the retention of a duplicated gene copy in teleosts

The HPX gene has been reported in different vertebrate lineages such as amphibians, reptiles, birds, and mammals. The ortholog WAP65 has been reported in fishes (Dooley, Buckingham et al. 2010), but in some teletost was found a duplicated copy of the gene (WAP65-1 and WAP65-2). The cartilaginous fishes only have one copy of the gene. Different studies considered WAP65-2 as an ortholog of the HPX, but this orthology have not been yet fully clarified (Dooley, Buckingham et al. 2010). Furthermore, there has not been any isoform of HPX reported in mammals or in cartilaginous fishes, suggesting that the duplication of the ancestral gene might have occurred after the divergence of teleosts and cartilaginous fishes. Here, we reported a putative WAP65 sequence in the artic lamprey, placing the emergence of the WAP65 in the jawless fishes. However, the orthology of the retrieved ESTs sequences remain to be fully clarified. Furthermore, we reported 20 out of 30 teleosts with the retention of a duplicated copy of WAP65. Consequently, it is important to understand the mechanism responsible by the retention in some teleosts genomes of a duplicated copy of this gene, which previous studies reported a fixation of the two copies to be more frequent in modern teleosts (Sarropoulou, Fernandes et al. 2010). It should be noted that the absence of the two copies of the WAP65 in some fish species might reflect also lacking of sequencing information in databases (ENSEMBL and GenBank). Interestingly, recent studies showed that some additional duplication occurred in the WAP65-1, namely in catfish (*I. punctatus*), where southern blot results indicated the presence of four WAP65-1 copies, but a single copy of WAP65-2 (Sha, Xu et al. 2008).

Therefore, after the WGD some additional tandem gene duplication occurred at least in the catfish.

Gene duplication has been described as one of the key factors driving genetic innovation, producing novel genetic variants (Conrad and Antonarakis 2007). A salient feature of the evolution of paralogous genes, is their divergence, which certainly involve accumulation of both advantageous and neutral or even mildly deleterious point mutations after the initial gene duplication (Ohta 1989). Under complete redundancy, where any number of functional copies confers the same fitness, selection on the paralogous genes will be relaxed (Cooke, Nowak et al. 1997; Nowak, Boerlijst et al. 1997). Several models of molecular evolution try to explain the factors contributing for the duplicated copies preservation and these models are different according to the dissimilarities in the evolutionary pattern after the duplication (Zhang, Wang et al. 2010). Additionally, it has been shown that "newly" evolved genes have an accelerated rate relatively to more ancient genes (Lynch and Conery 2000). Here, we reported that WAP65-1 exhibited an altered evolution rate, when compared with the paralogous gene, WAP65-2. Additionally, it has also been suggested that positive selection plays an important role in fixing specific amino acids in proteins after the main duplication events leading to the paralogous fixation in groups of a gene family (Martinez-Castilla and Alvarez-Buylla 2003). The preservation of the duplicated genes in the genome and partial functional relaxation caused by loss of ancestral functions subsequently provides the opportunity for advantageous mutations, which can lead to new functions (He and Zhang 2005).

Selective pressure acting on WAP65-1, WAP65-2 and HPX

An acceleration of the evolutionary rate may be implicated in the retention of both gene copies, and therefore may be the mechanism responsible for the retention of WAP65-1 and WAP65-2, contributing to the functional divergence of these paralogs. After duplication the genes may undergo functional divergence (Hahn 2009), and here we reported clearly a functional divergence between the two copies (WAP65-1 and WAP65-2). Accordingly, the neofunctionlization is usually supposed to include a stage of neutral evolution, with the functional divergence occurring in at least one step, including positive selection (Wagner 2008). It has been reported that WAP65-1 is expressed earlier in the development relative to WAP65-2 (Sarropoulou, Fernandes et al. 2010), and this different temporal expression is accompanied with a differential pattern of tissue expression. While WAP65-2 is only expressed in the liver, WAP65-1 is widely expressed (Sha, Xu et al. 2008). The differences in the expression pattern suggest that WAP65 paralogs underwent functional divergence, likely neofunctionalization (Sha, Xu et al. 2008), despite the difficulty

to distinguish between a model supporting subneofunctionalization and neofunctionilization. In addition, recent studies suggested that in the mud loach, *Misgurnus mizolepis*, the two paralogs underwent functional partitioning or subfunctionalization (Cho, Kim et al. 2012). Despite this fact, simulation studies suggest that subfunctionalization plays an important role, but as a transition state to neofunctionalization, rather than as a terminal fate of duplicated genes, since there is no apparent selective pressure to maintain redundancy and therefore the retention of duplicated genes in the genomes leads to neofunctionalization of the preserved copies (Rastogi and Liberles 2005). Furthermore it has been suggested that the WAP65-2 is associated with temperature adaptation and WAP65-1 is constitutively expressed (Sha, Xu et al. 2008), although previous works showed that the expression of WAP65-2 in the antarctic spiny plunderfish, *Harpagifer antarcticus*, is not up-regulated when the water temperature rise, suggestive that this acclimation function of WAP65 is phylogenetically constrained (Clark and Burns 2008). Indeed, we found that HPX also shows functional distinction (type I and type II functional divergence) relatively to both WAP65-1 and WAP65-2, but the lowest value obtained resulted from the comparison with the WAP65-2. This is in accordance with previous findings suggesting that the WAP65-2 is functionally similar to the mammalian HPX (Sha, Xu et al. 2008).

For both WAP65-1 and WAP65-2 no statistical evidence of relaxed purifying selection or positive selection was found in the PD branches. Duplicates that are being retained over long evolutionary time are more likely to experience strong purifying selection (Steinke, Salzburger et al. 2006). However, the p-value of the branch model in the case of WAP65-1 was near acceptance, 0.07, while in the WAP65-2 was near 0.37. This implies a different and asymmetrical evolutionary rate between the paralogs after the gene duplication. Indeed, the branch-site models also suggest that a few sites present signatures of selection and not the majority of the protein, making this approach more reliable to detect the episodic mode of evolution of these genes, WAP65-1 and WAP65-2. It can therefore expect that those sites might be implicated in the retention of the two copies after the duplication. Additionally, it was demonstrated that the measurement of positive selection is a powerful tool to identify divergence rates of duplicated genes and that this method has capacity to identify potentially interesting candidates for adaptive gene evolution (Steinke, Salzburger et al. 2006).

The selection signatures observed in WAP65-1, WAP65-2 and HPX could be a result of both functional adaptation and also might be implicated in the functional divergence among WAP65-1, WAP65-2 and HPX, where the accumulation of beneficial mutations leads to functional divergence. A previous work reported that WAP65-1 and WAP65-2 evolution is mainly due to purifying selection (Sha, Xu et al. 2008). More recently, (Sarropoulou, Fernandes et al. 2010) showed that the two paralogs are under moderate

positive selection suggestive of their evolutionary adaptation. However, this has been based in preliminary assessment of a small dataset (10 sequences of WAP65-1 and eight of WAP65-2), while here we reported 20 and 21 sequences for each paralog, respectively. By contrast, our data revealed a significant high number of positive selection events in WAP65-1 at codon level even after the more robust post-hoc test, the BEB analysis, and we point out also a high number of sites showing selection signatures at the amino acid level. Indeed, we implemented the same used M8, but we found several positively selected sites in WAP65-1 and two being highly significant, while in WAP65-2, despite the M8 fits better the data than the null model, the comparison between the M8 vs. M8a is not accepted, suggesting that WAP65-2 is evolving more under relaxed purifying selection rather than positive selection. This suggests that selection pressure follows different patterns in the two paralogs, while WAP65-1 was evolving slightly accelerated after the gene duplication (leading to a higher accumulation of non-synonymous mutations), in WAP65-2 the purifying selection or relaxed purifying selection was more influential in the evolution of this gene copy. Along with the asymmetrical evolutionary rate of the paralogs it is also relevant that HPX showed a higher d_N/d_S ratio when compared with the WAP65-1 and WAP65-2. It has been suggested that gene duplication have two trends, post-duplication acceleration and the generally slow evolutionary rate owing to the high level of functional constrains (Jordan, Wolf et al. 2004) and accordingly we reported a higher evolutionary rate in the mammalian singleton relatively to the duplicated copies in the teleosts.

The integration of the amino acids models suggests that many amino acids underwent positive selection at least in the physio-chemical properties in WAP65 paralogs. The codon-models are known to perform poorly when a substitution occurred only in a few species making these models conservative when applied to proteins that are subjected to purifying selection in the majority of the coding sequence. Here, we found that 16 sites showed at least three amino acid properties under selection in WAP65-2, but none of these sites correspond to the site showing significant positive selective pressure under M8 in CODEML. While WAP65-1 showed 20 sites under selection at the amino acid level (with more than three properties under selection) and two of those 20 sites also have been signalized as carrying signatures of selection at the codon level (24A and 405L, positions referred to the *D. labrax* WAP65-1 sequence). It is expectable that those sites might be of crucial relevance for the functional divergence of the two genes. The mammalian HPX have the higher amount of positively selected sites at codon based level but did not shown a significantly higher number at the amino acid level relatively to the WAP65-1 and WAP65-2, presenting 16 sites under selection at the amino acid level but only one of those sites is also signalized at the codon level (358D, position referred at the *Homo sapiens* HPX sequence).

Functional divergence and positive selection

The WAP65-1 protein showed five sites under positive selection that are contributing to the functional divergence relatively to the pairwise comparison with mammalian HPX (2 sites) and WAP65-2 (1 sites), and two sites are contributing to the functional divergence between both. While WAP65-2 shows six sites positive selected having a posterior probability above 0.8 of contributing to the functional divergence in the pairwise comparison with WAP65-1, but no sites under this criteria are signalized in pairwise comparison between HPX and WAP65-2. These results suggests that positive selection should have been important to drive the functional divergence between WAP65-1 and WAP65-2, but also between WAP65-1 and HPX, while nearly absent in the pairwise comparison between HPX and WAP65-2.

Structural similarity

The cysteine residues have a crucial contribution to the structural integrity of Hemopexin (Takahashi, Takahashi et al. 1985). Here, we reported that the model obtained for the WAP65 proteins shows that the paralogs are structurally similar and this 3D model similarity is of great relevance to understand the conservation of five cysteine residues between fishes and mammals. The 3D structures obtained for WAP65-1 and WAP65-2 were therefore topological similar, showing that positive selection and functional divergence is not causing considerable conformational divergences of the two 3D structures, even so the amino acid similarity of the two paralogs is only 58% in *D. labrax*. Remarkably, both WAP65-1 and WAP65-2 paralogs are also structurally similar to the mammalian HPX, despite the functional distinction among the three proteins.

It has been shown that recombinant rainbow trout hemopexin-like protein could bind to the free heme despite lacking the two histidines residues required for mammalian hemopexins to bind the free heme (de Monti, Miot et al. 1998). Similarly, the predicted binding residues in WAP65-1 did not shown these two histidines that are reported as essential in mammals to coordinate the heme binding. In the fish WAP65-1, two highly conserved histidines are present in the orthologs but in different evolutionary positions relatively to mammals. These two histidines highly conserve are in the positions 232 and 261 of the WAP65-1 in *D. labrax*. In WAP65-2, two histidines are present that align with the mammalian residues responsible for the heme coordination, in the positions 235H and 279H corresponding to *D. labrax* WAP65-2. Although it has been reported that the medaka, *O. latipes*, WAP65-2 revealed no affinity to the heme binding (Hirayama, Kobiyama et al. 2004), we did not found any evidence that one of the copies have lost the binding ability to

the free heme in *D. labrax*. The loss (or position alteration) of the two histidine residues in WAP65-1 coordinating the heme binding lead to predict that WAP65-1 protein bind to heme in a different manner from that of the mammalian HPX (Hirayama, Kobiyama et al. 2004). Indeed the heme-binding pocket seems to be under purifying selection and few sites showed functional divergence in the cluster of sites coordinating the ligand binding, suggestive that the functional distinctiveness between the paralogs is not due to any alteration in the binging ability although it might alter the overall affinity to the free heme.

The positive selection appears to have and important role in the functional divergence between the paralogs WAP65-1and WAP65-2. However, the retention of a duplicated copy in just some of the teleosts (20 out of the 30 teleosts here reported) suggests that the increase of fitness did not occur for all the species, as some have lost one of the duplicated copies. It will be therefore of great relevance to inspect the selective loss of the duplicated copy in some of the teleost species, as the paralogs perform different functions, and it would be of great interest to relate such events with the evolutionary history of the species and the different environmental conditions influencing the species fitness.

5.6 Conclusions

In this study, we assessed the evolutionary history of WAP65 in fishes and HPX in mammals. Statistical analyses of selection signatures suggest that positive selection and relaxed purifying selection have played important roles over evolutionary time in shaping the variation not only of the two paralogs (WAP65-1 and WAP65-2) but also the mammalian HPX. In contrast to other genes duplicated during the fish WGD, we detected a higher evolutionary rate in the mammalian singleton relatively to the teleosts paralogs. The detection of functional divergence between the fish paralogs and also the mammalian ortholog confirmed that these genes have evolved into different functional properties owing to rate shift of a small set of amino acids, which may explain the retention of the two WAP65 copies after the gene duplication in teleosts, as well as the overall functional divergence among WAP65 genes and HPX. The WAP65-2 seems to have retained the ancestral function of the protein, while the WAP65-1 underwent a higher functional divergence, suggestive of neofunctionalization or subneofunctionalization after the gene duplication. Indeed, we confirmed neofunctionalization of the two paralogs and pinpointed the sites that contributed to the functional distinctiveness of the two copies. We assessed by homology modeling the heme-binding pocket in both paralogs for *D. Labrax*, and both proteins seem to have retained the ability to bind to the free heme. The positively selected sites and those sites contributing to functional divergence between the paralogs are located outside the

heme pocket suggesting that the paralogs functional divergence and the preservation of both copies is not related with changes in the ability to bind the free heme.

5.7 Acknowledgements

JPM was funded by the PhD grant (SFRH/BD/65245/2009) from the Portuguese Fundação para a Ciência e a Tecnologia (FCT). AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490). This work was further supported by a grant from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism. We thank the Associate Editor, Dr. Stephen O'Brien and two anonymous reviewers for providing helpful comments to improve an earlier version of this manuscript. JPM thanks Marisa Silva from LEGE/CIIMAR for the careful read of the manuscript.

Chapter 6 - *Processed pseudogenes close to parent gene, or under a favorable expression context have higher chances to be functionally relevant*

6.1 Abstract

Pseudogenes were primarily described as “junk” DNA, dead copies of fully active genes that are present in the genome of different species. Two main different classes of pseudogenes have been described: processed and non-processed. While processed pseudogenes (*PΨgs*) are formed through retro-transposition, the non-processed typically arose after the decay following events of gene duplication. Some processed copies are expressed and since transcription is a costly process this raise the interest to understand their potential function. Given the persistence of *PΨgs* in mammalian genomes, here we surveyed annotated *PΨgs* from 18 mammalian genomes to assess their potential adaptive value. Indeed, the evolutionary analyses revealed evidences of positive selection in 40 out of 104 *PΨgs* (~38%). For five species, were retrieved 33,256 *PΨgs*, 29,563 were disabled i.e. premature stop codon. Mapping the chromosomal location of the *PΨgs*, and assuming the presence of stop codons as signatures of nonfunctionality, we found that those who are retro-transposed to the same chromosome of the parent gene have higher chances of potentially remain active (22%), while those that are allocated into a different chromosome have lesser changes to be active (10%). Additionally the location where the *PΨgs* are inserted seen associated to their potential expression, since those inserted in intragenic regions and in same strand of the “receiver” gene are more likely to be expressed. The origin/destination analysis of *PΨgs* shows that the larger chromosomes accommodate more *PΨgs*. While the analyses of the *PΨgs* inserted in the same chromosome of the parental gene, suggests a higher probability to be located closer to the parent gene than expected under a completely random process. Overall, our results provide evidences that natural selection action on *PΨgs*, influencing their distribution to be closer to the parental genes likely to a lower decay rate. While the expression of *PΨgs* is linked to a favorable expression context. Furthermore, those *PΨgs* that remain potentially active are subject to similar selection events as the coding sequences counterparts of fully active genes. Therefore, *PΨgs* do seen to have a function in adaptation processes, despite their fully role still remain unclear.

6.2 Introduction

In 1977 was first reported a structure termed as “pseudogene”, from a oocyte-type 5S RNA in *Xenopus laevis* (Jacq, Miller et al. 1977). Since them, several pseudogenes were described in a wide range of life forms from bacteria to vertebrates (Mighell, Smith et al. 2000). Pseudogenes are often described as non-functional sequences of genomic DNA

sequences, originally from functional genes (Balakirev and Ayala 2003). Two main processes may lead to the origin of pseudogenes from functional copies: (1) retro-transposition leading to the lack of a promoter ("dead on arrival") and (2) decay of functional genes (frame shifts and/or premature insertion of stop codons), mostly from duplicated copies but also from non-duplicated copies (Khachane and Harrison 2009). Therefore the pseudogenes are often separated into two different broad classes: (i) *PΨgs*, corresponding to those that have been retro-transposed back into a genome via an RNA intermediate; and (ii) nonprocessed pseudogenes, which are genomic remains of duplicated genes or residues of dead genes (Zheng, Frankish et al. 2007). Since they are often described as nonfunctional genomic fragments, and commonly referred as “defunct” genes, it is assumed that they evolve under neutrality (Martinez-Arias, Mateu et al. 2001). Although their presence might be functionally relevant supported either by transcription data and an unusual high conservation (Harrison, Zheng et al. 2005; Svensson, Arvestad et al. 2006).

Several estimates about transcribed pseudogenes number have been published in recent years (Harrison, Zheng et al. 2005; Frith, Wilming et al. 2006; Zheng, Frankish et al. 2007). It has been proposed that their abundance in genomes is associated and dependent on gene duplication and loss rate (Podlaha and Zhang 2010). Mammals have a high number of *PΨgs*, estimated to be around 8,000 (Zhang, Harrison et al. 2003; Zhang and Gerstein 2004). Two recent works have proposed that in the human genome 1286 out of 7849 *PΨgs* (~12.5%)(Navarro and Galante 2013), and 615 out of 4927 *PΨgs* (~16.4%) are expressed (Kabza, Ciomborowska et al. 2014). Contrary to the theoretical expectation that any gene can form a pseudogene, differences have been reported in the relative proportion, from gene to gene (Li, Yang et al. 2013). Housekeeping genes, highly expressed genes in germline cells and those participating in basic metabolic regulations show multiple corresponding pseudogenes (Frederiksen, Cao et al. 1997; Zhang and Zhang 2003; Zhang and Gerstein 2004; Pei, Sisu et al. 2012). This phenomenon may be due the high expression of these genes, which therefore can increase their likelihood to accumulate mutations (Park, Qian et al. 2012) or be retro-transposed back to the genome (Poliseno, Salmena et al. 2010). In addition to the expression level, the GC content of the genomic region where the pseudogenes were deposited will also affect the accumulation rate of mutations (Bustamante, Nielsen et al. 2002).

It is generally assumed that pseudogenes are nonfunctional units. However, previous reports proved that their sequence conservation strongly suggests a functional role. Remarkably *PΨgs* shown that their prevalence in the genomes have an important adaptive value in mammals, since here were reported sites falling in the category with an estimated $\omega > 1$, and previous works supported a main role of *PΨgs* as gene expression regulators. Here is shown that they are subjected to nonrandom processes such as the differential

retention as stop codons release, depending either on their destination and closer location to the parent gene as major determinants to decrease the rate of frame-shifts insertions.

6.3 Methods

Pseudogenes retrieval and identification

Pseudogene.org

The pseudogenes annotated in the pseudogene.org database were retrieving for: *Canis familaris* (dog, build 50), *Homo sapiens* (human, build 74), *Mus musculus* (mouse, build 74), *Pan troglodytes* (chimpanzee, build 50), and *Rattus norgevicus* (rat, build 74). The pseudogenes were divided relatively to their formation. In pseudogene.org were retrieved genes annotated as *PΨgs*, after excluding pseudogenes categorized as “ambiguous” and “duplicated”. Data were retrieved for the genes in mammals present in the database pseudogenes.org: *Canis familaris* (6001 out of 6126), *Mus musculus* (7923), *Homo sapiens* (8283 out of 9336), *Pan troglodytes* (7098 out of 7505) and *Rattus norgevicus* (7099). Performing 36,404 pseudogenes, and reduced to 33,256 after discarding those associated with mitochondrial DNA and those with inconclusive identity of the parental gene.

Ensembl

In *Ensembl* database pseudogenes annotated through meeting one the following criteria: 1) single exon gene and high similarity with a multi-exon transcript elsewhere in the genome; 2) transcript fully covered by repeat masker; 3) presence of multiple frame-shifts and no introns; 4) transcript contains multiple frame-shifts and the introns are >80% covered by repeat masks (Birney, Andrews et al. 2004). The pseudogenes sequences were retrieved through *Ensembl* Biomart (Kinsella, Kahari et al. 2011), and latter *PΨgs* have been separated from non-processed pseudogenes.

From the *Ensembl* v73 database, were retrieved pseudogenes from 18 mammalian species performing a total amount of 1871 sequences, annotated either as *PΨgs* or processed transcripts. The list of species and the number of available *PΨgs* encompass: *Ailuropoda melanoleuca* (9), *Bos taurus* (166), *Callithrix jacchus* (333), *Canis familiaris* (128), *Cavia porcellus* (14), *Equus caballus* (61), *Felis catus* (23), *Gorilla gorilla* (139), *Ictidomys tridecemlineatus* (18), *Loxodonta africana* (9), *Myotis lucifugus* (131), *Mus musculus* (254), *Nomascus leucogenys* (69), *Oryctolagus cuniculus* (109), *Otolemus garnetti* (142), *Pan troglodytes* (63) and *Rattus norgevicus* (203). Those *PΨgs* were identified through blast searches when the parent gene was unspecified in *Ensembl*. Those undefined PΨgs were

identified after their submission to BLASTn and BLASTx searches in NCBI. Searches in NCBI databases were performed with an "*in-house*" script through BLASTn to query the nucleotides database ("*Nucleotide collection*") and TBLASTx to search the protein database ("*Swissprot*"), using as threshold the cut-off, $10e^{-10}$ for the E-Value.

Expression of processed pseudogenes

The data from *PΨgs* expression in humans were retrieved from *RetrogeneDB* (Kabza, Ciomborowska et al. 2014) and *RCPedia* (Navarro and Galante 2013). In RetrogeneDB, based on *Ensembl* annotation, were retrieved information relatively to those expressed and showing conservation of the Original Reading Frame (ORF) (188 pseudogenes). While in *RCPedia*, based on University of California Santa Cruz (*UCSC*) annotation (Karolchik, Hinrichs et al. 2009), were retrieved all the 7849 pseudogenes assembling the information about their expression (1286 pseudogenes) and their genomic context (position relatively to other genes).

Pseudogenes concatenation and multiple sequences alignment

For genes represented by more than two sequences we built a multiple sequence alignment. This initial filtering procedure allowed the accommodation of 1798 genes under this criterion, while the remaining has been excluded from further analysis. The sequences have been aligned with MUSCLE (Edgar 2004; Edgar 2004) implemented in SeaView v4.4.2 (Gouy, Guindon et al. 2010) using protein sequences after translating nucleotides to amino-acids and the default parameters. Later the sequences were back-translated to nucleotides. For those multiple sequences alignment matching the criterion: ≥ 4 sequences and ≥ 2 different species, a phylogenetic tree was built using PhyML v3.1 (Guindon, Delsuc et al. 2009) implemented in SeaView v4.2, with aLRT as branch support (Anisimova, Gil et al. 2011). Redundant sequences (i.e. sequences with no dissimilarity at any nucleotide) provided by the two databases were removed at this stage. The total number of genes matching these criterions was 104 genes. Later those genes were used in the selection analysis, based in the previously obtained multiple sequence alignment and the topology estimated from the Maximum Likelihood tree.

For the *PΨgs* inserted in the same chromosome of the parent gene was built a pairwise alignment. Each parent gene was retrieved with *Ensembl* Biomart and aligned with TranslatorX (Abascal, Zardoya et al. 2010), applying a similar methodology as described above, translate to nucleotide, and back-translate to nucleotides, using the MUSCLE as alignment algorithm. The produced alignment was further used in YN00 implemented in

PAML v4.7 (Yang 2007), weighting=0 and common 3x4 = 0. The distance between the *PΨgs* and the parental genes has been used to rank from the closest to the more distant genes. A sliding windows analysis was performed using a window=10 genes and step-wise=3, *PΨgs* without stop codons were previously removed. The same analysis has been repeated for those *PΨgs* disabled but removing the sequence downstream appearance of the first stop codon.

Selection analysis

The positive selection analysis of the *PΨgs* matching the criteria described above (≥ 5 sequences and ≤ 2 sequences) was used to test the presence of selection signatures, using CODEML implemented in PAMLv4.7 (Yang 2007). The CODEML implement models that allow variation in d_N/d_S ratios along the codon sequences. To attain this purpose a maximum-likelihood approach is useable to test the significance of the presence of sites with a $d_N/d_S > 1$. This requires the use of two different models: 1) allowing a site class >1 (alternate model) and 2) restricting all sites ≤ 1(null model), while the likelihood of obtained for each model are compared using a likelihood ratio test (LRT). The LRT calculation were obtained after twice the difference of the log *likelihood* obtained between the alternate and null model (also called nested models). The significance of the difference between the models is obtained using a χ^2 distribution and considering the degrees of freedom as the difference in the number of parameters inherent in each model. A significant LRT (p-value ≤ 0.05) suggest that the alternate model (model allowing sites with a $d_N/d_S > 1$) is preferred relatively to the null model. Here we used a neutral model (M0) to access the single d_N/d_S ratio averaged across all sites to access the global d_N/d_S. The null model (M7) with the d_N/d_S ratio across sites calculated assuming as a beta distribution was compared to a selection model (M8) that add one additional site class with a d_N/d_S (allowed to be ≥ 1). After the pairwise comparison, we used a modified version of model 8, the model 8a, that restricts the site class (allowed to be above one) but in this model is forced to one. This test is particular relevant to distinguish from positive and neutral selection, therefore highly relevant to those comparisons (M7 vs M8) that have a significant LRT. To access the quality of the analysis and possible bias errors we have performed the same analysis using three methods (1-Filtering, 2-Removal of aberrant sequences, 3-Removal of possible frame-shifts);

1) The same analysis was re-done using the filtering tool, GBLOCKS (Talavera and Castresana 2007) to remove the positions in the more “gappy” alignment and those with more than 50% ambiguities and/or gaps. The remaining parameters were left as *default*. The alignment and phylogenetic tree were re-build using the same parameters without filtering. CODEML models were re-run using the new multiple

sequence alignment and phylogenetic tree estimation;

2) Furthermore we tested the amino acid composition in TREE-PUZZLE (Schmidt, Strimmer et al. 2002) that implements a chi-square test to distinguish relatively average base compositions, which taxon is identical and deviating taxa. The test considers sequences as homogenous those that obtained a score above 5%, but we consider a stricter empirical threshold of 25%. Those sequences considered as aberrant were removed, the alignment and tree were re-done using the same parameters as described above;
3) Removals of possible frame-shifting events were done using MACSE (Ranwez, Harispe et al. 2011) alignment. The character marking frame-shifts ("!") were replace with ("-"), the alignment obtained after the replacement was further used to re-build the phylogenetic tree and the evolutionary rate calculated in CODEML was re-done for those multiple sequence alignments presenting frame-shift errors.

Insertion simulation

For the genes inserted in the same chromosome of the parental gene, we conducted a simulation study to test if the observed distance between the *PΨgs* and the parent gene were expected under completely randomness or biased/skewed. To accomplish this test we used the method implemented in java "Math.random", following the steps: 1) generate randomly a chromosome number; 2) predict randomly an original position; and 3) random insertion of a gene in the same chromosome. The same simulation was conducted but using the original positions for each species and simulating only the random insertion location in the chromosome. For both scenarios was measured the distance for the parent gene (considered here as the center point in the genes, half-distance to both extremities). Both analyses were repeated three times to inspect the consistency in the obtained data. While the majority of the processed pseudogenes arose through retro-transposition events, some arose from duplication of previously retro-transposed genes, and this events leads to duplicated-processed pseudogenes. To avoid possible bias introduced by segmental duplication regions, highly rich in processed pseudogenes (Khurana, Lam et al. 2010). The analyses of the observed distances between the parent gene and the corresponding retro-copy were re-done after discard those *PΨgs* falling in that regions. Data from the processed pseudogenes and segmental duplication were retrieved at http://pseudogene.org/sdpgenes (Khurana, Lam et al. 2010). Since the data from pseudogenes.org was built in the assembly hg19, the corresponding coordinates were converted from hg18 to hg19 using the liftOver tool at UCSC (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Since the *PΨgs* were retrieved using different methods their coordinates did not perfectly match. To allow comparison we

built a script finding the middle point of the *PΨgs* gathered from http://pseudogene.org/sdpgenes and searching if that point was within the boundaries of the *PΨgs* annotated in pseudogenes.org, and those matching this criteria were considered the same *PΨgs*.

Gene evolution under neutrality

Pseudogenes are assumed to evolve under neutrality. If this prediction is correct then the reminiscent sequences of previous functional copies will "accept" all the mutations. The solely difference rely in the ratio of transitions and transversions. It is expected that the transitions double the number of transversions, and this is an explanatory model for sequences evolving under neutrality (Wakeley 1996). Here we tested the time and the mutation insertion effect of pseudogenes evolving under the emulated Kimura model (Kimura 1980), with the parameters of the model settled to a ts/tv (kappa value) = 2 which is a value frequently observed in mammalian genomes (Wakeley 1996). In each step, two random mutations were inserted, with doubled chances to insert a transition rather than a transversions to a maximum of 400 steps (800 mutations). For each step the sequences were submitted to blast searches (either BLASTn or BLASTx) and the genes were considered lost when the first of the following criteria were achieved: 1) the expected value for the gene of interest is below $1e^{-10}$; or 2) intrusion of an unrelated gene in the top-hit result. The mutation rate and the predictive time for gene loss were calculated under the assumption of 2.41 mutation/site/year (averaged value observed for fossil record and sequences) (Kumar and Subramanian 2002).

6.4 Results

Distribution in the Chromosomes

The *PΨgs* seen randomly distributed in the chromosomes, since it is observed a correlation between length and number of pseudogenes in each chromosome (r^2=0.552) (Figure 6-1). This phenomenon has been previously observed for human genome and has been described as "bombardment" along evolution (Zhang, Harrison et al. 2003). In the human genome, previous works have shown an association between chromosomes length and number of *PΨgs* (Zhang, Harrison et al. 2003). This seem a generalization in mammals since is evident a correlation between the chromosome length and the number of *PΨgs* incorporated among the five mammalian species studied (Figure 6-1). On the other hand

relatively to their origin, the correlation between number of *PΨgs* and chromosome length is nearly absent (r^2=0.156) (Figure 6-1).

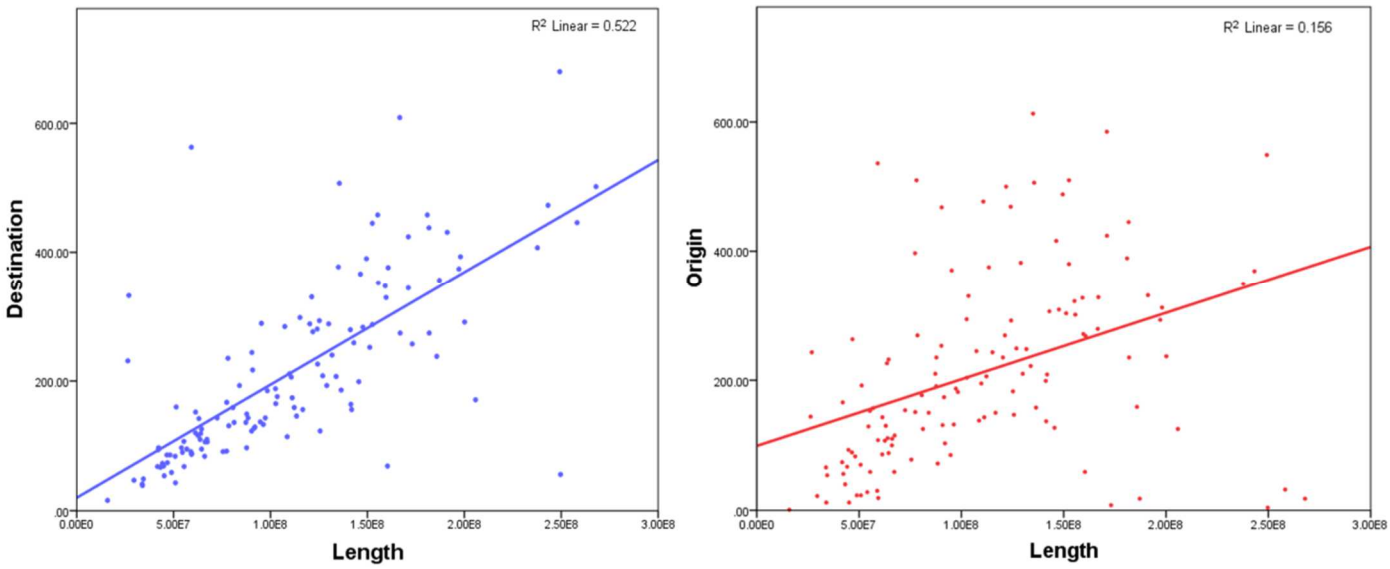

Figure 6-1. Relation between the chromosome length and the number of *PΨgs* in five mammalian genomes. The relation between the chromosome length and the number of genes *PΨgs* that are allocated in the chromosome (destination). The right panel shows the relation between the chromosome length and the number of *PΨgs* that are retrotansposed from that chromosome.

The position of parent gene and the placement location shows the origin and the destination of the *PΨgs* in the other four species (chimpanzee, dog, mouse and rat) (Figure 6-2A, Appendices VI: Figure S1). We detected dismantlement in about 90% of the *PΨgs* copies when they are re-inserted in a different chromosome relatively to the parental gene. However, when we considered exclusively the *PΨgs* re-inserted in the same chromosome, the dismantlement decrease to about 78% in 3415 *PΨgs*. In human, we have been able to use 7824 *PΨgs* out of 8283, while the remaining were discarded since we were unable to determine the parental genes. From those 7824 *PΨgs*, 6803 are inserted in a different chromosome while 1021 are inserted in the same chromosome relatively to the parent gene location (Figure 6-2A). The insertion in a different or the same chromosome (relatively to the parent gene) seems determinant of the absence of stop codons, since has been observed 716 (10.5%) in a different chromosome and 145 (14.2%) in the same chromosome, as maintaining the ORF (Figure 6-2). This suggests that in human genome is less evident this disproportion between *PΨgs* inserted in the same or different chromosome (free of stop codons). Even though, the results are statistical significant (Z-score, p-value<<0.01) favoring that the proportion of *PΨgs* with stop codons that are re-allocated in different chromosome is lesser than those inserted in the same chromosome relatively to the parent gene.

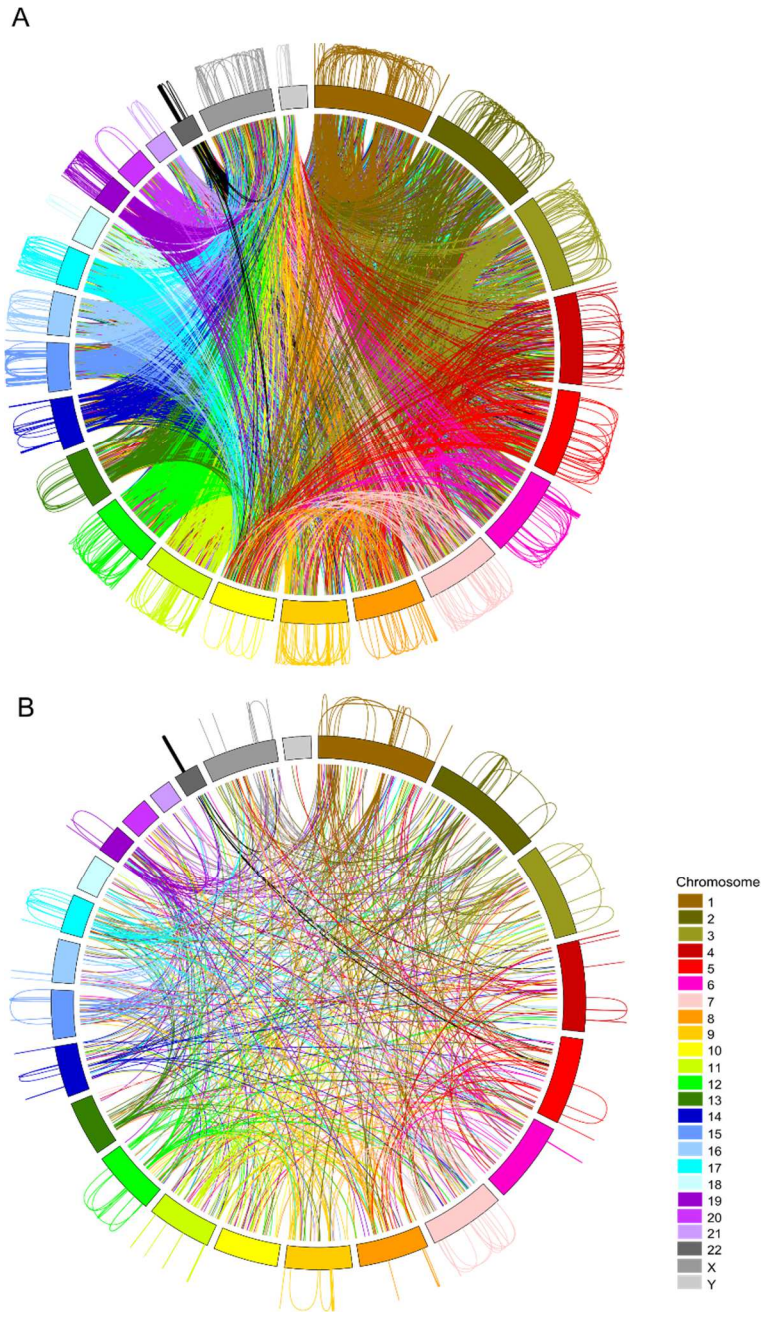

Figure 6-2. Processed *pseudogenes* in human genome.Human genome and the chromosome position of *PΨgs*, the linked position of the placement is colored accordingly to the parent gene chromosome, inner linker are those retro-transposed to a different chromosome, and external linker those retro-transposed to the same chromosome. A) Represent the annotated 7824 *PΨgs* in human; B) *PΨgs* without in-frame stop codons.

Insertion in the same Chromosome

Concerning those *PΨgs* placed within the same chromosome of the parent gene is observed a non-random location. As the placement location of those *PΨgs* was closer to the parent gene than a completely random placement will impose (Figure 6-3A). The Mann-Whitney test shows that the observed placement positions are significantly closer to the parent gene than a random placement (p-value < 0.01). This is also observed in simulations

using the original positions to constrain the initial placement (p-value < 0.01) (Figure 6-3B). Together, these two different simulation scenarios (original positions and completely random) revealed that the *PΨgs* are closer to the parent gene than a completely random distribution would predict (Figure 6-3A).

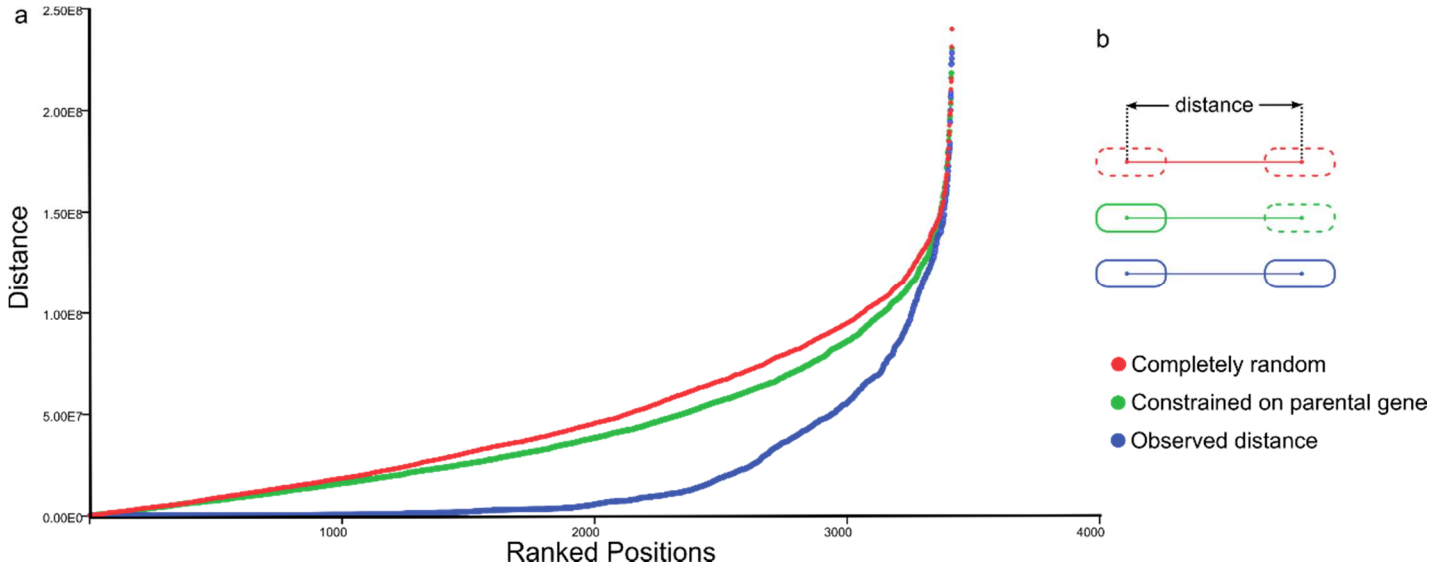

Figure 6-3. Distance from processed pseudogenes and parental gene. Distance of PΨgs transposed to the same chromosome. Blue represents the distance observed, and two random scenarios, red – completely random and green – constrained to the parent gene position. b) Schematic representation of the different scenarios, colored accordingly to the previous description. Dashed lines represents randomized positions while continuous represent lines observed position.

The presence of segmental duplications (SDs), may alter the results from the parent-*PΨg* measured distance analyses, particularly on those SD pairs (duplicated segments) were are present the pseudogene and the parent gene. Based on previous annotations of segmental duplications in human genome, Khurana *et al* (Khurana, Lam et al. 2010), the pseudogenes within this type of regions has been removed. This procedure reduced the number of *PΨgs* from 1021 to 879, therefore 142 *PΨgs* are within segmental duplications and in the same chromosome of the parent gene. Although this reduction did not affect significantly affect the analysis, Mann-Whitney-Wilcoxon test U=439658.5, p<0.2234 (Figure 6-4).

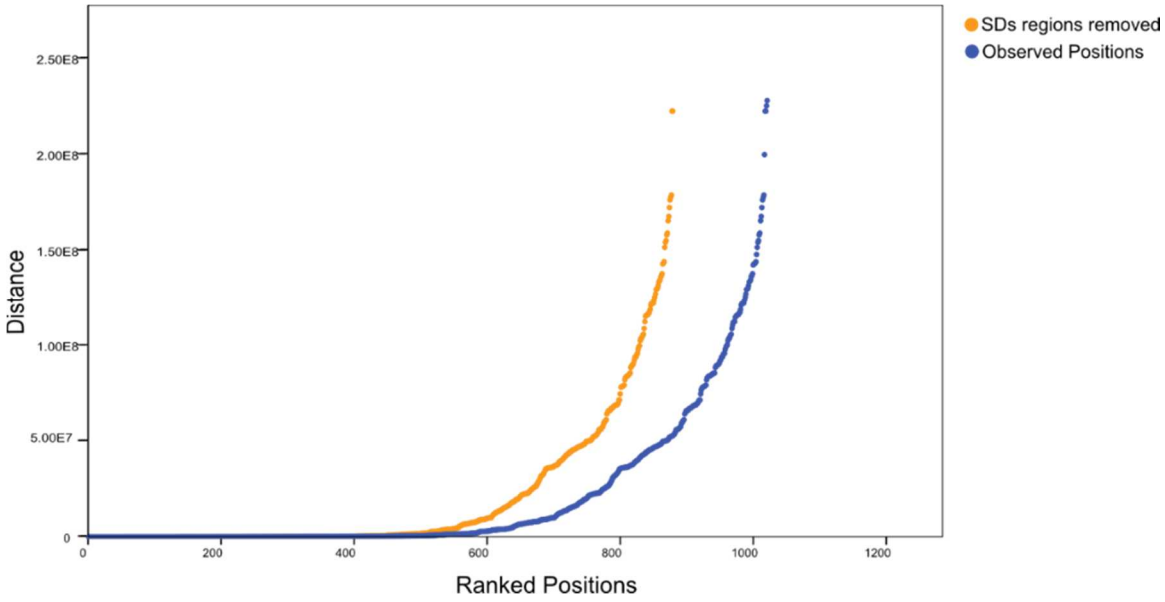

Figure 6-4. Observed distance from processed pseudogenes and parent gene after the removal segmental duplication regions. The observed distance between the parent gene and the processed pseudogene allocated in the same chromosome, including all regions, blue, and after the removal the regions associated with segmental duplications, orange.

Disablement of the *PΨgs* and their placement distance relatively to the parent gene follows also shown an observable trend distance-based, as the proximity shown most of the genes without premature stop codons are closer to the parent gene (Figure 6-5). Analyzing the distance of *PΨgs*, with or without in-frame stop codons, has shown that 50% of the *PΨgs* are distanced within 5Mbp from the parental genes (Figure 6-5). Yet, when comparing pseudogenes placed within 25Mbp relatively to the parental gene, the difference in the proportion of pseudogenes containing stop codons and those free of stop codons are dissimilar. Since has been observed ~70% containing in-frame stop codons and ~84% considering only those *PΨgs* without stop codons.

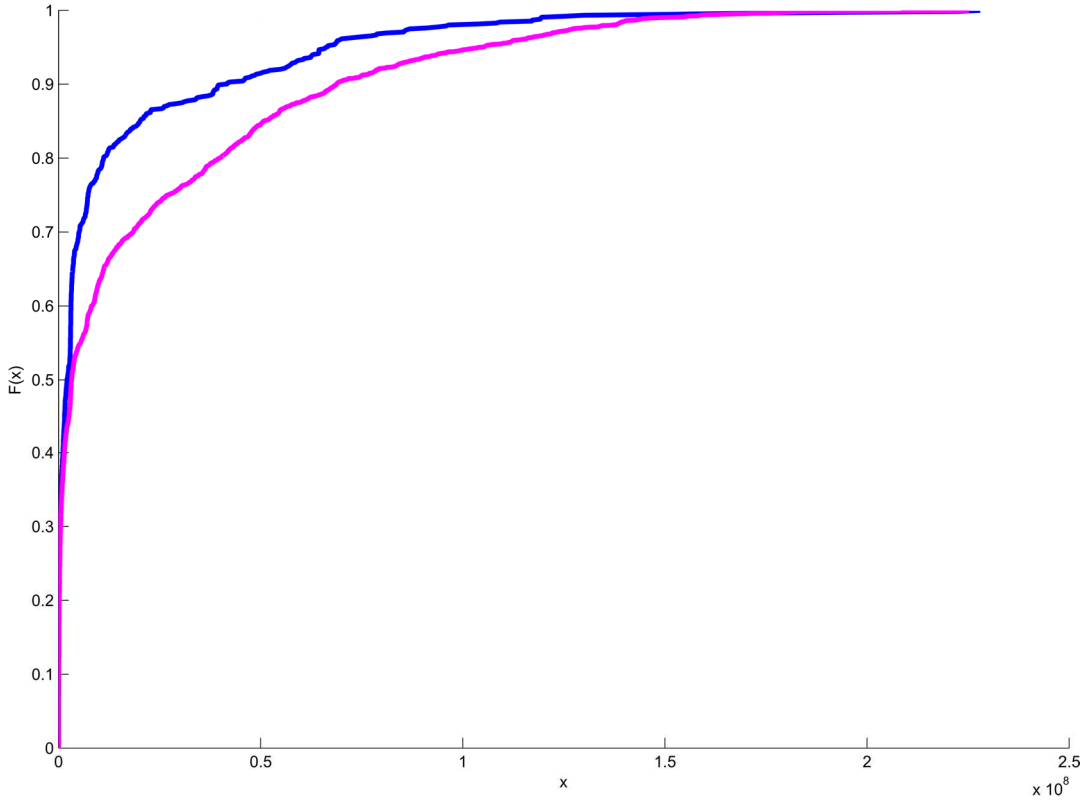

Figure 6-5. Comparative empirical cumulative distribution of processed pseudogenes inserted in the same chromosome. Purple line represents *PΨgs* containing stop codons and the blue line represents the *PΨgs* free of stop codons.

Analyzing the estimated d_N/d_S between the processed pseudogene and the parent gene shows that the closest genes have a lower degree of divergence, opposed to those farther placed, i.e. more distant of the parent gene (Figure 6-6). After deleting the sequence downstream the first stop codon, the increased divergence in course to ranked position is partially masked (Figure 6-6).

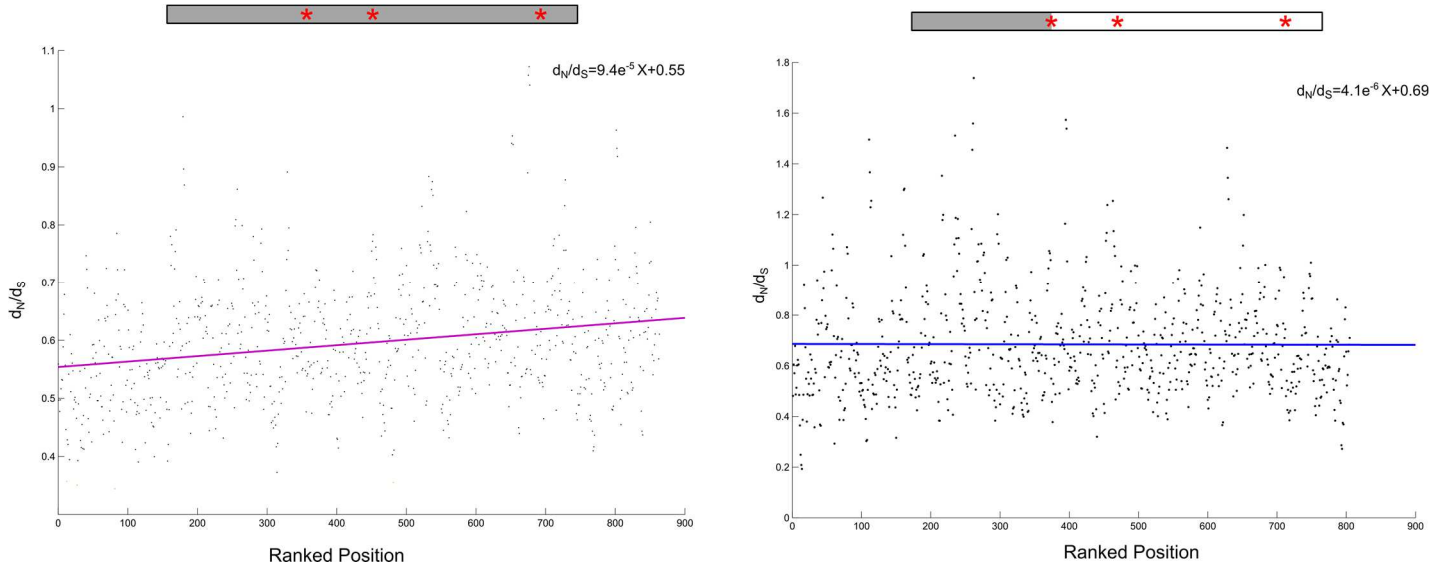

Figure 6-6. Processed pseudogenes and the calculated d_N/d_S for each parent gene accordingly to the ranked position. Purple line represents when considering the entire pseudogene and the blue line represents region upstream before the appearance of the first stop codon.

Processed Pseudogenes Expression

Along with the absence of stop codons, the *PΨgs* expression is typically associated with pseudogenes as possible functional elements. On the other hand, symptoms of non-functionality include frame disablement (premature stop codons), pseudogenes nucleotide sequence decay or incompleteness. Nonetheless several disabled *PΨgs* disabled are transcribed (Harrison, Zheng et al. 2005), concordantly the data from *RetrogeneDB* shows that from 615 expressed *PΨgs* (Figure 6-7A), from those only 188 *PΨgs* maintain the ORF, nearly 30.5% (Figure 6-7B). Interestingly the *RCPedia* shown an association between the genomic context and the fate of the *PΨg*, since from 0.26 genes that are expressed are inserted in intragenetic regions in the same strand. While those placed in intergenic positions and intragenic positions but on opposite strand, only 0.14 and 0.16 inserted in those regions are expressed (Figure 6-7C).

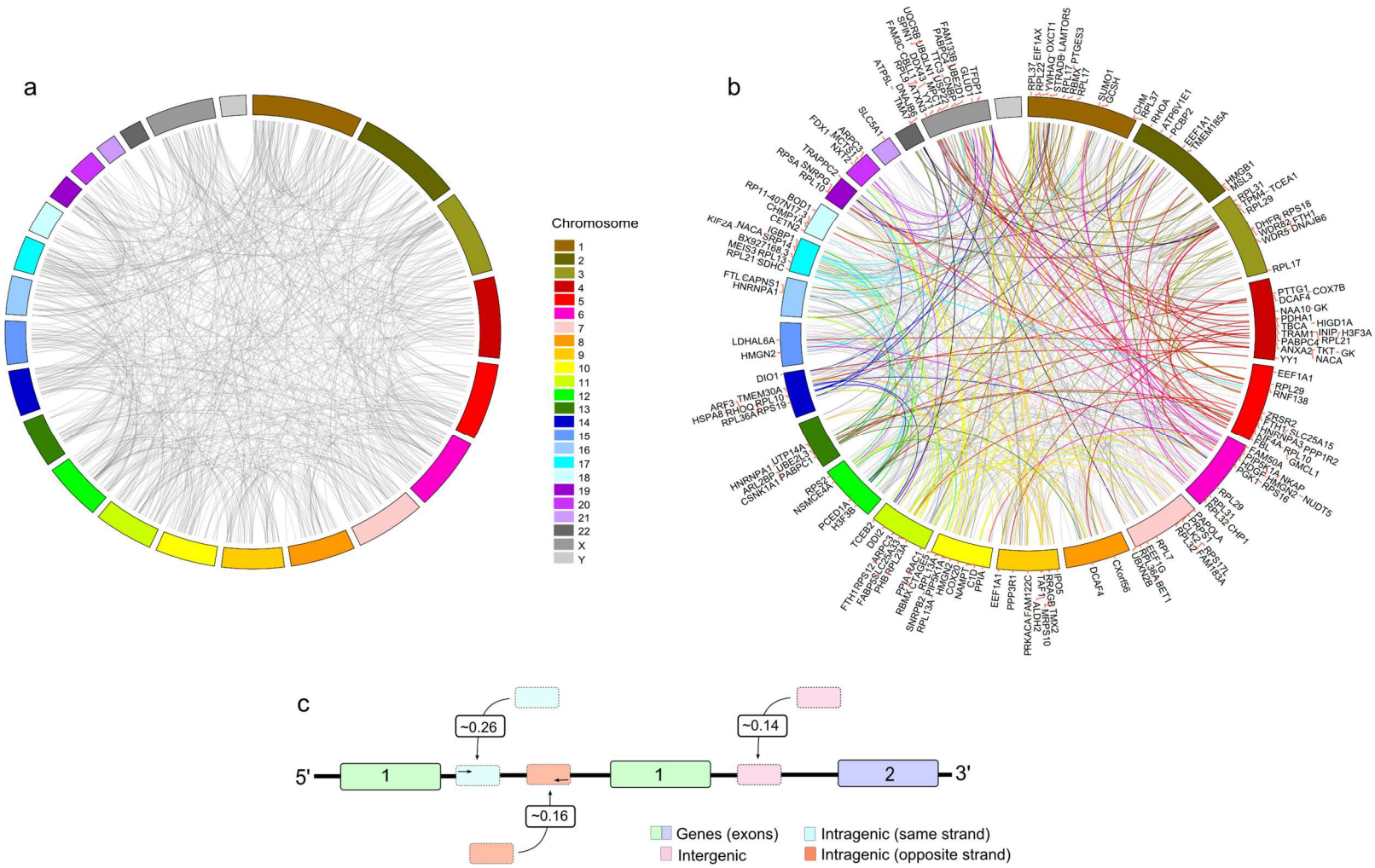

Figure 6-7. Processed pseudogenes expression and genomic context. Data retrieved from *RetrogeneDB*: A) Expressed pseudogenes; B) Expressed pseudogenes that conserve the ORF, text labels refers the parent gene name, the links are colored accordingly to the chromosome where the pseudogene is placed. Data retrieved from RCPedia C) Schematic representation of the expressed *PΨgs* and the genomic context.

Positive selection on Processed Pseudogenes

The data retrieved from *Ensembl* and pseudogene.org encompassing 18 species, performed a total of 33,256 and 1,871 *PΨgs*, respectively. After cleaning redundant sequences and those with stop codons inserted were produced 1,693 multiple sequence alignments, further used in selection analysis. The sequences from *pseudogenes.org* and *Ensembl* has been classified, and accordingly to their parental origin, were joined in a single dataset for each gene, while redundant sequences were removed. Those matching the criteria imposed here for the adaptive selection analysis (see Methods section for clarification) were aligned and used to reconstruct a gene-based phylogenetic tree. From the 104 *PΨgs* matching this criterion, nearly 40 genes showed signatures of positive selection, representing nearly 38% of total amount of the analyzed genes (Appendices VI: Table S1). The cleaning procedure through MACSE and the filtering using GBLOCKS were unsufficient to radically affected the number of positively selected genes, (Appendices VI: Table S2). Additionally, removal of sequences showing a dissimilar amino acid composition relatively to the remaining sequences, also did not alter the results significantly (Appendices VI: Table S3). The use of both nested models, M7 vs M8 followed by M8 vs M8a, constitute a reasonable statistical framework that ensures the corrected detection of positive selection in the sequences. Since the simultaneous acceptance of the alternate model in both comparisons suggests that the genes are changing more than the neutrality will impose. This is particularly relevant for analyses in pseudogenes since the combined acceptance of the pairwise tests, M7 vs M8 and M8 vs M8a, ensures that the site-class under positive selection ($\omega > 1$) is significantly above one (i.e. evolving under neutrality ≈ 1). It is worthy to explain that in several multiple sequence alignments are included "*in-paralogs*" but was ensured that at least two species were present in the selection analyses. Additionally to the presence of *in-paralogs* the retrieved sequences corresponds most likely to temporally varied insertions in the genome. Therefore later was tested for two randomly chosen genes, *BTF3* and *PPIA*, the possibility of an incorrect detection of positive selection given the different number of sequences and the temporal distant retro-transpositions events. Although in both simulations for 93 and 97 valid replicates has been obtained six and 12 false positives tests, respectively. This shows a false positive rate of 0.1%, even though clearly bellow the reported proportion of positively selected genes 0.38.

Codon position and evolutionary rate

Additionally we tested the distance (p-distance) in the pairwise comparison (parental vs processed pseudogene) relatively to each nucleotide position in codon. In disabled and non-disabled pseudogenes has been observed a higher distance in the 3th position, almost the double, relatively to the 1th and 2th position (Appendices VI: Table S4). This differences suggests a higher evolutionary rate in the third position relatively to the first two positions.

Genes evolving under neutrality

Since it was detected positive selection on *PΨgs* it is relevant to access the time course for a gene while evolving under complete neutrality to be undetectable through blast searchers (BLASTn or BLASTx). Using the cut-off value $1x10^{-10}$ and assuming an evolutionary rate of $2.415x10^{-9}$ mutations/site/year (Kumar and Subramanian 2002) we estimated for eight genes (BTF3, CYC1, GADPH, H3F3B, PPIA, RPL7A, SCOP, RPL39) the estimated time needed for the blast searches to be below the empirical cut-off (Appendices VI: Figure S2). Our data supports that will take nearly 229.46 ± 100.82 Myr for a gene to be undetectable trough blast searches but the insertion of the first stop codon occurs around 26.23 ± 22.95 Myr (Appendices VI: Table S5).

6.5 Discussion

Previous studies reported random placement of human *PΨgs*, described as a "bombardment" along evolution (Zhang, Harrison et al. 2003), turning retro-transposition an efficient process introduce regulatory elements into the genome "in search" of new target genes (Brosius 1999). Consistently, this "stated war" was also observed for the other mammals (here we reported this effect on four additional mammalian species). This seems a general trend for mammals, since larger chromosomes consistently tend to accumulate a higher number of *PΨgs*. Similarly, as previously observed in the human genome, pseudogenes allocation seen a random process, mainly affected by the chromosome length (Zhang, Harrison et al. 2003).

However, not all genes are prone to form *PΨgs* and several factors affecting which genes can constitute pseudogenes have been proposed, such as expression level and gene length (Li, Yang et al. 2013), or mRNA stability (Pavlicek, Gentles et al. 2006). Additionally gene length has been associated as a major determinant for the pseudogenization process, since longer genes tend to "produce" more non-processed pseudogenes, whereas *PΨgs* are typically formed by short protein-coding genes (Goncalves, Duret et al. 2000; Zhang,

Harrison et al. 2002; Khachane and Harrison 2009). At first glance, this intrinsic properties seen to be irrelevant to adaptation process, or unrelated with the process involved in the formation of "new" process pseudogenes. Although this is suggestive that the raw material to form "new" *PΨgs* appears to be highly associated with the nature of genes or mRNA molecules. Since these properties has been shown to have an adaptive value, such as mRNA stability (Yamanaka and Inouye 2001; Dressaire, Picard et al. 2013), gene expression (Lopez-Maury, Marguerat et al. 2008; Fraser 2013) and gene length (Lipman, Souvorov et al. 2002). Hence, despite the random insertion of the *PΨgs*, the current data is suggestive that their formation might be under selective regime. It is therefore plausible to expect constrains in gene evolution to maintain properties that potentiate retro-transposition events of those genes (parental genes). Likewise, this partially explain the nearly absent correlation between chromosome length and the number of *PΨgs* in each chromosome (considering the origin in the chromosome location of the parental gene). Although the generalization that only functional genes lead to the arousal of "new" *PΨgs* constitute an over-simplification of this process. An alternative path can be attained through duplications of previously retro-transposed genes, generating duplicated-processed pseudogenes (Zhang, Harrison et al. 2003). This distinct *PΨgs* origin could potentially influence the estimations in the frequency of mRNA transcripts retro-transposition events, as on his basis template for the retro-transposition differ from the typical *PΨgs* (Zhang, Cayting et al. 2008). Yet the results when excluding and including those *PΨgs* reveal a similar trend, suggestive that the deposition of *PΨgs* is mainly influenced by the chromosome length, while the origin depends on factor such as mRNA transcript properties.

Despite the *PΨgs* random re-allocation sites, their placement location seems to be associated with the occurrence frequency of premature of stop codons, indication of putative loss of function (Harrison, Zheng et al. 2005). Our results supports that those *PΨgs* incorporated in nearest locations relatively to the parent gene accumulate lesser differences relatively to the parent gene, and their allocation in the same chromosome lead to a higher retention (i.e. no perceived disablement of the processed pseudogene). Those placed in the same chromosome showed a decreased probability to accumulate stop codons, and furthermore conversely those inserted in the same chromosome of the parental gene tend to be distributed closer to the parent gene that a completely random distribution will impose. This is highly unexpected, as the association between chromosome length and number of *PΨgs* is suggestive of a random placement. While has been suggests that some *PΨgs* close to their parents on chromosomes and adjacent genes tend to be co-regulated (Cabili, Trapnell et al. 2011): Later has been shown an absence of co-regulation due to their close chromosomal proximity of the parent gene and pseudogene but rather an correlation between *PΨgs* and parent gene irrespective to their distance (Guo, Lin et al. 2014).

Based on the analysis of *PΨgs* in human and mouse genomes, it has been estimated that the formation of "new" *PΨgs* occurs in a rate of about 1-2% per gene per million years (Sakai, Koyanagi et al. 2007). While gene duplications occurs at a predicted rate of 0.9% per gene per million years based on estimation in the human genome (Sakai, Koyanagi et al. 2007). This suggests an important role in the arousal of "new genes" through *PΨgs* even more relevant than gene duplications. Their maintenance also appears to be of prime relevance since nearly 40% of the processed pseudogenes are shared by human and mouse, i.e. formed before these two species diverged and maintained (Zhang, Carriero et al. 2004), and many of them maintain intact the coding regions. Furthermore the loss of ORF (that implies a loss of the ability to encode a functional protein and that), and if expressed, the ancestral tissue expression patterns and gene expression levels show tendencies to be retained (Marques, Tan et al. 2012). The maintenance of the expression patterns might be important for regulatory processes (Tam, Aravin et al. 2008; Poliseno, Salmena et al. 2010) and probably enabling the possibility of pseudogenes functional resurrection, described as frequent in human and mouse pseudogenes (Sakai, Koyanagi et al. 2007).

Furthermore we detected an adaptive value for their presence in the mammalian genomes, as the presence of selection signatures in nearly 40% in a total of 104 analyzed *PΨgs*. Functional relevancy in the presence of pseudogenes was previously reported (Allmendinger and Knowles 2013), and integrative analyses highlighted their functional role and potential adaptive value (Korrodi-Gregorio, Abrantes et al. 2013). Previous reports shows that there is positive selection acting on pseudogenes as an high d_N/d_S is indicative of an adaptive value (Korrodi-Gregorio, Abrantes et al. 2013) but may also be indicative of functional divergence relatively to parent gene and used frequently to classify them as pseudogenes. This could be partially explained as the similarity to the functional genes is critical to detected pseudogenes in the genomes, therefore their retrieval is strongly biased. Moreover, if those sequences evolve under an accelerated rate, this implies that their resemblance with functional copies is lower, and therefore difficult to detect and to characterize. Even so, under a neutral evolutionary pace, it will take ~244Myr for a gene to become "undetectable" while the hypothetical frame-shift might occurs much earlier (~26Myr). This implies that pseudogenes, either processed or non-processed that are detected through blast searches are predictability above ~244 Myr were under functional constrains during their evolution. An implication is that, for instance, genes detected as pseudogenes from the teleost specific whole genome duplication were, if not currently, functionally constrained during time. Along with *PΨgs* and the arguments about their predicted or proofed functionality, a recent work showed that some nonprocessed pseudogenes can be expressed, raising the interest about their role as functional units (Branca, Orre et al. 2014).

The lack of an appropriate regulatory environment often lead to the degeneration of the *PΨgs* (D'Errico, Gadaleta et al. 2004), and we found that the integration of the processed pseudogene near the parent gene lead to an increased probability to be retained. Concordantly and interdependence of a parent gene and its pseudogene was reported for the gene *ABCC6* and the *ABCC6P1*, since their co-expression result in a decrease of the expression *ABCC6* (Piehler, Hellum et al. 2008). Moreover, there increasing evidences that gene order in eukaryotes does not follow a random location, as those closely placed tend to be co-expressed and co-functional (Michalak 2008). Interestingly, since shorter genes tend either to evolve faster and are typically younger genes (proteins) when compared to longer genes. This is partially contradictory as the older genes tend to be highly expressed and therefore tend to be subjected to retro-transposition events. Although is reasonable to suspect that the retro-transposition constitutes a potential mechanism that regulates "newly" evolved genes.

Here we reported three relevant aspects of *PΨgs*: 1) pseudogenes retro-transposed to the same chromosome have higher chances to "survive"; 2) pseudogenes retro-transposed to the same chromosome of the parental genes are placed closer to the parent gene than a completely random process will impose or have a lower decay rate; and 3) detected selection signatures on nearly 40% of 104 tested genes using d_N/d_S approach implemented in CODEML.

6.6 Conclusions

The search for functional elements in genomes is of prime relevance for evolutionary biology, and during the last years, several arguments about *PΨgs* functionally have been shown. The processes that lead to their arousal and pervasive presence in the mammalian genomes is therefore of prime relevance to understand their function. While *PΨgs* seem randomly placed and dispersed in the genomes, their location apparently determine their chances to acquire a functional relevancy. Those retro-transposed closer to the parent gene (same chromosome) twice their chances to survive (i.e. no perceived introduction of stop codons). Also those placed in intragenic regions, in the same strand of the 'receiver' gene have higher chances to be expressed. Thus, this highlight a mechanism of natural selection acting on those *PΨgs*, revealing a functional relevancy in their maintenance and placing them as potential functional units.

6.7 Acknowledgements

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to JPM (SFRH/BD/65245/2009). This work was further supported by a grant from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism. AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013 and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490).

Chapter 7 Discussion

In this thesis was aimed to study vertebrates under differential selective regimes. Thus, several comparative studies of vertebrate's genes were conducted in mammals, birds and fishes, and their potential adaptive value determined in order to understand at the molecular level how natural selection operates. Considering that neutrality can be excluded, then insight is obtained on the action of natural selection on genes. The thesis work included the detection of natural selection in bone and tooth associated genes (**Chapter 2**, **3** and **4**). Later it was studied the role of positive selection in duplicated genes maintenance and functional diversification (**Chapter 5**). Finally, regions of presumed "junk" DNA were studied, mirroring the potential adaptive value of processed pseudogenes (**Chapter 6**).

Despite the high conservation in terms of Vertebrates body plans they have diversified lifestyles and life habitats. Therefore if at the molecular level neutrality can be excluded then genes might present molecular evidences of the species diversified modes of locomotion, diets or life habitats.

In this purpose the **Chapter 2** was focus on MEPE and their evolution in birds and mammals. Despite often described as a multifunctional protein, in MEPE only two main functional domains were described (dentonin and ASARM). However, the analyses of selection signatures revealed 20 sites positively selected at simultaneously two levels, codon and amino acids based approaches. Those highly supported positively selected sites suggest a crucial role of other residues outside the two main domains, since positive selection is often clearly associated with functional relevance (Yokoyama, Tada et al. 2008). Moreover, directional selection was detected toward "intrinsic disorder" residues accumulation. This seem crucial for biomineralization, since was later observed that the growth of mineral phase is attained trough proteins structurally highly disordered (Kalmar, Homola et al. 2012). Concordantly in **Chapter 4** the work in tooth associated genes, also reported a prevalence of positively selected sites in "intrinsic disorder" regions.

The comparative analyzes of MEPE sequences revealed that the functional domain RGD is not fully conserved in Chiroptera, neither in the *Tursiops truncatus* or *Procavia capensis.* The few rodents surveyed also reported alteration in RGD motif, however as explained in **Chapter 2**, they present relatively to other mammals a significant altered evolutionary rate. The alterations observed in RGD of Chiroptera and *T. truncatus* seem relevant. Since as further examined in **Chapter 3**, the bats ability to flight constrained their bone associated genes evolution, as they present different selective pattern relatively to other mammals. If the altered RGD in bats is a consequence of flight ability remains an open question. Although it is interesting to observe that in *T. truncatus* RGD motif is also altered,

since there is an association between mechanical loading and the formation of “new” bone (Tanaka, Sun et al. 2004) and MEPE is highly regulated mechanical loading (Reijnders, van Essen et al. 2013). This is particular relevant since bone homeostasis in terrestrial animals endeavor challenges of a higher gravity and calcium-poor environment when compared with the aquatic environment that subject bones to reduced mechanical loading (Bouillon and Suda 2014). Additionally this can partially explain the higher number of SIBLING members observed in birds, mammals and reptiles vertebrates when compared to fishes and amphibians.

On **Chapter 3** the main focus where on the impact of flight ability modulating bone associated genes. The results strongly suggest that bone remodeling genes are under selective and diversifying selection in birds. Remarkably bats show, relatively to the other mammals, an unusual evolutionary rate. This remarks that the transition from terrestrial to aerial locomotion impose a similar great challenges to the bone structure and the different selective pressure models the evolution of bone to attain flight efficiency, a light but stiff bone structure. It is noteworthy to observe that positive selection “target” specific gene ontology processes such as bone associated remodeling genes. Similarly preferred positive selection “targets” were observed on the transitions and adaptation of Cetaceans to aquatic life (Nery, Gonzalez et al. 2013).

Similar as skeleton, tooth reflect the high diversified lifestyles observed in vertebrates. Within this taxonomic group, it is among mammals where the highest variety of number, shape and structure is observed. They reflect the variety of different strategies to chop, grind or chew a wide range of diverse food. The results from the **Chapter 4** suggest that the diverse phenotypes observed in mammals are the results of positive selection pervasive in “newly” emerged genes, since mammalian specific genes showed higher evolutionary rate. Typically new genes emerge performing as redundant or highly specialized in terms of function (Domazet-Loso and Tautz 2003), and therefore can easily become lost if the selective regime is altered. Although if they are later co-opted to another function, will result in a prevalence of positive selection, which would be detectable trough an elevated d_N/d_S ratio (Vishnoi, Kryazhimskiy et al. 2010). Remarkably the gene age seems also intimately associated with the timeframe increased gene expression, as new genes tend to be more expressed in later developmental stages. On **Chapter 4** was also reported a correlation between the intronic and exonic evolutionary rate. The higher departure from neutrality observed in introns of positively selected genes remarks that some of the adaptations are not in the exonic regions but rather rely in the intronic regions.

The gene duplication generates “new” raw material but also impose great challenges. Given the presence of two copies of a gene perform a similar function that could result in imbalance of the dose of a protein or their product. On **Chapter 5** was studied the

WAP65, showing that subtle differences outside the main functional domains (heme-pocket) explain their retention in some teleosts. The results show also a counter-example to the more common acceleration of a paralogue relatively to the singletons, since it was observed that HPX have a higher evolutionary rate to both WAP65-1 and WAP65-2. The subfunctionalization is a transitory state that leads to neofunctionalization, forward the emergence of new functions in one of the paralogs. Despite the uncertainty about the key advantage for teleosts that possess two copies. The teleosts specific whole-genome duplication and the selective alteration in the evolutionary rate of one copy explain their retention and was a major contributor for teleosts success (Philip, Machado et al. 2012).

As discussed in **Chapter 5**, gene duplication is a key event in generating evolutionary novelty. Recently, are raising evidences on the functional importance and relevance of processed pseudogenes. Moreover, they are probably also important in the emergence of new genes (Kaessmann 2010). The **Chapter 6** shows the presence of signatures of positive selection acting on processed pseudogenes. Moreover, is suggestive that despite their distribution in the mammalian genome seems random, those allocated in the same chromosome have higher chances to be retained as functional. Thus, should processed pseudogenes still be considered “junk” DNA?

Summarizing, the results show that the diversification of life habits, diets and mode of locomotion lead signatures of positive Darwinian selection on gene protein encoding regions in Vertebrates. But the classical view that adaptive changes are observable solely in those regions probably needs a reformulation.

References

Abascal, F., R. Zardoya, et al. (2010). "TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations." Nucleic Acids Res **38**(Web Server issue): W7-13.

AccelrysSoftwareInc. (2012). Discovery Studio Modeling Environment. R. 3.5. http://accelrys.com/, Accelrys Software Inc.

Addison, W. N., D. L. Masica, et al. (2010). "Phosphorylation-dependent inhibition of mineralization by osteopontin ASARM peptides is regulated by PHEX cleavage." J Bone Miner Res **25**(4): 695-705.

Addison, W. N., Y. Nakano, et al. (2008). "MEPE-ASARM peptides control extracellular matrix mineralization by binding to hydroxyapatite: an inhibition regulated by PHEX cleavage of ASARM." J Bone Miner Res **23**(10): 1638-1649.

Alexander, R. M. (1998). "All-time giants: the largest animals and their problems." Palaeontology **41**: 1231-1245.

Alexander, R. P., G. Fang, et al. (2010). "Annotating non-coding regions of the genome." Nat Rev Genet **11**(8): 559-571.

Alexopoulou, O., J. Jamart, et al. (2006). "Bone density and markers of bone remodeling in type 1 male diabetic patients." Diabetes Metab **32**(5 Pt 1): 453-458.

Aliza, D., I. S. Ismail, et al. (2008). "Identification of Wap65, a human homologue of hemopexin as a copper-inducible gene in swordtail fish, Xiphophorus helleri." Fish Physiol Biochem **34**(2): 129-138.

Allmendinger, R. and J. Knowles (2013). "On handling ephemeral resource constraints in evolutionary search." Evol Comput **21**(3): 497-531.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Andolfatto, P. (2005). "Adaptive evolution of non-coding DNA in Drosophila." Nature **437**(7062): 1149-1152.

Anisimova, M. and O. Gascuel (2006). "Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative." Syst Biol **55**(4): 539-552.

Anisimova, M., M. Gil, et al. (2011). "Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes." Syst Biol **60**(5): 685-699.

Anisimova, M., R. Nielsen, et al. (2003). "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites." Genetics **164**(3): 1229-1236.

Antunes, A. and M. J. Ramos (2007). "Gathering computational genomics and proteomics to unravel adaptive evolution." Evol Bioinform Online **3**: 207-209.

Argiro, L., M. Desbarats, et al. (2001). "Mepe, the gene encoding a tumor-secreted protein in oncogenic hypophosphatemic osteomalacia, is expressed in bone." Genomics **74**(3): 342-351.

Armfield, B. A., Z. Zheng, et al. (2013). "Development and evolution of the unique cetacean dentition." PeerJ **1**: e24.

Armstrong, S., A. Pereverzev, et al. (2009). "Activation of P2X7 receptors causes isoform-specific translocation of protein kinase C in osteoclasts." J Cell Sci **122**(Pt 1): 136-144.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Bae, Y. S., J. H. Lee, et al. (2009). "Macrophages generate reactive oxygen species in response to minimally oxidized low-density lipoprotein: toll-like receptor 4- and spleen tyrosine kinase-dependent activation of NADPH oxidase 2." Circ Res **104**(2): 210-218, 221p following 218.

Balakirev, E. S. and F. J. Ayala (2003). "Pseudogenes: are they "junk" or functional DNA?" Annu Rev Genet **37**: 123-151.

Bardet, C., S. Delgado, et al. (2010). "MEPE evolution in mammals reveals regions and residues of prime functional importance." Cell Mol Life Sci **67**(2): 305-320.

Barja, G. (1998). "Mitochondrial Free Radical Production and Aging in Mammals and Birdsa." Ann N Y Acad Sci **854**(1): 224-238.

Barrett, T., T. O. Suzek, et al. (2005). "NCBI GEO: mining millions of expression profiles--database and tools." Nucleic Acids Res **33**(Database issue): D562-566.

Barrett, T., S. E. Wilhite, et al. (2013). "NCBI GEO: archive for functional genomics data sets--update." Nucleic Acids Res **41**(Database issue): D991-995.

Bassett, J. H., A. Gogakos, et al. (2012). "Rapid-throughput skeletal phenotyping of 100 knockout mice identifies 9 new genes that determine bone strength." PLoS Genet **8**(8): e1002858.

Baud'huin, M., N. Solban, et al. (2012). "A soluble bone morphogenetic protein type IA receptor increases bone mass and bone strength." Proc Natl Acad Sci U S A **109**(30): 12207-12212.

Benson, R. B., R. J. Butler, et al. (2012). "Air-filled postcranial bones in theropod dinosaurs: physiological implications and the 'reptile'-bird transition." Biol Rev Camb Philos Soc **87**(1): 168-193.

Benson, R. B., R. J. Butler, et al. (2012). "Air-filled postcranial bones in theropod dinosaurs: physiological implications and the ‘reptile’–bird transition." Biological Reviews **87**(1): 168-193.

Binns, D., E. Dimmer, et al. (2009). "QuickGO: a web-based tool for Gene Ontology searching." Bioinformatics **25**(22): 3045-3046.

Birney, E., T. D. Andrews, et al. (2004). "An overview of Ensembl." Genome Res **14**(5): 925-928.

Bouillon, R. and T. Suda (2014). "Vitamin D: calcium and bone homeostasis during evolution." Bonekey Rep **3**: 480.

Bouzat, J. L. (2000). "The importance of control populations for the identification and management of genetic diversity." Genetica **110**(2): 109-115.
Brame, L. A., K. E. White, et al. (2004). "Renal phosphate wasting disorders: clinical features and pathogenesis." Semin Nephrol **24**(1): 39-47.
Branca, R. M., L. M. Orre, et al. (2014). "HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics." Nat Methods **11**(1): 59-62.
Braun, E. J. and K. L. Sweazea (2008). "Glucose regulation in birds." Comp Biochem Physiol B Biochem Mol Biol **151**(1): 1-9.
Brosius, J. (1999). "Genomes were forged by massive bombardments with retroelements and retrosequences." Genetica **107**(1-3): 209-238.
Brudno, M., C. B. Do, et al. (2003). "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA." Genome Res **13**(4): 721-731.
Brunet-Rossinni, A. K. (2004). "Reduced free-radical production and extreme longevity in the little brown bat (< i> Myotis lucifugus</i>) versus two non-flying mammals." Mechanisms of ageing and development **125**(1): 11-20.
Bustamante, C. D., R. Nielsen, et al. (2002). "A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents." Mol Biol Evol **19**(1): 110-117.
Butler, W. T. (1998). "Dentin matrix proteins." Eur J Oral Sci **106 Suppl 1**: 204-210.
Cabili, M. N., C. Trapnell, et al. (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." Genes Dev **25**(18): 1915-1927.
Canals, M., C. Atala, et al. (2005). "Relative size of hearts and lungs of small bats." Acta Chiropterologica **7**(1): 65-72.
Capra, J. A., A. G. Williams, et al. (2012). "ProteinHistorian: tools for the comparative analysis of eukaryote protein origin." PLoS Comput Biol **8**(6): e1002567.
Casasoli, M., L. Federici, et al. (2009). "Integration of evolutionary and desolvation energy analysis identifies functional sites in a plant immunity protein." Proc Natl Acad Sci U S A **106**(18): 7666-7671.
Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Mol Biol Evol **17**(4): 540-552.
Caviedes-Vidal, E., T. J. McWhorter, et al. (2007). "The digestive adaptation of flying vertebrates: high intestinal paracellular absorption compensates for smaller guts." Proc Natl Acad Sci U S A **104**(48): 19132-19137.
Chatr-Aryamontri, A., B. J. Breitkreutz, et al. (2013). "The BioGRID interaction database: 2013 update." Nucleic Acids Res **41**(Database issue): D816-823.
Chen, D., M. Zhao, et al. (2004). "Bone morphogenetic proteins." Growth Factors **22**(4): 233-241.
Chen, S., L. Chen, et al. (2008). "Expression and processing of small integrin-binding ligand N-linked glycoproteins in mouse odontoblastic cells." Arch Oral Biol **53**(9): 879-889.
Cherayil, B. J. (2011). "The role of iron in the immune response to bacterial infection." Immunol Res **50**(1): 1-9.
Chevin, L. M. and A. P. Beckerman (2012). "From adaptation to molecular evolution." Heredity (Edinb) **108**(4): 457-459.
Cho, Y. S., B. S. Kim, et al. (2012). "Modulation of warm-temperature-acclimation-associated 65-kDa protein genes (Wap65-1 and Wap65-2) in mud loach (Misgurnus mizolepis, Cypriniformes) liver in response to different stimulatory treatments." Fish Shellfish Immunol **32**(5): 662-669.
Choi, C. Y., K. W. An, et al. (2008). "Expression of warm temperature acclimation-related protein 65-kDa (Wap65) mRNA, and physiological changes with increasing water temperature in black porgy, Acanthopagrus schlegeli." J Exp Zool A Ecol Genet Physiol **309**(4): 206-214.
Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-218.
Clark, A. G., S. Glanowski, et al. (2003). "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios." Science **302**(5652): 1960-1963.
Clark, M. S. and G. Burns (2008). "Characterisation of the warm acclimated protein gene (wap65) in the Antarctic plunderfish (Harpagifer antarcticus)." DNA Seq **19**(1): 50-55.
Cleve, F. M. (1965). Anaximander of Miletus. The Giants of Pre-Sophistic Greek Philosophy, Springer**:** 144-165.
Cohen, B. A., R. D. Mitra, et al. (2000). "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression." Nat Genet **26**(2): 183-186.
Collins, M. O., L. Yu, et al. (2008). "Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder." Mol Cell Proteomics **7**(7): 1331-1348.
Colosimo, P. F., K. E. Hosemann, et al. (2005). "Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles." Science **307**(5717): 1928-1933.
Conrad, B. and S. E. Antonarakis (2007). "Gene duplication: a drive for phenotypic diversity and cause of human disease." Annu Rev Genomics Hum Genet **8**: 17-35.
Cook, L. (2000). "Changing views on melanic moths." Biological Journal of the Linnean Society **69**(3): 431-441.
Cooke, J., M. A. Nowak, et al. (1997). "Evolutionary origins and maintenance of redundant gene expression during metazoan development." Trends Genet **13**(9): 360-364.
Cornish-Bowden, A. (1985). "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984." Nucleic Acids Res **13**(9): 3021-3030.
Corp., I. (Released 2011). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Cubo, J. and A. Casinos (1994). "Scaling of skeletal element mass in birds." Belgian Journal of Zoology **124**.
Cubo, J. and A. Casinos (2000). "Incidence and mechanical significance of pneumatization in the long bones of birds." Zoological Journal of the Linnean Society **130**(4): 499-510.
Currey, J. D. (2003). "The many adaptations of bone." J Biomech **36**(10): 1487-1495.
D'Errico, I., G. Gadaleta, et al. (2004). "Pseudogenes in metazoa: origin and features." Brief Funct Genomic Proteomic **3**(2): 157-167.
Darwin, C. (1859). "On the origins of species by means of natural selection." London: Murray.
Dasmeh, P., A. W. Serohijos, et al. (2013). "Positively selected sites in cetacean myoglobins contribute to protein stability." PLoS Comput Biol **9**(3): e1002929.
Dateki, M., T. Horii, et al. (2005). "Neurochondrin negatively regulates CaMKII phosphorylation, and nervous system-specific gene disruption results in epileptic seizure." J Biol Chem **280**(21): 20503-20508.
Davit-Beal, T., A. S. Tucker, et al. (2009). "Loss of teeth and enamel in tetrapods: fossil record, genetic data and morphological adaptations." J Anat **214**(4): 477-501.
De Beur, S. M., R. B. Finnegan, et al. (2002). "Tumors associated with oncogenic osteomalacia express genes important in bone and mineral metabolism." J Bone Miner Res **17**(6): 1102-1110.
de Monti, M., S. Miot, et al. (1998). "Caractérisation d'une hémopexine sérique de truite par utilisation d'une protéine recombinante." Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie **321**(4): 299-304.
Dececchi, T. A. and H. C. Larsson (2011). "Assessing arboreal adaptations of bird antecedents: testing the ecological setting of the origin of the avian flight stroke." PLoS One **6**(8): e22292.
Dececchi, T. A. and H. C. Larsson (2013). "Body and limb size dissociation at the origin of birds: uncoupling allometric constraints across a macroevolutionary transition." Evolution **67**(9): 2741-2752.
Dee, K. C., T. T. Andersen, et al. (1998). "Design and function of novel osteoblast-adhesive peptides for chemical modification of biomaterials." Journal of Biomedical Materials Research **40**(3): 371-377.
Dehouck, Y., J. M. Kwasigroch, et al. (2011). "PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality." BMC Bioinformatics **12**: 151.
Delanghe, J. R. and M. R. Langlois (2001). "Hemopexin: a review of biological aspects and the role in laboratory medicine." Clin Chim Acta **312**(1-2): 13-23.
Dobbie, H., D. G. Shirley, et al. (2003). "Infusion of the Bone-Derived protein MEPE causes phosphaturia in rats." Journal of the American Society of Nephrology **14**: 467a-468a.
Dobzhansky, T. (1973). "Nothing in biology makes sense except in the light of evolution." The american biology teacher **35**(3): 125-129.
Domazet-Loso, T. and D. Tautz (2003). "An evolutionary analysis of orphan genes in Drosophila." Genome Res **13**(10): 2213-2219.
Dooley, H., E. B. Buckingham, et al. (2010). "Emergence of the acute-phase protein hemopexin in jawed vertebrates." Mol Immunol **48**(1-3): 147-152.
Doron-Faigenboim, A., A. Stern, et al. (2005). "Selecton: a server for detecting evolutionary forces at a single amino-acid site." Bioinformatics **21**(9): 2101-2103.
Dressaire, C., F. Picard, et al. (2013). "Role of mRNA stability during bacterial adaptation." PLoS One **8**(3): e59059.
Dumont, E. R. (2010). "Bone density and the lightweight skeletons of birds." Proc Biol Sci **277**(1691): 2193-2198.
Dutheil, J. (2008). "Detecting site-specific biochemical constraints through substitution mapping." Journal of Molecular Evolution **67**(3): 257-265.
Dutheil, J. (2008). "Detecting site-specific biochemical constraints through substitution mapping." J Mol Evol **67**(3): 257-265.
Dutheil, J. Y., N. Galtier, et al. (2012). "Efficient selection of branch-specific models of sequence evolution." Mol Biol Evol **29**(7): 1861-1874.
Edgar, R. C. (2004). "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." BMC Bioinformatics **5**: 113.
Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.
Fastovsky, D. E. and D. B. Weishampel (2005). The Evolution and Extinction of the Dinosaurs, Cambridge University Press.
Fedducia, A. (1996). "The origin and evolution of birds." New Haven: Yale University.
Field, D. J., C. Lynner, et al. (2013). "Skeletal Correlates for Body Mass Estimation in Modern and Fossil Flying Birds." PLoS One **8**(11): e82000.
Fisher, L. W. (2011). "DMP1 and DSPP: Evidence for Duplication and Convergent Evolution of Two SIBLING Proteins." Cells Tissues Organs **194**(2-4): 113-118.
Fisher, L. W. and N. S. Fedarko (2003). "Six genes expressed in bones and teeth encode the current members of the SIBLING family of proteins." Connect Tissue Res **44 Suppl 1**: 33-40.
Fisher, L. W., D. A. Torchia, et al. (2001). "Flexible structures of SIBLING proteins, bone sialoprotein, and osteopontin." Biochem Biophys Res Commun **280**(2): 460-465.
Flicek, P., I. Ahmed, et al. (2013). "Ensembl 2013." Nucleic Acids Res **41**(Database issue): D48-55.
Flicek, P., M. R. Amode, et al. (2014). "Ensembl 2014." Nucleic Acids Res **42**(1): D749-755.
Flicek, P., M. R. Amode, et al. (2012). "Ensembl 2012." Nucleic Acids Res **40**(Database issue): D84-90.

Fraczkiewicz, R. and W. Braun (1998). "Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules." Journal of Computational Chemistry **19**(3): 319-333.
Franzosa, E. A. and Y. Xia (2009). "Structural determinants of protein evolution are context-sensitive at the residue level." Mol Biol Evol **26**(10): 2387-2395.
Fraser, G. J., C. D. Hulsey, et al. (2009). "An ancient gene network is co-opted for teeth on old and new jaws." PLoS Biol **7**(2): e31.
Fraser, H. B. (2013). "Gene expression drives local adaptation in humans." Genome Res **23**(7): 1089-1096.
Frazer, K. A., L. Pachter, et al. (2004). "VISTA: computational tools for comparative genomics." Nucleic Acids Research **32**: W273-W279.
Frederiksen, S., H. Cao, et al. (1997). "The rat 5S rRNA bona fide gene repeat maps to chromosome 19q12-->qter and the pseudogene repeat maps to 12q12." Cytogenet Cell Genet **76**(1-2): 101-106.
Freeman, S., J. C. Sharp, et al. (2008). Biological science, Benjamin Cummings.
Friedman, M. S., S. M. Oyserman, et al. (2009). "Wnt11 promotes osteoblast maturation and mineralization through R-spondin 2." Journal of biological chemistry **284**(21): 14117-14125.
Frith, M. C., L. G. Wilming, et al. (2006). "Pseudo-messenger RNA: phantoms of the transcriptome." PLoS Genet **2**(4): e23.
Fukumoto, S. (2008). "Physiological regulation and disorders of phosphate metabolism--pivotal role of fibroblast growth factor 23." Intern Med **47**(5): 337-343.
Gabaldón, T. and E. V. Koonin (2013). "Functional and evolutionary implications of gene orthology." Nature Reviews Genetics **14**(5): 360-366.
Gazave, E., T. Marques-Bonet, et al. (2007). "Patterns and rates of intron divergence between humans and chimpanzees." Genome Biol **8**(2): R21.
Gill, F. B. (2007). Ornithology. 3rd, New York: WH Freeman. xxvi.
Gluhak-Heinrich, J., S. P. Kotha, et al. (2004). "In-vivo site-specific correlation of dentin matrix protein 1 (DMPI) and matrix extracellular phosphoglycoprotein (MEPE) gene expression: Effect of overload." Journal of Bone and Mineral Research **19**: S73-S73.
Goettsch, C., A. Babelova, et al. (2013). "NADPH oxidase 4 limits bone mass by promoting osteoclastogenesis." J Clin Invest **123**(11): 4731-4738.
Goldman, M., B. Craft, et al. (2013). "The UCSC Cancer Genomics Browser: update 2013." Nucleic Acids Res **41**(Database issue): D949-954.
Goncalves, I., L. Duret, et al. (2000). "Nature and structure of human genes that generate retropseudogenes." Genome Res **10**(5): 672-678.
Gouy, M., S. Guindon, et al. (2010). "SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Mol Biol Evol **27**(2): 221-224.
Gowen, L. C., D. N. Petersen, et al. (2003). "Targeted disruption of the osteoblast/osteocyte factor 45 gene (OF45) results in increased bone formation and bone mass." J Biol Chem **278**(3): 1998-2007.
Green, P., D. Lipman, et al. (1993). "Ancient conserved regions in new gene sequences and the protein databases." Science **259**(5102): 1711-1716.
Grubb, B. R. (1983). "Allometric relations of cardiovascular function in birds." American Journal of Physiology-Heart and Circulatory Physiology **245**(4): H567-H572.
Gu, X. (1999). "Statistical methods for testing functional divergence after gene duplication." Mol Biol Evol **16**(12): 1664-1674.
Gu, X. (2001). "Maximum-likelihood approach for gene family evolution under functional divergence." Mol Biol Evol **18**(4): 453-464.
Gu, X. (2003). "Functional divergence in protein (family) sequence evolution." Genetica **118**(2-3): 133-141.
Gu, X. and K. Vander Velden (2002). "DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family." Bioinformatics **18**(3): 500-501.
Guindon, S., F. Delsuc, et al. (2009). "Estimating maximum likelihood phylogenies with PhyML." Methods Mol Biol **537**: 113-137.
Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol **52**(5): 696-704.
Guo, X., M. Lin, et al. (2014). "Characterization of human pseudogene-derived non-coding RNAs for functional potential." PLoS One **9**(4): e93972.
Gutzwiller, S. C., A. Su, et al. (2013). "Postcranial pneumaticity and bone structure in two clades of neognath birds." Anat Rec (Hoboken) **296**(6): 867-876.
Hadjidakis, D. J. and Androulakis, II (2006). "Bone remodeling." Ann N Y Acad Sci **1092**: 385-396.
Hahn, M. W. (2009). "Distinguishing among evolutionary models for the maintenance of gene duplicates." Journal of Heredity **100**(5): 605-617.
Haldane, J. (1933). "The part played by recurrent mutation in evolution." American Naturalist: 5-19.
Haldane, J. B. S. (1957). "The cost of natural selection." Journal of Genetics **55**(3): 511-524.
Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symposium Series **41**: 95-98.
Hamrick, M. W. (2010). "Myostatin (GDF-8) as a therapeutic target for the prevention of osteoporotic fractures." IBMS BoneKEy **7**(1): 8-17.
Hare, M. P. and S. R. Palumbi (2003). "High intron sequence conservation across three mammalian orders suggests functional constraints." Mol Biol Evol **20**(6): 969-978.
Harris, H. (1966). "Enzyme polymorphisms in man." Proc R Soc Lond B Biol Sci **164**(995): 298-310.

Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32**(Database issue): D258-261.
Harrison, P. M., D. Zheng, et al. (2005). "Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability." Nucleic Acids Res **33**(8): 2374-2383.
Hasegawa, M., J. L. Thorne, et al. (2003). "Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution." Genes Genet Syst **78**(4): 267-283.
Hayashibara, T., T. Hiraga, et al. (2004). "A synthetic peptide fragment of human MEPE stimulates new bone formation in vitro and in vivo." J Bone Miner Res **19**(3): 455-462.
Haygood, R., C. C. Babbitt, et al. (2010). "Contrasts between adaptive coding and noncoding changes during human evolution." Proceedings of the National Academy of Sciences **107**(17): 7853-7857.
Hazel, J. R. and C. L. Prosser (1974). "Molecular mechanisms of temperature compensation in poikilotherms." Physiol Rev **54**(3): 620-677.
He, X. and J. Zhang (2005). "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution." Genetics **169**(2): 1157-1164.
Hedges, S. B., J. Dudley, et al. (2006). "TimeTree: a public knowledge-base of divergence times among organisms." Bioinformatics **22**(23): 2971-2972.
Henricson, A., K. Forslund, et al. (2010). "Orthology confers intron position conservation." BMC Genomics **11**: 412.
Hillier, L. W., W. Miller, et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.
Hincke, M. T., J. Gautron, et al. (1999). "Molecular cloning and ultrastructural localization of the core protein of an eggshell matrix proteoglycan, ovocleidin-116." J Biol Chem **274**(46): 32915-32923.
Hirayama, M., A. Kobiyama, et al. (2004). "The occurrence of two types of hemopexin-like protein in medaka and differences in their affinity to heme." J Exp Biol **207**(Pt 8): 1387-1398.
Hirayama, M., M. Nakaniwa, et al. (2003). "Primary structures and gene organizations of two types of Wap65 from the pufferfish <i>Takifugu rubripes</i>." Fish Physiol Biochem **29**(3): 211-224.
Hirsh, A. E. and H. B. Fraser (2001). "Protein dispensability and rate of evolution." Nature **411**(6841): 1046-1049.
Hogg, D. A. (1984). "The distribution of pneumatisation in the skeleton of the adult domestic fowl." J Anat **138 (Pt 4)**: 617-629.
Hone, D. W., G. J. Dyke, et al. (2008). "Body size evolution in Mesozoic birds." J Evol Biol **21**(2): 618-624.
Houlston, R. S., E. Webb, et al. (2008). "Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer." Nat Genet **40**(12): 1426-1435.
Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.
Huang da, W., B. T. Sherman, et al. (2007). "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists." Nucleic Acids Res **35**(Web Server issue): W169-175.
Hudson, R. R., M. Kreitman, et al. (1987). "A test of neutral molecular evolution based on nucleotide data." Genetics **116**(1): 153-159.
Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.
Huerta-Cepas, J. and T. Gabaldon (2011). "Assigning duplication events to relative temporal scales in genome-wide studies." Bioinformatics **27**(1): 38-45.
Hurst, L. D. (2006). "Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals." J Mol Evol **63**(2): 174-182.
Hvidberg, V., M. B. Maniecki, et al. (2005). "Identification of the receptor scavenging hemopexin-heme complexes." Blood **106**(7): 2572-2579.
Ishida, T. and K. Kinoshita (2007). "PrDOS: prediction of disordered protein regions from amino acid sequence." Nucleic Acids Res **35**(Web Server issue): W460-464.
Jacq, C., J. R. Miller, et al. (1977). "A pseudogene structure in 5S DNA of Xenopus laevis." Cell **12**(1): 109-120.
Jarvinen, E., M. Tummers, et al. (2009). "The role of the dental lamina in mammalian tooth replacement." J Exp Zool B Mol Dev Evol **312B**(4): 281-291.
Jernvall, J. and I. Thesleff (2012). "Tooth shape formation and tooth renewal: evolving with the same signals." Development **139**(19): 3487-3497.
Jonsson, J. J., M. D. Foresman, et al. (1992). "Intron requirement for expression of the human purine nucleoside phosphorylase gene." Nucleic Acids Res **20**(12): 3191-3198.
Jordan, I. K., Y. I. Wolf, et al. (2004). "Duplicated genes evolve slower than singletons despite the initial rate increase." BMC Evol Biol **4**: 22.
Jorgensen, N. R., L. B. Husted, et al. (2012). "Single-nucleotide polymorphisms in the P2X7 receptor gene are associated with post-menopausal bone loss and vertebral fractures." Eur J Hum Genet **20**(6): 675-681.
Kabza, M., J. Ciomborowska, et al. (2014). "RetrogeneDB--A Database of Animal Retrogenes." Mol Biol Evol.
Kabza, M., J. Ciomborowska, et al. (2014). "RetrogeneDB - a database of animal retrogenes." Mol Biol Evol.
Kaessmann, H. (2010). "Origins, evolution, and phenotypic impact of new genes." Genome Res **20**(10): 1313-1326.

Kalmar, L., D. Homola, et al. (2012). "Structural disorder in proteins brings order to crystal growth in biomineralization." Bone **51**(3): 528-534.
Karolchik, D., A. S. Hinrichs, et al. (2009). "The UCSC Genome Browser." Curr Protoc Bioinformatics **Chapter 1**: Unit1 4.
Katzman, S., A. D. Kern, et al. (2007). "Human genome ultraconserved elements are ultraselected." Science **317**(5840): 915.
Kawasaki, K. (2009). "The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues." Dev Genes Evol **219**(3): 147-157.
Kawasaki, K., A. V. Buchanan, et al. (2009). "Biomineralization in humans: making the hard choices in life." Annu Rev Genet **43**: 119-142.
Kawasaki, K., T. Suzuki, et al. (2004). "Genetic basis for the evolution of vertebrate mineralized tissue." Proc Natl Acad Sci U S A **101**(31): 11356-11361.
Kawasaki, K. and K. M. Weiss (2006). "Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family." J Exp Zool B Mol Dev Evol **306**(3): 295-316.
Kawasaki, K. and K. M. Weiss (2008). "SCPP gene evolution and the dental mineralization continuum." J Dent Res **87**(6): 520-531.
Keane, T. M., C. J. Creevey, et al. (2006). "Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified." BMC Evol Biol **6**(1): 29.
Keane, T. M., C. J. Creevey, et al. (2006). "Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified." Bmc Evolutionary Biology **6**.
Kelley, J. L., J. Madeoy, et al. (2006). "Genomic signatures of positive selection in humans and the limits of outlier approaches." Genome Res **16**(8): 980-989.
Kelley, J. L. and W. J. Swanson (2008). "Dietary change and adaptive evolution of enamelin in humans and among primates." Genetics **178**(3): 1595-1603.
Kelm, D. H., R. Simon, et al. (2011). "High activity enables life on a high-sugar diet: blood glucose regulation in nectar-feeding bats." Proc Biol Sci **278**(1724): 3490-3496.
Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.
Khachane, A. N. and P. M. Harrison (2009). "Assessing the genomic evidence for conserved transcribed pseudogenes under selection." BMC Genomics **10**: 435.
Khachane, A. N. and P. M. Harrison (2009). "Strong association between pseudogenization mechanisms and gene sequence length." Biol Direct **4**: 38.
Khurana, E., H. Y. Lam, et al. (2010). "Segmental duplications in the human genome reveal details of pseudogene formation." Nucleic Acids Res **38**(20): 6997-7007.
Kikuchi, K., S. Watabe, et al. (1997). "The Wap65 gene expression of goldfish (Carassius auratus) in association with warm water temperature as well as bacterial lipopolysaccharide (LPS)." Fish Physiol Biochem **17**(1-6): 423-432.
Kikuchi, K., S. Watabe, et al. (1998). "Isolation of a 65-kDa protein from white muscle of warm temperature-acclimated goldfish (Carassius auratus)." Comp Biochem Physiol B Biochem Mol Biol **120**(2): 385-391.
Kim, K. M., J. Lim, et al. (2012). "Gene expression profiling of oral epithelium during tooth development." Arch Oral Biol **57**(8): 1100-1107.
Kimchi-Sarfaty, C., J. M. Oh, et al. (2007). "A "silent" polymorphism in the MDR1 gene changes substrate specificity." Science **315**(5811): 525-528.
Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**(5129): 624-626.
Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-120.
Kin, T. and Y. Ono (2007). "Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat." Bioinformatics **23**(21): 2945-2946.
King, J. L., T. H. Jukes, et al. (1969). Non-darwinian evolution, Bobbs-Merrill.
Kinoshita, S., S. Itoi, et al. (2001). "cDNA cloning and characterization of the warm-temperature-acclimation-associated protein Wap65 from carp, Cyprinus carpio." Fish Physiol Biochem **24**(2): 125-134.
Kinsella, R. J., A. Kahari, et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space." Database (Oxford) **2011**: bar030.
Knight, R., P. Maxwell, et al. (2007). "PyCogent: a toolkit for making sense from sequence." Genome Biol **8**(8): R171.
Korrodi-Gregorio, L., J. Abrantes, et al. (2013). "Not so pseudo: the evolutionary history of protein phosphatase 1 regulatory subunit 2 and related pseudogenes." BMC Evol Biol **13**: 242.
Kosakovsky Pond, S. L. and S. D. Frost (2005). "Not so different after all: a comparison of methods for detecting amino acid sites under selection." Mol Biol Evol **22**(5): 1208-1222.
Kosakovsky Pond, S. L., B. Murrell, et al. (2011). "A random effects branch-site model for detecting episodic diversifying selection." Mol Biol Evol **28**(11): 3033-3043.
Kosakovsky Pond, S. L., A. F. Poon, et al. (2008). "A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus." Mol Biol Evol **25**(9): 1809-1824.

Kosakovsky Pond, S. L., D. Posada, et al. (2006). "GARD: a genetic algorithm for recombination detection." Bioinformatics **22**(24): 3096-3098.
Kosiol, C., T. Vinar, et al. (2008). "Patterns of positive selection in six Mammalian genomes." PLoS Genet **4**(8): e1000144.
Koussoulakou, D. S., L. H. Margaritis, et al. (2009). "A curriculum vitae of teeth: evolution, generation, regeneration." Int J Biol Sci **5**(3): 226-243.
Kryazhimskiy, S. and J. B. Plotkin (2008). "The population genetics of dN/dS." PLoS Genet **4**(12): e1000304.
Krylov, D. M., Y. I. Wolf, et al. (2003). "Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution." Genome Res **13**(10): 2229-2235.
Kumar, S. (2005). "Molecular clocks: four decades of evolution." Nat Rev Genet **6**(8): 654-662.
Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proc Natl Acad Sci U S A **99**(2): 803-808.
Lachke, S. A., J. W. Ho, et al. (2012). "iSyTE: integrated Systems Tool for Eye gene discovery." Invest Ophthalmol Vis Sci **53**(3): 1617-1627.
Lan, Y., P. D. Kingsley, et al. (2001). "Osr2, a new mouse gene related to Drosophila odd-skipped, exhibits dynamic expression patterns during craniofacial, limb, and kidney development." Mech Dev **107**(1-2): 175-179.
Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
Landin, M. A., M. Shabestari, et al. (2012). "Gene Expression Profiling during Murine Tooth Development." Front Genet **3**: 139.
Lartillot, N. and R. Poujol (2011). "A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters." Mol Biol Evol **28**(1): 729-744.
Laulederkind, S. J., G. T. Hayman, et al. (2013). "The Rat Genome Database 2013--data, tools and users." Brief Bioinform **14**(4): 520-526.
Lee, J. Y., P. J. Taub, et al. (2009). "Identification of CITED2 as a negative regulator of fracture healing." Biochemical and biophysical research communications **387**(4): 641-645.
Levasseur, A., P. Gouret, et al. (2006). "Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family." Bmc Evolutionary Biology **6**: 92.
Lewontin, R. C. and J. L. Hubby (1966). "A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura." Genetics **54**(2): 595-609.
Li, Q., H. Pan, et al. (2012). "CITED2 mutation links congenital heart defects to dysregulation of the cardiac gene VEGF and PITX2C expression." Biochem Biophys Res Commun **423**(4): 895-899.
Li, W., B. Liu, et al. (2009). "Phylogenetic analysis, structural evolution and functional divergence of the 12-oxo-phytodienoate acid reductase gene family in plants." BMC Evol Biol **9**: 90.
Li, W., W. Yang, et al. (2013). "Pseudogenes: pseudo or real functional elements?" J Genet Genomics **40**(4): 171-177.
Li, W. H., D. L. Ellsworth, et al. (1996). "Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis." Mol Phylogenet Evol **5**(1): 182-187.
Liao, C., J. Hsu, et al. (2013). "Selective inhibition of spleen tyrosine kinase (SYK) with a novel orally bioavailable small molecule inhibitor, RO9021, impinges on various innate and adaptive immune responses: implications for SYK inhibitors in autoimmune disease therapy." Arthritis Res Ther **15**(5): R146.
Lin, D., Y. Huang, et al. (2007). "Expression survey of genes critical for tooth development in the human embryonic tooth germ." Dev Dyn **236**(5): 1307-1312.
Lindblad-Toh, K., M. Garber, et al. (2011). "A high-resolution map of human evolutionary constraint using 29 mammals." Nature **478**(7370): 476-482.
Lindblad-Toh, K., M. Garber, et al. (2011). "A high-resolution map of human evolutionary constraint using 29 mammals." Nature **478**(7370): 476-482.
Lipman, D. J., A. Souvorov, et al. (2002). "The relationship of protein conservation and sequence length." BMC Evol Biol **2**: 20.
Liu, H., W. Li, et al. (2004). "Dentonin, a fragment of MEPE, enhanced dental pulp stem cell proliferation." J Dent Res **83**(6): 496-499.
Liu, S., H. Wang, et al. (2009). "MEPE/OF45 protects cells from DNA damage induced killing via stabilizing CHK1." Nucleic Acids Res **37**(22): 7447-7454.
Lopez-Maury, L., S. Marguerat, et al. (2008). "Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation." Nat Rev Genet **9**(8): 583-593.
Louis, A., M. Muffato, et al. (2013). "Genomicus: five genome browsers for comparative genomics in eukaryota." Nucleic Acids Res **41**(D1): D700-705.
Loytynoja, A. (2014). "Phylogeny-aware alignment with PRANK." Methods Mol Biol **1079**: 155-170.
Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-1155.
Lynn, D. J., A. T. Lloyd, et al. (2004). "Evidence of positively selected sites in mammalian alpha-defensins." Mol Biol Evol **21**(5): 819-827.
Maina, J. (2000). "What it takes to fly: the structural and functional respiratory refinements in birds and bats." Journal of Experimental Biology **203**(20): 3045-3064.

Majewski, J. and J. Ott (2002). "Distribution and characterization of regulatory elements in the human genome." Genome Res **12**(12): 1827-1836.
Maldonado, E., J. Y. Dutheil, et al. (2011). "IMPACT: integrated multiprogram platform for analyses in ConTest." Journal of Heredity **102**(3): 366-369.
Maldonado, E., I. Khan, et al. (2013). "EASER: Ensembl Easy Sequence Retriever." Evol Bioinform Online **9**: 487-490.
Margoliash, E. (1963). "Primary Structure and Evolution of Cytochrome C." Proc Natl Acad Sci U S A **50**: 672-679.
Markova-Raina, P. and D. Petrov (2011). "High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes." Genome Res **21**(6): 863-874.
Marques, A. C., J. Tan, et al. (2012). "Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs." Genome Biol **13**(11): R102.
Martin, A., V. David, et al. (2008). "Degradation of MEPE, DMP1, and release of SIBLING ASARM-peptides (minhibins): ASARM-peptide(s) are directly responsible for defective mineralization in HYP." Endocrinology **149**(4): 1757-1772.
Martin, G. and T. Lenormand (2006). "The fitness effect of mutations across environments: a survey in light of fitness landscape models." Evolution **60**(12): 2413-2427.
Martinez-Arias, R., E. Mateu, et al. (2001). "Profiles of accepted mutation: from neutrality in a pseudogene to disease-causing mutation on its homologous gene." Hum Genet **109**(1): 7-10.
Martinez-Castilla, L. P. and E. R. Alvarez-Buylla (2003). "Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny." Proc Natl Acad Sci U S A **100**(23): 13407-13412.
Massingham, T. and N. Goldman (2005). "Detecting amino acid sites under positive selection and purifying selection." Genetics **169**(3): 1753-1762.
MATLAB. (2012). MATLAB and Statistics Toolbox Release 2012b. Natick, Massachusetts, United States.
Maurer, B. A., J. H. Brown, et al. (2004). "Similarities in body size distributions of small-bodied flying vertebrates." Evolutionary Ecology Research **6**(6): 783-797.
McClellan, D. A., E. J. Palfreyman, et al. (2005). "Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins." Mol Biol Evol **22**(3): 437-455.
McClellan, D. A., E. J. Palfreyman, et al. (2005). "Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins." Mol Biol Evol **22**(3): 437-455.
McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-654.
McKnight, D. A. and L. W. Fisher (2009). "Molecular evolution of dentin phosphoprotein among toothed and toothless animals." Bmc Evolutionary Biology **9**: -.
Meredith, R. W., J. E. Janecka, et al. (2011). "Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification." Science **334**(6055): 521-524.
Messier, W. and C. B. Stewart (1997). "Episodic adaptive evolution of primate lysozymes." Nature **385**(6612): 151-154.
Michalak, P. (2008). "Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes." Genomics **91**(3): 243-248.
Mighell, A. J., N. R. Smith, et al. (2000). "Vertebrate pseudogenes." FEBS Lett **468**(2-3): 109-114.
Mikkola, M. L. (2009). "Controlling the number of tooth rows." Sci Signal **2**(85): pe53.
Milinkovitch, M. C., R. Helaers, et al. (2010). "2x genomes--depth does matter." Genome Biol **11**(2): R16.
Mina, M. and E. J. Kollar (1987). "The induction of odontogenesis in non-dental mesenchyme combined with early murine mandibular arch epithelium." Arch Oral Biol **32**(2): 123-127.
Mitsiadis, T. A. and D. Graf (2009). "Cell fate determination during tooth development and regeneration." Birth Defects Res C Embryo Today **87**(3): 199-211.
Montgomery, M. K., A. Hulbert, et al. (2012). "Does the oxidative stress theory of aging explain longevity differences in birds? I. Mitochondrial ROS production." Experimental gerontology **47**(3): 203-210.
Montoya-Burgos, J. I. (2011). "Patterns of positive selection and neutral evolution in the protein-coding genes of Tetraodon and Takifugu." PLoS One **6**(9): e24800.
Morgan, C. C., N. B. Loughran, et al. (2010). "Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins." Bmc Evolutionary Biology **10**: 39.
Morgan, C. C., K. Shakya, et al. (2012). "Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions." BMC Evol Biol **12**: 114.
Morgan, W. T., H. H. Liem, et al. (1976). "Transfer of heme from heme-albumin to hemopexin." Biochim Biophys Acta **444**(2): 435-445.
Morris, C. R., F. E. Nelson, et al. (2010). "The metabolic power requirements of flight and estimations of flight muscle efficiency in the cockatiel (Nymphicus hollandicus)." J Exp Biol **213**(Pt 16): 2788-2796.
Morrison, D. A. (2014). "On the Organic Law of Change: A Facsimile Edition and Annotated Transcription of Alfred Russel Wallace's Species Notebook of 1855–1859.—Edited and annotated by James T. Costa." Syst Biol: syt112.
Moury, B. and V. Simon (2011). "dN/dS-Based Methods Detect Positive Selection Linked to Trade-Offs between Different Fitness Traits in the Coat Protein of Potato virus Y." Mol Biol Evol **28**(9): 2707-2717.
Muller-Eberhard, U. (1988). "Hemopexin." Methods Enzymol **163**: 536-565.
Muller, H. (1935). "The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere." Genetica **17**(3): 237-252.

Munshi-South, J. and G. S. Wilkinson (2010). "Bats and birds: Exceptional longevity despite high metabolic rates." Ageing Res Rev **9**(1): 12-19.
Muse, S. V. and B. S. Gaut (1997). "Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test." Genetics **146**(1): 393-399.
Myers, P., R. Espinosa, et al. (2014). "The Animal Diversity Web (online)." Accessed at http://animaldiversity.org. Retrieved 2014.
Nakaniwa, M., M. Hirayama, et al. (2005). "Genomic sequences encoding two types of medaka hemopexin-like protein Wap65, and their gene expression profiles in embryos." J Exp Biol **208**(Pt 10): 1915-1925.
Navarro, F. C. and P. A. Galante (2013). "RCPedia: a database of retrocopied genes." Bioinformatics **29**(9): 1235-1237.
Near, T. J., R. I. Eytan, et al. (2012). "Resolution of ray-finned fish phylogeny and timing of diversification." Proc Natl Acad Sci U S A **109**(34): 13698-13703.
Nehrt, N. L., W. T. Clark, et al. (2011). "Testing the ortholog conjecture with comparative functional genomic data from mammals." PLoS Comput Biol **7**(6): e1002073.
Nery, M. F., D. J. Gonzalez, et al. (2013). "How to Make a Dolphin: Molecular Signature of Positive Selection in Cetacean Genome." PLoS One **8**(6): e65491.
Nielsen, R., C. Bustamante, et al. (2005). "A scan for positively selected genes in the genomes of humans and chimpanzees." PLoS Biol **3**(6): e170.
Nielsen, R. and Z. Yang (1998). "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene." Genetics **148**(3): 929-936.
Nilsson, J., M. Grahn, et al. (2011). "Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins." Genome Biol **12**(7): R65.
Nishihara, H., M. Hasegawa, et al. (2006). "Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions." Proc Natl Acad Sci U S A **103**(26): 9929-9934.
Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-217.
Nowak, M. A., M. C. Boerlijst, et al. (1997). "Evolution of genetic redundancy." Nature **388**(6638): 167-171.
Nylander, J. A. A. (2004). MrAIC.pl. Program distributed by the author. Evolutionary Biology Centre, Uppsala.
Nylander, J. A. A. (2004). MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.
Ogbureke, K. U. and L. W. Fisher (2007). "SIBLING expression patterns in duct epithelia reflect the degree of metabolic activity." J Histochem Cytochem **55**(4): 403-409.
Ohta, T. (1989). "Role of gene duplication in evolution." Genome **31**(1): 304-310.
Ohta, T. (1992). "The Nearly Neutral Theory of Molecular Evolution." Annual Review of Ecology and Systematics **23**: 263-286.
Oliveros, J. C. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html.
Olson, S. L. and A. Feduccia (1979). "Flight capability and the pectoral girdle of Archaeopteryx."
Oshima, R. G., L. Abrams, et al. (1990). "Activation of an intron enhancer within the keratin 18 gene by expression of c-fos and c-jun in undifferentiated F9 embryonal carcinoma cells." Genes Dev **4**(5): 835-848.
Osorio, D. S., A. Antunes, et al. (2007). "Structural and functional implications of positive selection at the primate angiogenin gene." Bmc Evolutionary Biology **7**: -.
Padian, K. and L. M. Chiappe (1998). "The origin of birds and their flight." Scientific American **278**(2): 28-37.
Page, A. E., K. Fuller, et al. (1993). "Purification and characterization of a tripeptidyl peptidase I from human osteoclastomas: evidence for its role in bone resorption." Arch Biochem Biophys **306**(2): 354-359.
Paoli, M., B. F. Anderson, et al. (1999). "Crystal structure of hemopexin reveals a novel high-affinity heme site formed between two beta-propeller domains." Nat Struct Biol **6**(10): 926-931.
Paoli, M., J. Marles-Wright, et al. (2002). "Structure-function relationships in heme-proteins." DNA Cell Biol **21**(4): 271-280.
Parfitt, A. M. (2002). "Targeted and nontargeted bone remodeling: relationship to basic multicellular unit origination and progression." Bone **30**(1): 5-7.
Park, C., W. Qian, et al. (2012). "Genomic evidence for elevated mutation rates in highly expressed genes." EMBO Rep **13**(12): 1123-1129.
Parker, J., G. Tsagkogeorga, et al. (2013). "Genome-wide signatures of convergent evolution in echolocating mammals." Nature **502**(7470): 228-231.
Pavlicek, A., A. J. Gentles, et al. (2006). "Retroposition of processed pseudogenes: the impact of RNA stability and translational control." Trends Genet **22**(2): 69-73.
Peatman, E., P. Baoprasertkul, et al. (2007). "Expression analysis of the acute phase response in channel catfish (Ictalurus punctatus) after infection with a Gram-negative bacterium." Dev Comp Immunol **31**(11): 1183-1196.
Peatman, E., J. Terhune, et al. (2008). "Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with Edwardsiella ictaluri." Mol Immunol **45**(2): 553-566.
Pei, B., C. Sisu, et al. (2012). "The GENCODE pseudogene resource." Genome Biol **13**(9): R51.
Pemberton, T. J., F. Y. Li, et al. (2007). "Identification of novel genes expressed during mouse tooth development by microarray gene expression analysis." Dev Dyn **236**(8): 2245-2257.

Penn, O., E. Privman, et al. (2010). "An alignment confidence score capturing robustness to guide tree uncertainty." Mol Biol Evol **27**(8): 1759-1767.
Perelman, P., W. E. Johnson, et al. (2011). "A Molecular Phylogeny of Living Primates." PLoS Genet **7**(3).
Petersen, D. N., G. T. Tkalcevic, et al. (2000). "Identification of osteoblast/osteocyte factor 45 (OF45), a bone-specific cDNA encoding an RGD-containing protein that is highly expressed in osteoblasts and osteocytes." J Biol Chem **275**(46): 36172-36180.
Philip, S., J. P. Machado, et al. (2012). "Fish lateral line innovation: insights into the evolutionary genomic dynamics of a unique mechanosensory organ." Mol Biol Evol **29**(12): 3887-3898.
Piehler, A. P., M. Hellum, et al. (2008). "The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference." BMC Genomics **9**: 165.
Podlaha, O. and J. Zhang (2010). "Pseudogenes and their evolution." eLS.
Podulka, S., R. W. Rohrbaugh, et al. (2004). Handbook of bird biology, Cornell Lab of Ornithology Ithaca, New York.
Poliseno, L., L. Salmena, et al. (2010). "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology." Nature **465**(7301): 1033-1038.
Pollard, K. S., M. J. Hubisz, et al. (2010). "Detection of nonneutral substitution rates on mammalian phylogenies." Genome Res **20**(1): 110-121.
Pond, S. K. and S. V. Muse (2005). "Site-to-site variation of synonymous substitution rates." Mol Biol Evol **22**(12): 2375-2385.
Pond, S. L. and S. D. Frost (2005). "Datamonkey: rapid detection of selective pressure on individual sites of codon alignments." Bioinformatics **21**(10): 2531-2533.
Pond, S. L., S. D. Frost, et al. (2005). "HyPhy: hypothesis testing using phylogenies." Bioinformatics **21**(5): 676-679.
Ponting, C. P. and G. Lunter (2006). "Signatures of adaptive evolution within human non-coding sequence." Hum Mol Genet **15**(suppl 2): R170-R175.
Poon, A. F., S. D. Frost, et al. (2009). "Detecting signatures of selection from DNA sequences using Datamonkey." Methods Mol Biol **537**: 163-183.
Porollo, A. A., R. Adamczak, et al. (2004). "POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins." Bioinformatics **20**(15): 2460-2462.
Porter, M. L., T. W. Cronin, et al. (2007). "Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins." Mol Biol Evol **24**(1): 253-268.
Posada, D. (2008). "jModelTest: phylogenetic model averaging." Mol Biol Evol **25**(7): 1253-1256.
Posada, D. and K. A. Crandall (1998). "MODELTEST: testing the model of DNA substitution." Bioinformatics **14**(9): 817-818.
Prange, H. D., J. F. Anderson, et al. (1979). "Scaling of skeletal mass to body mass in birds and mammals." American Naturalist: 103-122.
Pride, D. T. and M. J. Blaser (2002). "Concerted evolution between duplicated genetic elements in Helicobacter pylori." J Mol Biol **316**(3): 629-642.
Privman, E., O. Penn, et al. (2012). "Improving the performance of positive selection inference by filtering unreliable alignment regions." Mol Biol Evol **29**(1): 1-5.
Prosdocimi, F., B. Linard, et al. (2012). "Controversies in modern evolutionary biology: the imperative for error detection and quality control." BMC Genomics **13**: 5.
Proux, E., R. A. Studer, et al. (2009). "Selectome: A database of positive selection." Nucleic Acids Research **37**(SUPPL. 1): D404-D407.
Puttick, M. N., G. H. Thomas, et al. (2014). "High Rates of Evolution Preceded the Origin of Birds." Evolution.
Quarles, L. D. (2003). "FGF23, PHEX, and MEPE regulation of phosphate homeostasis and skeletal mineralization." Am J Physiol Endocrinol Metab **285**(1): E1-9.
Ramialison, M., B. Bajoghli, et al. (2009). "Identification of Starmaker-Like in Medaka as a Putative Target Gene of Pax2 in the Otic Vesicle." Developmental Dynamics **238**(11): 2860-2866.
Ranwez, V., S. Harispe, et al. (2011). "MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons." PLoS One **6**(9): e22594.
Rastogi, S. and D. A. Liberles (2005). "Subfunctionalization of duplicated genes as a transition state to neofunctionalization." BMC Evol Biol **5**: 28.
Reijnders, C. M., H. W. van Essen, et al. (2013). "Increased expression of matrix extracellular phosphoglycoprotein (MEPE) in cortical bone of the rat tibia after mechanical loading: identification by oligonucleotide microarray." PLoS One **8**(11): e79672.
Renaut, S., A. W. Nolte, et al. (2010). "Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae)." Mol Ecol **19 Suppl 1**: 115-131.
Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-277.
Ridout, K. E., C. J. Dixon, et al. (2010). "Positive selection differs between protein secondary structure elements in Drosophila." Genome Biol Evol **2**: 166-179.
Riggs, B. L. (2000). "The mechanisms of estrogen regulation of bone resorption." Journal of clinical Investigation **106**(10): 1203-1204.
Rivadeneira, F., U. Styrkarsdottir, et al. (2009). "Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies." Nat Genet **41**(11): 1199-1206.

Robinson-Rechavi, M. and D. Huchon (2000). "RRTree: relative-rate tests between groups of sequences on a phylogenetic tree." Bioinformatics **16**(3): 296-297.
Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-1574.
Roots, C. (2006). Flightless birds, Greenwood Publishing Group.
Rowe, P. S. (2004). "The wrickkened pathways of FGF23, MEPE and PHEX." Crit Rev Oral Biol Med **15**(5): 264-281.
Rowe, P. S. (2012). "Regulation of Bone− Renal Mineral and Energy Metabolism: The PHEX, FGF23, DMP1, MEPE ASARM Pathway." Critical Reviews™ in Eukaryotic Gene Expression **22**(1).
Rowe, P. S., P. A. de Zoysa, et al. (2000). "MEPE, a new gene expressed in bone marrow and tumors causing osteomalacia." Genomics **67**(1): 54-68.
Rowe, P. S., Y. Kumagai, et al. (2004). "MEPE has the properties of an osteoblastic phosphatonin and minhibin." Bone **34**(2): 303-319.
Rowe, P. S. N., Y. Kumagai, et al. (2003). "MEPE regulates bone mineralization and phosphate transport: PHEX and the MEPE ASARM-peptide." Journal of Bone and Mineral Research **18**: S24-S24.
Rowe, P. S. N., N. Matsumoto, et al. (2005). "MEPE-ASARM-peptide associated mineralization defects in X-linked hypophosphatemic rickets (hyp) is corrected by protease-inhibitors." Journal of Bone and Mineral Research **20**(9): S42-S42.
Sakai, H., K. O. Koyanagi, et al. (2007). "Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes." Gene **389**(2): 196-203.
Salazar-Ciudad, I. and J. Jernvall (2004). "How different types of pattern formation mechanisms affect the evolution of form and development." Evol Dev **6**(1): 6-16.
Sarropoulou, E., J. M. Fernandes, et al. (2010). "Evolution of a multifunctional gene: The warm temperature acclimation protein Wap65 in the European seabass Dicentrarchus labrax." Mol Phylogenet Evol **55**(2): 640-649.
Schepelmann, K. (1990). "Erythropoietic bone marrow in the pigeon: development of its distribution and volume during growth and pneumatization of bones." Journal of morphology **203**(1): 21-34.
Schmidt-Wellenburg, C. A., S. Engel, et al. (2008). "Energy expenditure during flight in relation to body mass: effects of natural increases in mass and artificial load in Rose Coloured Starlings." J Comp Physiol B **178**(6): 767-777.
Schmidt, H. A., K. Strimmer, et al. (2002). "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics **18**(3): 502-504.
Schultz, J., R. R. Copley, et al. (2000). "SMART: a web-based tool for the study of genetically mobile domains." Nucleic Acids Res **28**(1): 231-234.
Sears, K. E., R. R. Behringer, et al. (2006). "Development of bat flight: morphologic and molecular evolution of bat wing digits." Proc Natl Acad Sci U S A **103**(17): 6581-6586.
Sha, Z., P. Xu, et al. (2008). "The warm temperature acclimation protein Wap65 as an immune response gene: its duplicates are differentially regulated by temperature and bacterial infections." Mol Immunol **45**(5): 1458-1469.
Shatsky, M., R. Nussinov, et al. (2004). "A method for simultaneous alignment of multiple protein structures." Proteins **56**(1): 143-156.
Shen, B., X. Han, et al. (2012). "Adaptive evolution in the glucose transporter 4 gene Slc2a4 in Old World fruit bats (family: Pteropodidae)." PLoS One **7**(4): e33197.
Shen, Y. Y., L. Liang, et al. (2010). "Adaptive evolution of energy metabolism genes and the origin of flight in bats." Proc Natl Acad Sci U S A **107**(19): 8666-8671.
Shen, Y. Y., P. Shi, et al. (2009). "Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability." Genome Res **19**(10): 1760-1765.
Shi, Y. H., J. Chen, et al. (2010). "Molecular cloning of liver Wap65 cDNA in ayu (Plecoglossus altivelis) and mRNA expression changes following Listonella anguillarum infection." Mol Biol Rep **37**(3): 1523-1529.
Shimoyama, M., J. R. Smith, et al. (2011). "RGD: a comparative genomics platform." Hum Genomics **5**(2): 124-129.
Shoichet, B. K., W. A. Baase, et al. (1995). "A relationship between protein stability and protein function." Proc Natl Acad Sci U S A **92**(2): 452-456.
Simon-Sanchez, J., C. Schulte, et al. (2009). "Genome-wide association study reveals genetic risk underlying Parkinson's disease." Nat Genet **41**(12): 1308-1312.
Smith, F. A., J. H. Brown, et al. (2004). "Similarity of mammalian body size across the taxonomic hierarchy and across space and time." Am Nat **163**(5): 672-691.
Smith, M. M. and M. I. Coates (1998). "Evolutionary origins of the vertebrate dentition: phylogenetic patterns and developmental evolution." Eur J Oral Sci **106 Suppl 1**: 482-500.
Smith, N. D. (2012). "Body mass and foraging ecology predict evolutionary patterns of skeletal pneumaticity in the diverse "waterbird" clade." Evolution **66**(4): 1059-1078.
Smith, S. A. and C. W. Dunn (2008). "Phyutility: a phyloinformatics tool for trees, alignments and molecular data." Bioinformatics **24**(5): 715-716.
Smith, T. D., J. B. Rossie, et al. (2005). "Secondary pneumatization of the maxillary sinus in callitrichid primates: insights from immunohistochemistry and bone cell distribution." Anat Rec A Discov Mol Cell Evol Biol **285**(2): 677-689.

Sollner, C., M. Burghammer, et al. (2003). "Control of crystal size and lattice formation by starmaker in otolith biomineralization." Science **302**(5643): 282-286.
Somarowthu, S. and M. J. Ondrechen (2012). "POOL server: machine learning application for functional site prediction in proteins." Bioinformatics **28**(15): 2078-2079.
Springer, M. S., W. J. Murphy, et al. (2003). "Placental mammal diversification and the Cretaceous-Tertiary boundary." Proc Natl Acad Sci U S A **100**(3): 1056-1061.
SPSS, S., v20 (Released 2011). IBM SPSS Statistics for Windows, Armonk, NY: IBM Corp.
Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-539.
Steinke, D., W. Salzburger, et al. (2006). "Many genes in fish have species-specific asymmetric rates of molecular evolution." BMC Genomics **7**: 20.
Stern, A., A. Doron-Faigenboim, et al. (2007). "Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach." Nucleic Acids Research **35**: W506-W511.
Strimmer, K. and A. von Haeseler (1997). "Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment." Proc Natl Acad Sci U S A **94**(13): 6815-6819.
Subramanian, S. and S. Kumar (2004). "Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome." Genetics **168**(1): 373-381.
Suchard, M. A. and B. D. Redelings (2006). "BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny." Bioinformatics **22**(16): 2047-2048.
Sun, Y. B., W. P. Zhou, et al. (2013). "Genome-wide scans for candidate genes involved in the aquatic adaptation of dolphins." Genome Biol Evol **5**(1): 130-139.
Svensson, O., L. Arvestad, et al. (2006). "Genome-wide survey for biologically functional pseudogenes." PLoS Comput Biol **2**(5): e46.
Swanson, W. J. (2003). "Adaptive evolution of genes and gene families." Current opinion in genetics & development **13**(6): 617-622.
Swanson, W. J., R. Nielsen, et al. (2003). "Pervasive adaptive evolution in mammalian fertilization proteins." Mol Biol Evol **20**(1): 18-20.
Szwergold, B. S. and C. B. Miller (2013). "Potential of birds to serve as a pathology-free model of Type 2 Diabetes, 1: Is the apparent absence of the RAGE gene a factor in the resistance of avian organisms to chronic hyperglycemia?" Rejuvenation research(ja).
Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-595.
Takahashi, N., Y. Takahashi, et al. (1985). "Complete amino acid sequence of human hemopexin, the heme-binding protein of serum." Proc Natl Acad Sci U S A **82**(1): 73-77.
Takano, T., Z. Sha, et al. (2008). "The two channel catfish intelectin genes exhibit highly differential patterns of tissue expression and regulation after infection with Edwardsiella ictaluri." Dev Comp Immunol **32**(6): 693-705.
Talavera, G. and J. Castresana (2007). "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments." Syst Biol **56**(4): 564-577.
Tam, O. H., A. A. Aravin, et al. (2008). "Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes." Nature **453**(7194): 534-538.
Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**(8): 1596-1599.
Tamura, K., M. Nei, et al. (2004). "Prospects for inferring very large phylogenies by using the neighbor-joining method." Proc Natl Acad Sci U S A **101**(30): 11030-11035.
Tanaka, S. M., H. B. Sun, et al. (2004). "Bone formation induced by a novel form of mechanical loading on joint tissue." Biol Sci Space **18**(2): 41-44.
Taylor, A. C., M. Horvat-Gordon, et al. (2013). "The effects of melatonin on the physical properties of bones and egg shells in the laying hen." PLoS One **8**(2): e55663.
Thesleff, I. (2003). "Epithelial-mesenchymal signalling regulating tooth morphogenesis." J Cell Sci **116**(Pt 9): 1647-1648.
Thesleff, I. (2006). "The genetic basis of tooth development and dental defects." Am J Med Genet A **140**(23): 2530-2535.
Thomas, S. P., D. B. Follette, et al. (1995). "Metabolic and ventilatory adjustments and tolerance of the bat Pteropus poliocephalus to acute hypoxic stress." Comp Biochem Physiol A Physiol **112**(1): 43-54.
Thompson, A. A. and L. T. Nguyen (2000). "Amegakaryocytic thrombocytopenia and radio-ulnar synostosis are associated with HOXA11 mutation." Nat Genet **26**(4): 397-398.
Thompson, J. D., T. J. Gibson, et al. (2002). "Multiple sequence alignment using ClustalW and ClustalX." Curr Protoc Bioinformatics **Chapter 2**: Unit 2 3.
Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.
Tobimatsu, K., T. Noguchi, et al. (2009). "Overexpression of the transcriptional coregulator Cited2 protects against glucocorticoid-induced atrophy of C2C12 myotubes." Biochemical and biophysical research communications **378**(3): 399-403.
Toll-Riera, M., D. Bostick, et al. (2012). "Structure and age jointly influence rates of protein evolution." PLoS Comput Biol **8**(5): e1002542.

Vaidya, G., D. J. Lohman, et al. (2011). "SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information." Cladistics **27**(2): 171-180.
Vaquero, A. R., N. E. Ferreira, et al. (2012). "Using gene-network landscape to dissect genotype effects of TCF7L2 genetic variant on diabetes and cardiovascular risk." Physiological genomics **44**(19): 903-914.
Vishnoi, A., S. Kryazhimskiy, et al. (2010). "Young proteins experience more variable selection pressures than old proteins." Genome Res **20**(11): 1574-1581.
Wagner, A. (2008). "Neutralism and selectionism: a network-based reconciliation." Nat Rev Genet **9**(12): 965-974.
Wakeley, J. (1996). "The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance." Trends Ecol Evol **11**(4): 158-162.
Wall, D. P., A. E. Hirsh, et al. (2005). "Functional genomic analysis of the rates of protein evolution." Proc Natl Acad Sci U S A **102**(15): 5483-5488.
Wang, Z. and J. Zhang (2009). "Why is the correlation between gene importance and gene evolutionary rate so weak?" PLoS Genet **5**(1): e1000329.
Ward, S., C. M. Bishop, et al. (2002). "Heart rate and the rate of oxygen consumption of flying and walking barnacle geese (Branta leucopsis) and bar-headed geese (Anser indicus)." J Exp Biol **205**(Pt 21): 3347-3356.
Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.
Wedel, M. J. and M. P. Taylor (2013). "Caudal pneumaticity and pneumatic hiatuses in the sauropod dinosaurs giraffatitan and apatosaurus." PLoS One **8**(10): e78213.
Wells, J. (2013). "Not junk after all: non-protein-coding DNA carries extensive biological information." Biological Information—New Perspectives. World Scientific, Singapore: 210-231.
Wesselius, A., M. J. Bours, et al. (2013). "Association of P2X7 receptor polymorphisms with bone mineral density and osteoporosis risk in a cohort of Dutch fracture patients." Osteoporos Int **24**(4): 1235-1246.
Witmer, L. M. (1997). "The evolution of the antorbital cavity of archosaurs: a study in soft-tissue reconstruction in the fossil record with an analysis of the function of pneumaticity." Journal of Vertebrate Paleontology **17**(S1): 1-76.
Woo, Y. H., M. Walker, et al. (2010). "Coordinated expression domains in mammalian genomes." PLoS One **5**(8): e12158.
Woolley, S., J. Johnson, et al. (2003). "TreeSAAP: selection on amino acid properties using phylogenetic trees." Bioinformatics **19**(5): 671-672.
Wu, C. H., R. Apweiler, et al. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-191.
Wu, C. I. and W. H. Li (1985). "Evidence for higher rates of nucleotide substitution in rodents than in man." Proc Natl Acad Sci U S A **82**(6): 1741-1745.
Xia, X. and Z. Xie (2001). "DAMBE: Software package for data analysis in molecular biology and evolution." Journal of Heredity **92**(4): 371-373.
Yamada, K. M. (1991). "Adhesive recognition sequences." J Biol Chem **266**(20): 12809-12812.
Yamanaka, K. and M. Inouye (2001). "Selective mRNA degradation by polynucleotide phosphorylase in cold shock adaptation in Escherichia coli." J Bacteriol **183**(9): 2808-2816.
Yang, Y. J., Y. B. Wang, et al. (2007). "AHSG gene polymorphisms are associated with bone mineral density in Caucasian nuclear families." Eur J Epidemiol **22**(8): 527-532.
Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-556.
Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.
Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." Trends Ecol Evol **15**(12): 496-503.
Yang, Z. and W. J. Swanson (2002). "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes." Mol Biol Evol **19**(1): 49-57.
Yang, Z., W. S. Wong, et al. (2005). "Bayes empirical bayes inference of amino acid sites under positive selection." Mol Biol Evol **22**(4): 1107-1118.
Yang, Z. H. (2007). "PAML 4: Phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.
Yang, Z. H., W. S. W. Wong, et al. (2005). "Bayes empirical Bayes inference of amino acid sites under positive selection." Mol Biol Evol **22**(4): 1107-1118.
Yap, V. B., H. Lindsay, et al. (2010). "Estimates of the effect of natural selection on protein-coding content." Mol Biol Evol **27**(3): 726-734.
Yi, H., L. Xue, et al. (2010). "Gene expression atlas for human embryogenesis." FASEB J **24**(9): 3341-3350.
Yokoyama, S., T. Tada, et al. (2008). "Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates." Proc Natl Acad Sci U S A **105**(36): 13480-13485.
Yoshitake, H., S. R. Rittling, et al. (1999). "Osteopontin-deficient mice are resistant to ovariectomy-induced bone resorption." Proc Natl Acad Sci U S A **96**(14): 8156-8160.
Zhang, G. "Comparative Genomics Across Modern Bird Species Reveal Insights into Avian Genome Evolution and Adaptation." Submitted.
Zhang, G., C. Cowled, et al. (2013). "Comparative analysis of bat genomes provides insight into the evolution of flight and immunity." Science **339**(6118): 456-460.

Zhang, J., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Mol Biol Evol **22**(12): 2472-2479.
Zhang, J., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Mol Biol Evol **22**(12): 2472-2479.
Zhang, J., Y.-p. Zhang, et al. (2002). "Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey." Nature Genetics **30**(4): 411-415.
Zhang, J. and Y. P. Zhang (2003). "Pseudogenization of the tumor-growth promoter angiogenin in a leaf-eating monkey." Gene **308**: 95-101.
Zhang, J. Z., R. Nielsen, et al. (2005). "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level." Mol Biol Evol **22**(12): 2472-2479.
Zhang, Q. and S. V. Edwards (2012). "The evolution of intron size in amniotes: a role for powered flight?" Genome Biol Evol **4**(10): 1033-1043.
Zhang, T., E. Faraggi, et al. (2012). "SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method." J Biomol Struct Dyn **29**(4): 799-813.
Zhang, X., M. Wang, et al. (2010). "Molecular selection and functional divergence of HIF-alpha proteins in vertebrates." Genetica **138**(11-12): 1241-1250.
Zhang, Y. (2008). "I-TASSER server for protein 3D structure prediction." BMC Bioinformatics **9**: 40.
Zhang, Y. D., Z. Chen, et al. (2005). "Making a tooth: growth factors, transcription factors, and stem cells." Cell Res **15**(5): 301-316.
Zhang, Z., N. Carriero, et al. (2004). "Comparative analysis of processed pseudogenes in the mouse and human genomes." Trends Genet **20**(2): 62-67.
Zhang, Z. and M. Gerstein (2004). "Large-scale analysis of pseudogenes in the human genome." Curr Opin Genet Dev **14**(4): 328-335.
Zhang, Z., P. Harrison, et al. (2002). "Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome." Genome Res **12**(10): 1466-1482.
Zhang, Z., P. M. Harrison, et al. (2003). "Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome." Genome Res **13**(12): 2541-2558.
Zhang, Z. D., P. Cayting, et al. (2008). "Analysis of nuclear receptor pseudogenes in vertebrates: how the silent tell their stories." Mol Biol Evol **25**(1): 131-143.
Zheng, D., A. Frankish, et al. (2007). "Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution." Genome Res **17**(6): 839-851.
Zheng, Y., D. Xu, et al. (2007). "Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits." J Exp Zool B Mol Dev Evol **308**(1): 85-96.
Zhou, X., F. Sun, et al. (2013). "Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations." Nature Communications **4**.
Zirkle, C. (1941). Natural selection before the Origin of species, American Philosophical Society.
Zuckercandl, E. and L. Pauling (1965). "Evolutionary divergence and convergence in proteins."

Appendices

7.1 Appendices II - Chapter 2

http://goo.gl/pCUSLJ

7.2 Appendices III - Chapter 3

http://goo.gl/FNusK6

7.3 Appendices IV - Chapter 4

http://goo.gl/VNbGRb

7.4 Appendices V - Chapter 5

http://goo.gl/yjqVN5

7.5 Appendices VI - Chapter 6

http://goo.gl/d518fb

7.6 Appendices VII – Code developed

http://goo.gl/42ctZ9