# EVOLUTIONARY GENOMICS AND ADAPTIVE EVOLUTION OF THE HEDGEHOG GENE FAMILY IN VERTEBRATES

JOANA MARIA SOARES PEREIRA

Dissertação de Mestrado em Bioquímica

Universidade do Porto
Faculdade de Ciências
Instituto de Ciências Biomédicas Abel Salazar

2011/2012

JOANA MARIA SOARES PEREIRA

# EVOLUTIONARY GENOMICS AND ADAPTIVE EVOLUTION OF THE HEDGEHOG GENE FAMILY IN VERTEBRATES

Dissertação de candidatura ao grau de Mestre em Bioquímica da Universidade do Porto

Orientador – Prof. Doutor Agostinho Antunes

Categoria – Investigador Auxiliar e Professor Auxiliar Convidado

Afiliação – CIIMAR/FCUP

Co-orientador – Prof. Doutor Vitor Vasconcelos

Categoria – Professor Catedrático

Afiliação – CIIMAR/FCUP

**2012**

## ACKNOWLEDGEMENTS

After all those years working on the Laboratory of Ecotoxicology, Genomics and Evolution (LEGE) at CIIMAR, I have a great list of people who contributed in some way for this thesis, and to which I would like to express my gratitude.

First of all, I would like thank my supervisor, Prof. Doutor Agostinho Antunes, for his insights, helpful discussions and text corrections, interest and cooperation during this work, and to my co-supervisor, Prof. Doutor Victor Vasconcelos, for kindly welcoming me in his lab and co-supervising me during this time. However, I also would like to thank to my supervisors and CIIMAR for all the logistical support and financial aid for this work to be possible and to participate in scientific meetings, workshops and international conferences. I am extremely grateful to all LEGE members, for all the support and excellent working environment. I specially would like to thank Drs. Rui Borges, João Paulo Machado and Siby Philip for helpful discussions and support regarding the computational methodologies applied and Drs. Cidália Gomes, Bárbara Frazão, Anoop Alex and Andreia Fernandes for all the cooperation and insightful experimental tips and discussions. Rui, Cidália and Daniela Almeida will always have a special place on this list due to their amazing friendship.

My family and friends were also essential for this thesis to be completed. I want to thank to my parents, Maria de Fátima Soares and José Paulo Pereira, and my brother, João Paulo Pereira, for their unconditional love, patience and support during my under and graduate studies, but also during all my life. To my uncles, Alice Soares and Domingos Ferreira, and cousins, Carla Ferreira and Pedro Ferreira, for all their words of hope and strength, their roof and their help on providing me chicken samples. And to my college friends, Ana Rita Carvalho and Cátia Carvalho, for all the good years at FCUP, all the lunches, all the words, all the support and all the hugs. Finally but not least, I owe my deepest gratitude to my boyfriend, João Barros, not only for all the days beside me and for his unconditional friendship, but also for all the computational help and informatical skills that he provided me and made possible some of the analysis to be completed. To all of them I dedicate this thesis!

Finally, I would like to thank to the two external institutions who supplied us a great part of the data studied: to BGI, who kindly provided us access to draft genome assemblies resulted from their Avian Phylogenomics Project, and to Parque Biológico de Gaia, who supplied us with peregrine falcon fresh blood; and Fundação Para a Ciência e Tecnologia (FCT), who financed this study with the projects PTDC/BIA-BDE/69144/2006 (FCOMP-01-

0124-FEDER-007065) and PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610).

## ABSTRACT

### Evolutionary Genomics and Adaptive Evolution of the Hedgehog Gene Family in Vertebrates

The Hedgehog gene family is one of the most important family of genes involved in key developmental and homeostatic events, encoding a class of highly conserved secreted proteins that act as signaling molecules in all metazoans. These proteins play numerous roles in the regulation of cell growth and patterning during the embryonic and postembryonic development of several animals, from simple invertebrates to humans. Most bilaterians, with the exception of *C. elegans,* have been shown to possess at least one *Hh* gene, with the genome expansions in vertebrates giving rise to at least three *Hh* genes with different functional roles: *Shh, Ihh* and *Dhh*, which likely favoured the increased complexity of vertebrates and their successful diversification.

In this study, we characterized the evolutionary genomics of the Hedgehog gene family in vertebrates, at the gene and protein levels. We used synteny analyses to better characterize and understand the genomic evolution of this family on vertebrate genomes, showing that this genes share syntenic features that may have evolved together at least since the origin of Deuterostomes. Detailed comparative genomic analyses suggested that these features may be present on avian genomes but probably located on microchromosomes, regions difficult to sequence and map. We also performed adaptive selection and functional divergence analyses in around 50 Hh gene and protein sequences, and we found that the vertebrate Hh paralogs are evolving under strong purifying constraints, mainly at the signaling domain. Different Hh paralogs, however, are under different purifying selective pressures, probably related with their different physiological roles. Also, functional divergence analysis showed that a small number of negatively selected residues located on the two Hh main domains significantly count for functional divergence between vertebrate Hh paralogs. A significant number of these residues are already annotated as mutation hotspots causing disease in human and are also related to important signaling events on the Hh signaling pathway. Interestingly, adaptive evolution analysis at the protein-level showed evidences of positive selection acting over the two main domains that comprise Hh proteins, mainly at the protein surface. This can be hypothesized to be responsible for different protein-protein interactions, explaining new sources for the distinct functional roles observed for each of the vertebrate members of this family, in addition to their distinct expression patterns.

## RESUMO

### Genómica Evolutiva e Evolução Adaptativa da Família de Genes Hedgehog em Vertebrados

A família de genes Hedgehog (*Hh*) é uma das mais importantes famílias de genes envolvidos em eventos homeostáticos e de desenvolvimento, codificando uma classe de proteínas secretadas altamente conservadas que atuam como moléculas sinalizadoras em todos os metazoários. Estas proteínas desempenham vários papeis na regulação do crescimento e da diferenciação celulares durante o desenvolvimento embrionário e pós-embrionário de vários animais, dos mais simples invertebrados até aos humanos. Mostrou-se já que a maior parte dos bilateria, com exceção de *C. elegans,* possui pelo menos um gene *Hh,* com as expansões do genoma em vertebrados a originar pelo menos três genes *Hh* com diferentes papeis funcionais: *Shh, Ihh* e *Dhh*, que possivelmente favoreceram o aumento da complexidade dos vertebrados e a sua diversificação.

Neste estudo, caracterizamos a genómica evolutiva e a evolução adaptativa da família de genes Hedgehog em vertebrados, ao nível do gene e da proteína. Usamos análises de sintenia para melhor caracterizar e compreender a evolução genómica desta família em genomas vertebrados, mostrando que estes genes partilham características sinténicas que possivelmente evoluíram em conjunto pelo menos desde a origem dos Deuterostomes. Análises de genómica comparativa detalhada sugeriram que estes estarão presentes em genomas de aves mas provavelmente localizadas em microcromossomas, regiões de difícil sequenciação e mapeamento. Também realizamos análises de seleção adaptativa e divergência funcional sobre cerca de 50 sequências de genes e proteínas Hh, e descobrimos que os parálogos vertebrados Hh encontram-se a evoluir sob fortes constrições purificantes, majoritariamente ao nível do domínio sinalizador. Diferentes parálogos Hh, contudo, encontram-se sob diferentes pressões seletivas purificantes, provavelmente devido aos seus diferentes papeis fisiológicos. Ainda, análises de divergência funcional mostraram que um pequeno número de resíduos selecionados negativamente, localizados nos dois principais domínios Hh, participam significativamente na divergência funcional entre parálogos vertebrados Hh. Um número significativo destes resíduos encontra-se já anotados como locais de mutação em diversas doenças humanas e estão também relacionados com importantes eventos sinalizadores da via de sinalização Hh. Curiosamente, análises de evolução adaptativa ao nível da proteína mostraram evidências de seleção positiva sobre os dois principais domínios que compõem as proteínas Hh, principalmente na superfície da proteína. Uma

hipótese é de estes resíduos serem responsáveis por diferentes interações proteína-proteína, explicando novas fontes para os distintos papeis funcionais observados para cada membro desta família em vertebrados, para além dos seus distintos padrões de expressão.

## INDEX

## LIST OF ABBREVIATIONS AND SYMBOLS

°C – Celsius degrees

µL – Microliter

AIC – Akaike Information Criterion

BEB – Bayes Empirical Bayes

BLAST – Basic Local Alignment Search Tool

BOC – Bother of CDO

BOI – Brother of Interference hedgehog

BSA – Bovine serum albumin

Ci – Cubitus interruptus

CKI – Cyclin-dependent kinase inhibitor protein

COS2 – Costal-2

Dhh – Desert hedgehog

Disp – Dispatched

$d_N$ – Non-synonymous substitution rates

DNA – Deoxyribonucleic acid

dNTP – Deoxyribonucleotide

dpc – Days postcoitum

Dpp – Decapentaplegic

$d_S$ – Synonymous substitution rates

EDTA – Ethylenediaminetetraacetic acid

En – Engrailed

ER – Endoplasmic reticulum

E-value – Expected value

FEL – Fixed Effects Likelihood

FSGD – Fish-specific genome duplication

FU – Fused

g – Gram

Gas1 – Growth-arrest-specific I

gDNA – Genomic DNA

GPI – Glycosylphosphatidylinositol

GSK3 – Glycogen synthase kynase-3

HCl – Hydrogen chloride

Hh – Hedgehog

HHAT – Hedgehog acyltransferase

HhC – Carboxy-terminal Hog domain

HhN – Amino-terminal Hedge domain

Hip – Hedgehog-interacting protein

ID – Identity

Ihh – Indian hedgehog

IHog – Interference hedgehog

Iro – Iroquois

$I_{SS}$ – Observed saturation index

$I_{SS.C}$ – Saturation index when assuming full saturation

kDa – Kilodalton

LMBR1 – Limb region 1 protein

LMBR1L – Limb region 1 protein-like

LRT – Likelihood ratio-test

MCMC – Markov chain Monte Carlo

Meg – Megalin

mg – Milligram

$MgCl_2$ – Magnesium chloride

min - minute

ML – Maximum Likelihood

mL – Milliliter

MLL – Mixed-lineage leukemia

mM – Millimolar

mya – Million years ago

NJ – Neighbor-Joining

Oxy – Oxysterols

$P(S_1|S_0)$ – Posterior probability for the functional divergence for each position in the alignment

PCR – Polymerase chain reaction

PI4P – Phosphatidylinositol-4-phosphate

PKA – Protein kinase A

PNS – Perypheral Nervous System

Ptc – Patched

*r* – Rate correlation between two duplicate genes

$R(S_1|S_0)$ – Posterior odd ratio

RHEB – Ras homolog enriched in the brain

RHEBL – Ras homolog enriched in the brain-like

rpm – Revolutions per minute

SDS – Sodium dodecyl sulfate

sec - second

Shh – Sonic hedgehog

Ski – Skinny hedgehog

SLAC – Single Likelihood Ancestor Counting

Smo – Smoothened

SRR – Sterol recognition region

SRY – Sex-determining region Y

SS – Signaling sequence

SUFU – Suppressor of fused

TGF-β – Transforming Growth Factor-β

Trx – Trithorax

TS – Theiler stage

w/v – Mass volume percentage

Wg – Wingless

WGD – Wide-genome duplications

ZPA – Zone of polarizing activity

$\theta$ – Coefficient of functional divergence

$\hat{\theta}$ – Estimated coefficient of functional divergence

$\hat{\theta}_I$ – Estimated coefficient of Type I functional divergence

$\hat{\theta}_{II}$ – Estimated coefficient of Type II functional divergence

$\omega$ – Non-synonymous to synonymous substitutions rate ratio

## LIST OF FIGURES

## LIST OF TABLES

# 1. INTRODUCTION

## 1.1. Prelude

In the book “Your Inner Fish: A Journey into the 3.5-Billion-Year History of the Human Body” [1], we can read “It turns out that being a paleontologist is a huge advantage in teaching human anatomy. Why? The best road maps to human bodies lie in the bodies of other animals. (...) The reason is that *the bodies of these creatures are often simpler versions of ours*”. This sentence reveals two points: first, the evolution of animals is a crucial topic in understanding human evolution and the human body; second, that we must share with other animals genes involved in development. Homologous genes involved in adaptation and development processes, like bone, brain, digits and other structures formation, are found in a wide range of animals, from fishes to mammals. Indeed, homologous developmental genes can even be found between humans and invertebrate species and the evolution of these genes can be influenced by several factors, such as mutation, recombination, gene duplication, and even gene transfer, which can provide advantageous features to the individual that are preserved through positive selection during the evolution of the lineage where it appeared, providing the ability of the species to adapt to different environments [2]. Deciphering signatures of adaptation in protein-coding genes can be challenging, but increasingly powerful genomics and proteomics tools may be the ultimate bridge between structural biology and molecular evolution [3].

## 1.2. The Hedgehog Gene Family

Cell signaling is an important event for the development and survival of multicellular organisms and evolution has worked with a limited number of signaling pathways and signaling molecules to generate the outstanding diversity and complexity of life [4]. Metazoans use many distinct signaling proteins for cell-to-cell communication encoded by a small number of gene families and, among the central group of developmental signaling pathways, the Hedgehog (Hh) signaling pathway is one of the most enigmatic [4, 5]. Since their isolation in the early 1990s, the members of the Hh family of intercellular signaling proteins have come to be recognized as key mediators of many fundamental processes in embryonic development and tissue homeostasis. Their activities are central to growth, patterning, and morphogenesis of many different regions within the body plans of vertebrates and invertebrates. In some contexts, Hh signals act as morphogens in the dose-dependent induction of distinct cell fates within a target field, in others as mitogens regulating cell proliferation or as inducing factors controlling the form of a developing organ [6].

*Hh* genes owe their discovery to the pioneering work of Nüsslein-Volhard and Wieschaus [6, 7]. In their screen for mutations that disrupt the *Drosophila* larval body plan, these authors identified in 1980 several that cause the duplication of denticles (spiky cuticular processes that decorate the anterior half of each body segment) and an accompanying loss of naked cuticle, characteristic of the posterior half of each segment. The ensuing appearance of a continuous lawn of denticles projecting from the larval cuticle suggested the spine of a hedgehog to the discoverers, hence the origin of the name of this family. Other loci identified by mutants with this phenotype included *armadillo, gooseberry,* and *wingless* (*wg*) and, on the basis of these mutant phenotypes, Nüsslein-Volhard and Wieschaus [7] proposed that these segment-polarity genes regulate pattern within each of the segments of the larval body [6].

Later, most bilaterians, with the exception of *C. elegans* [8], have been shown to possess at least one *Hh* gene and vertebrate *Hh* genes were first reported in 1993, following a cross-species (fish, chick and mouse) collaborative effort involving three groups [9-11] and additional reports of *Hh* homologs appeared the following year [12, 13]. Interestingly, unlike *Drosophila melanogaster*, which carries a single *Hh* gene, three *Hh* genes are usually found on vertebrate genomes: *Desert hedgehog* (*Dhh*), *Indian hedgehog* (*Ihh*) and *Sonic hedgehog* (*Shh*). While in *Drosophila* the only known *Hh* gene patterns many of the developing embryo stages [14], the vertebrate members of the *Hh* family each have different roles which depends from different expression patterns [15]: *Shh* has a central role in the development and patterning of the nervous and skeletal systems [6], *Ihh* mediates endochondral bone formation and vasculangiogenesis, and *Dhh* is essential for the formation of the peripheral neural system [16] and is involved in the differentiation of peritubular myoid cells and consequent formation of the testis cord [17].

### 1.2.1. Structural Features of The Hedgehog Proteins

Hh proteins are synthesized as approximately 45 kDa pro-proteins (about 400-460 amino acids long) and comprise several highly conserved motifs and domains (Fig. 1A): a signal peptide for protein export, a secreted N-terminal “Hedge” domain (HhN) that acts as a signaling molecule, and an autocatalytical C-terminal “Hog” domain (HhC) that is involved on the processing of the mature signaling peptide [18]. The fact that purified Hh proteins from a bacterial source can undergo cleavage *in vitro* first indicated that this is an autoproteolytic process [19], and the concentration-independent kinetics of the reaction further suggested that it occurs by an intramolecular mechanism [20] (Fig. 1B).

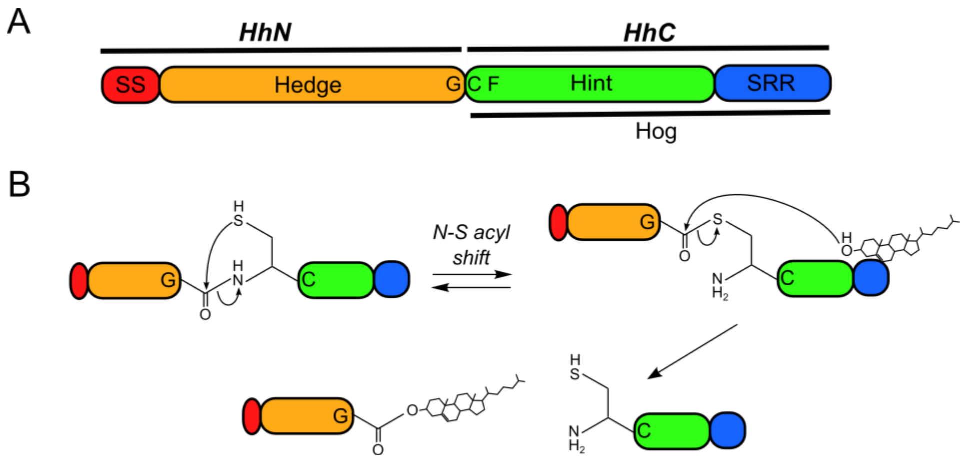


**Figure 1. Structural features of Hh proteins.** (a) The hedgehog proteins are composed by two main domains: the Hedge (N-terminal) and Hog (C-terminal) domains. The Hedge domain forms the HhN portion of the Hh proteins (together with the signaling sequence, SS) and is separated from the Hog domain by a GCF motif that forms the boundary between the two main parts of the Hh proteins. The sterol-recognition region (SRR) forms the C-terminal region of the Hog domain [21]. (b) The intramolecular autoprocessing of the Hh proteins occurs on a two-step reaction. First, the thiol group of the cysteine at the cleavage site makes a nucleophilic attack on the carbonyl group of the preceding residue, glycine, resulting in a thioester intermediate. Second, the SRR region recognizes a cholesterol moiety and its 3-β-hydroxyl group attacks this thioester to form an ester-linked adduct to the HhN and free HhC [21]. Figure adapted from [18].

Based on the analysis of different forms of mutant Hh proteins, HhC was found to be the catalytic domain, whereas most of HhN is dispensable for the reaction [19, 20]. On the other hand, all of the signaling activity of the Hh proteins is performed by the HhN fragment and the only known function of the Hog domain is to promote the autocleavage reaction. It was noticed that the Hog domain has sequence similarity with self-splicing Inteins [22] (protein sequences that autocatalytically splice themselves out of a longer protein precursor) and the shared region was called “Hint” [23]. Therefore, HhC bind cholesterol in the sterol-recognition region (SRR) [21] and the catalytic activity of the Hint module cleaves Hh over a highly conserved GCF (Glycine-Cysteine-Phenylalanine) motif that forms the boundaries between the two main domains in a two-step reaction (Fig. 1) [21].

Until today, the structure of HhC was only solved for the *Drosophila melanogaster* Hh protein, by Hall *et al.* in 1997 [23]. The structure is globular, composed of β-strands and starts with the cysteine residue critical for auto-processing (Fig. 2). However, the overall structure found only represents the Hint region and do not comprise the SRR region from the Hog domain [23]. It folds to form a unique hydrophobic core with the catalytic center being located on a deep groove within the interior of the peptide. This active site is composed by the highly conserved cysteine residue as well by two absolutely conserved histidine and threonine residues, crucial for thioester formation, and by a third residue that can either be an aspartic acid or an histidine residue and is essential for sterol transfer [23] (Fig. 2a).

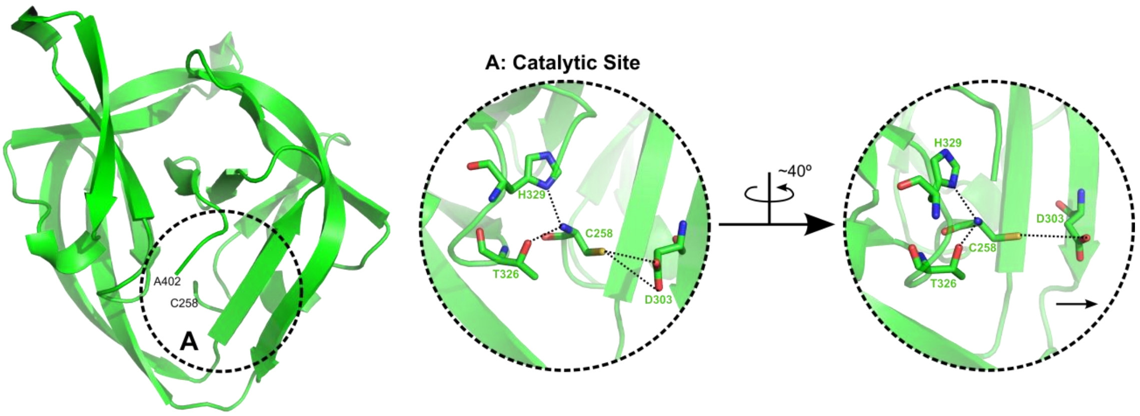


**Figure 2. Tridimensional structure of HhC peptides. T**he tridimensional structure of the *Drosophila melanogaster* HhC peptide (PDB: 1AT0) is represented in green cartoon. The peptide is incomplete at C-termini, missing the sterol recognition region (therefore, only the Hint region is represented). (a) The catalytic site is composed by residues Cys258, Asp303, Thr326 and His 329 (numbered according to the *Drosophila melanogaster* Hh sequence) buried on the surface of the peptide.

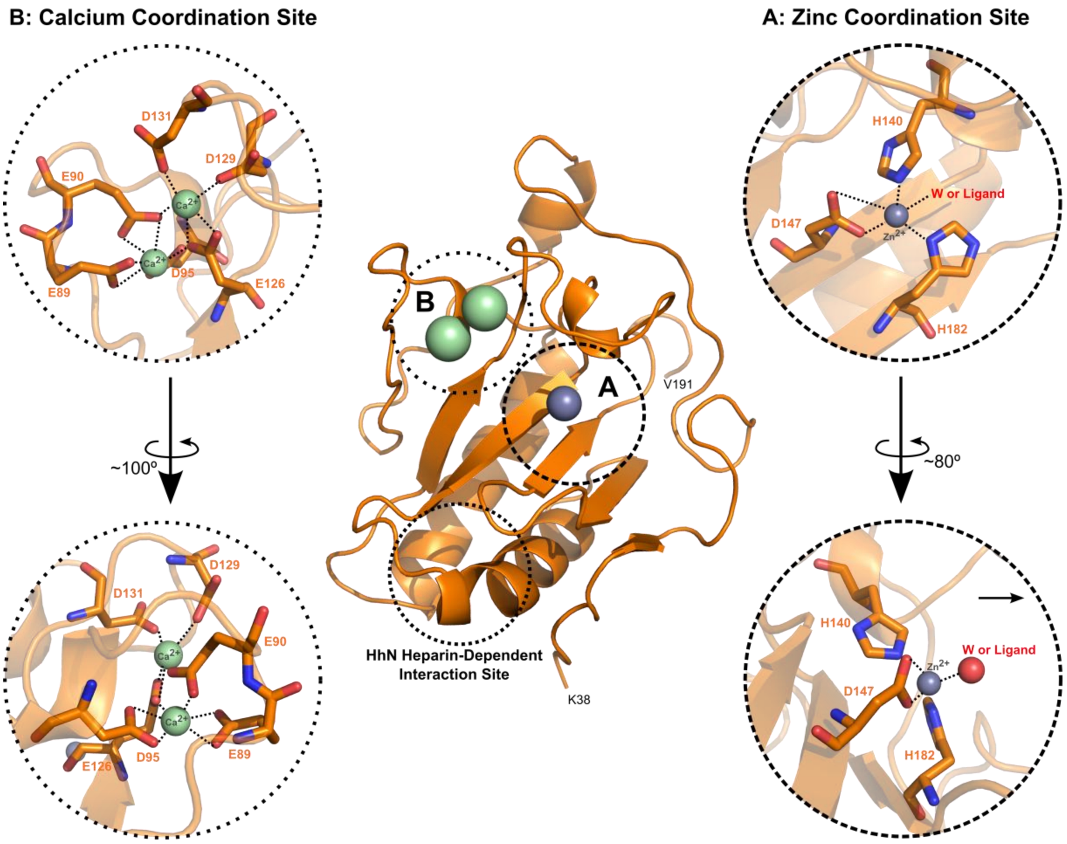


**Figure 3. Tridimensional structure of HhN peptides.** In the centre is represented, as an example, the tridimensional structure of the human ShhN peptide (PDB: 3HO5), in orange cartoon. In grey sphere is represented the zinc atom and in palegreen the two calcium atoms. The peptide is incomplete both at its N- and C-termini. Circles mark the position of the three main interaction regions known for HhN peptide: the highly conserved (a) mononuclear zinc coordination site and (b) binuclear calcium coordination site, only present on vertebrate HhN peptides, and the vertebrate equivalent position of the heparin-dependent binding site only present on the invertebrate members of the Hh family. (a) The tetrahedrally zinc coordination site is located at the base of a large cleft formed by several β-strands surrounded by loops, and is composed by residues His140, Asp147 and His182 (numbered according to the human Shh sequence) and by a fourth ligand, that can be either a water moiety (W) or the lateral chain of a residue on a binding protein [24]. A horizontal arrow marks the exit of the cleft. (b) The binuclear calcium coordination site is located next to the zinc coordination site, and is composed by six highly acidic amino acid residues: Glu89, Glu90, Asp95, Glu126, Asp129 and Asp131 (numbered according to the human Shh sequence).

On the other hand, the crystal structure of an HhN peptide was first determined in 1995 by Hall *et al.* for the murine Shh protein and it revealed a relatively globular structure with two antiparallel α-helixes and several β-strands wrapping one face of the helixes (Fig. 3) [25]. Recently, the same structural features where described for additional human, murine and *Drosophila* Hh proteins, highlighting a highly conserved structure among HhN paralogs [24]. Interestingly, the HhN peptide structure revealed two conserved ion coordination sites found only on the vertebrate peptides [24]: a zinc coordination site and a calcium coordination site (Fig. 3).

The HhN zinc coordination site shares a high homology with the active site of zinc hydrolases, with the zinc ion being coordinated by two histidines and an aspartate at the base of a large cleft formed by the β-strands, and by a water molecule with a potential role on catalysis (Fig. 3A) [25]. This exciting finding suggested the possible contribution of an intrinsic hydrolytic activity on the signaling activity of HhN peptides but mutagenesis studies discarded this possibility [26, 27]. In fact, the zinc coordination site plays an important structural and functional role on the signaling activity of the HhN peptide, being responsible for its stability [26, 27], but acts also as a recognition site for Hh-protein receptors with the substitution of the water moiety by a residue from the receptor protein on the moment of biding [28, 29]. Equally, the calcium coordination site is crucial for the interaction of HhN peptides with the majority of its receptor proteins. It is located apart from the zinc coordination site and is composed by two calcium ions coordinated by six acidic amino acids and by none from the interacting pattern (Fig. 3) [24, 28]. Interestingly, this binuclear coordination site is not found on the *Drosophila* HhN peptide, who requires heparin as a binding-cofactor for the interacting protein and promotes the interaction on a different peptide region (Fig. 3) [24, 30].

In addition, the HhN fragments also undergo palmitoylation at their first N-termini residue, a modification that is promoted by an acyl transferase encoded by the *skinny hedgehog* (*ski*) gene in *Drosophila melanogaster* [31] and in vertebrates by its orthologue *hedgehog acyltransferase* (*HHAT*) [32] (Fig. 4). This dual lipid modification of the Hh signaling protein has important effects on its properties, both enhancing its membrane association [33] and potentiating its secretion and range of activity. The modification is crucial for the extracellular movement of the signal following secretion [34, 35], as it promotes the formation of freely diffusible multimeric complexes [36, 37] and its incorporation into lipoprotein particles that seem to mediate its long-range transport [38].

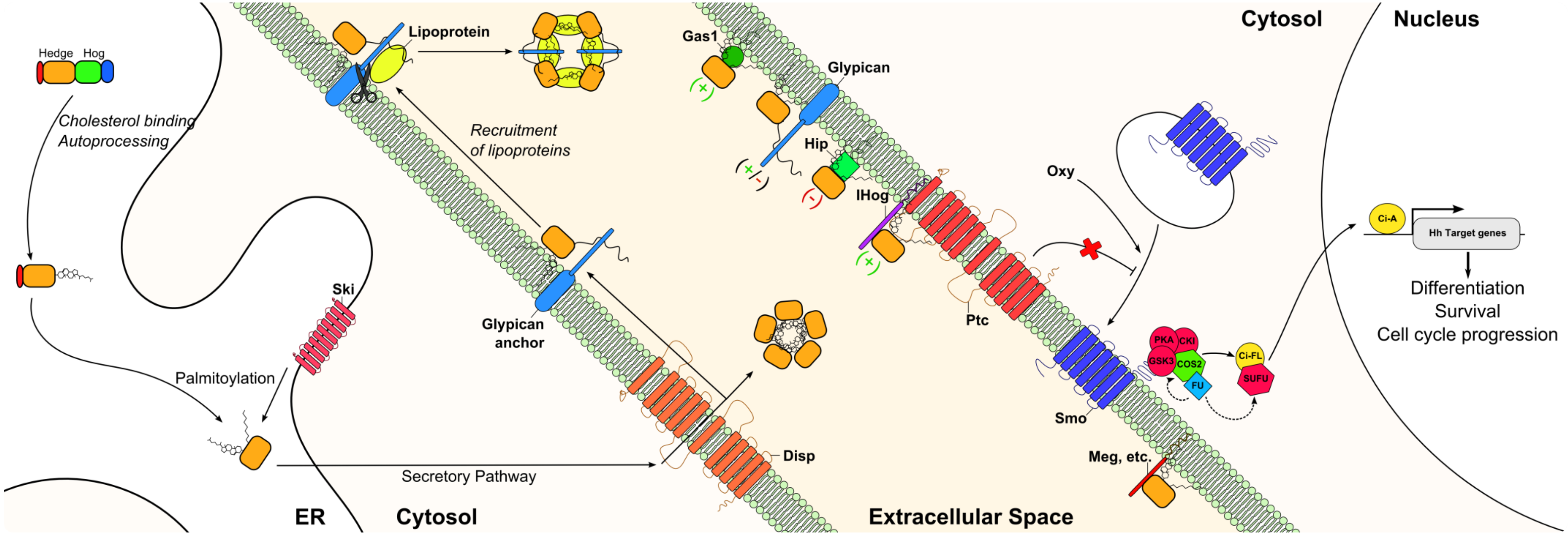


**Figure 4. A simplified Hh signaling pathway, constructed from combined *Drosophila* and mammalian data.** Following its translation, full-length Hedgehog undergoes autoproteolysis in the endoplasmic reticulum (ER), resulting in its covalent coupling to cholesterol, and is further palmitoylated at its N-terminal by the *Drosophila* transmembrane acyl transferase Skinny Hedgehog (Ski) and by its vertebrate orthologue Hedgehog acyltransferase (HHAT). Release of the modified HhN peptide by the secretory pathway requires the activity of the multipass transmembrane protein Dispatched (Disp), which probably transports the protein across the plasma membrane. Once on the outer surface of the cell, modified HhN peptides can form multimers or associate with lipoproteins. The association of modified HhN peptides with lipoproteins requires the association of HhN with heparin sulphate moieties of glypicans, which recruit the apo-lipoprotein lipophorin that, together with HhN, becomes assembled into lipoprotein particles. Release of these particles might be mediated by the phospholipase C-like Notum, which cleaves the GPI anchors from the glypicans (indicated by scissors). A number of molecules can interact with the modified HhN peptides and propagate or modulate [(+): positive regulation; (-): negative regulation] its trafficking: glypicans, the Hedgehog interacting protein (Hip), the Growth-arrest-specific I protein (Gas1), Megalin (Meg), etc. On the other hand, Interference Hedgehog (IHog) and its homologs (BOI, COD and BOC) act as co-receptors for modified HhN peptides, presenting the signal to its receptor, Patched (Ptc). Modified HhN peptides repressed the function of Ptc, a 12-transmembrane protein related to Disp, resulting in the internalization of the receptor-ligand complex and further destruction (not shown). Ptc inhibits the 7-pass membrane receptor Smoothened (Smo) and when the inhibitory function of Ptc is released by HhN, Smo can translocate to the plasma membrane or to the primary cilium, and active Smo is phorphorylated by Protein kinase A (PKA), Glycogen synthase kynase-3 (GSK3) and Cyclin-dependent kinase inhibitor protein (CKI). Oxysterols (Oxy) can also indirectly activate Smo. Smo phosphorylation causes a conformational change in the Smo C-terminal domain, enhancing its interaction with Costal-2 (COS2), who phosphorylates Fused (FU, dashed arrow) and causes Ci to be released from the Hedgehog Signaling Complex. Fu-dependent phosphorylation of Suppressor of fused (SUFU; dashed arrows) promotes its dissociation from Ci-FL, allowing Ci-FL to translocate to the nucleus, where it undergoes further modification to its activated form (Ci-A) ans thus promotes the transcriptional activation of Hh target genes, involved in differentiation, survival and cell cycle progression. Figure adapted from [5], [18] and [24].

### 1.2.2. The Hedgehog Signaling Pathway

Originally defined through genetic analysis in *Drosophila melanogaster* [39], the components of the Hh signaling pathway have subsequently been characterized in several vertebrate species (mouse, zebrafish and human) and have also been identified in species from a wide range of phyla. These studies have revealed a high level of conservation of the 'core' components of the signal transduction pathway that is likely to extend across the eumetazoa [5] and, although recent studies have suggested a role for Hh in modulating the cytoskeleton via SRC family kinases [40], the most widespread and best-studied response of cells to Hh signaling is the upregulation of target genes, mainly involved in differentiation, survival and cell cycle progression (Fig. 4) [5].

Figure 4 shows a summary of the canonic Hh pathway built from combined *Drosophila* and mammalian data. Following translation, the Hh pro-proteins undergo autoproteolysis in the endoplasmic reticulum (ER) [41], resulting in its covalent coupling to cholesterol (Fig. 1), and HhN is further modified through N-terminal palmitoylation, promoted in *Drosophila* by the transmembrane acyl transferase Skinny hedgehog (Ski) [31] and in vertebrates by its homolog Hedgehog acyltransferase (HHAT) [32]. Release of this doubly lipid-conjugated form of HhN requires the activity of the 12-pass transmembrane protein Dispatched (Disp), which probably transports the protein across the plasma membrane [42]. Once on the outer surface of the cell, the modified HhN peptides can follow two fates: they can form freely diffusible multimeric complexes [36, 37] or be incorporated into lipoprotein particles that seem to mediate their long-range transport [38], which depends if the modified HhN peptide is basally or apically released from the producing cell [43].

The assembly of the modified HhN peptides into lipoproteins is promoted by interaction with lipophorin, an apo-lipoprotein that is recruited to HhN secreting cells by its interaction with the heparin sulphate moieties of the glypicans Dally and Dally-like [44]. These proteoglycans, which can also interact with HhN [5], localize to the apical surface of epithelial cells via GPI anchors (a glycolipid, glycosylphosphatidylinositol, linked to the C-terminal amino acid of proteins anchoring them to the outer leaflet of the plasma membrane), the cleavage of which by the phospholipase C-like Notum seems to be required for effective long-range HhN signaling [43]. Therefore, glypicans promote the assembly of modified HhN-lipophorin particles at the plasma membrane and the cleavage of their GPI anchor facilitates the release and dispersal of modified HhN from producing cells [5] (Fig. 4).

Over the receiving cell, the modified HhN peptide can interact with multiple cell surface proteins, which can be implicated in receiving or modulating responses to Hh signals (Fig.

4). The key function of the modified HhN peptide as an extracellular signal is to inhibit the activity of the receptor Patched (Ptc) at the primary cilium [45, 46], a 12-pass transmembrane protein related to Disp (Fig. 4) [47]. Ptc specifically binds the modified HhN and is a 1500 amino acid glycoprotein with 12 membrane-spanning domains [48, 49] with two large extracellular loops that are required for Hedgehog binding [50]. This interaction is promoted in *Drosophila* by the transmembrane proteins Interference Hedgehog (IHog) and Brother of Interference Hedgehog (BOI) [51], and in vertebrates by their orthologues CDO and Brother of CDO (BOC) [52].

However, Hh signaling can be further regulated or modulated by several other cell surface components (Fig. 4), mainly: vertebrate and invertebrate glypicans, which can have a positive or negative effect and can affect either responsiveness to HhN or the tissue distribution of HhN [53-56]; and the vertebrate cell surface proteins Growth-arrest-specific I (Gas1) and Hedgehog-interacting protein (Hip), positive and negative modulators of the Hh signaling pathway, respectively [57-59]. The interaction between HhN with its co-receptors IHog/CDO/BOC and Hip was already characterized and it was shown that vertebrate HhN peptides bind CDO/BOC by the calcium coordination site and Hip by the zinc coordination site, while the *Drosophila* HhN peptide bind IHog with the aid of heparin over the Heparin-dependent interaction site (Fig. 3). Inversely, none zinc coordination site is found on the *Drosophila melanogaster* HhN peptide and any Hip identified homolog is present in this species [24]. Several other proteins, including Megalin [60], Vitronectin [61], Perlecan [62], Scube2 [63] and Shifted [64, 65], have been reported to bind HhN peptides, but their interactions with HhN have been less well characterized [24].

HhN interaction with its modulators and co-receptors does not activate any known signaling pathway [24] but the transmembrane domains of Ptc shows an intriguing homology to the “cholesterol sensing” motifs of transporters involved in cholesterol homeostasis and this motif may have a broader role in intracellular trafficking of receptors and their ligands [66]. In fact, HhN binding causes endocytosis of the Hedge-Ptc complex and decrease in the total amount of Ptc protein in the cell, likely due to lysossomal degradation [67, 68]. In the absence of HhN binding, Ptc represses a signaling pathway that acts through Smoothened (Smo) [67, 69], a 115 kDa seven-pass protein with structural similarity to serpentine G-protein coupled receptors (Fig. 4) [70, 71]. Smo is negatively regulated by pro-vitamin $D_3$ and it is positively, but indirectly, regulated by oxysterols (oxygenated derivatives of cholesterol) [72, 73]. Thus, Ptc may secret pro-vitamin $D_3$ or related compounds to inhibit Smo [74], which is supported by the discovery that the steroidal alkaloid cyclopamine binds and inhibits Smo activity [75]. In addition, recent studies showed that Ptc is responsible for cholesterol efflux, which may modulate the activation of Smo [76], and also that the phospholipid phosphatidylinositol-4-phosphate (PI4P) is implicated in the regulatory relationship between Ptc and Smo, suggesting that Smo is activated by an increase in intracellular PI4P levels and that Ptc modulates these levels by inhibiting the activity of the kinase that is responsible for PI4P synthesis [77]. Conversely, when HhN binds to Ptc, the complex is internalized while Smo translocates to the cell membrane and oxysterols can indirectly activate Smo [73].

Activated Smo is phosphorylated and signals via a cascade of microtubule-associated proteins to the nucleus, where the transcription factor Cubitus interruptus (Ci) in *Drosophila melanogaster* or its mammalian counterparts, the Gli transcription factors, activate or repress target genes (Fig. 4). Only a few such targets have been described in detail, but recent genome-wide analyses suggest that there are several hundred [5]. Some examples are *Ptc*, *decapentaplegic* (*dpp*), *engrailed* (*en*), *iroquois* (*iro*), *wingless* (*wg*), *cyclins D* and *E*, *Myc, Gli1* and *Hip,* which comprise regulators of the Hh signaling pathway, as well as cell cycle, differentiation and survival controllers [78, 79] and links the Hedgehog signaling pathway to several congenital and hereditary diseases (e.g., holoprosencephaly and cyclopia [80, 81], acrocapitofemoral dysplasia [82] and gonadal dysgenesis with minifascicular neuropathy [83]), but also to tumerogenesis (e.g., basal cell carcinoma, medulloblastoma and breast and liver cancers [84, 85]).

### 1.2.3. Members of The Hedgehog Family

In *Drosophila melanogaster*, the Hh protein is a central patterning signal in the wing [86, 87], leg [88] and eye discs [89, 90], as well as in regulating several other processes, including germ-cell migration [91], and development of the optic lamina [92, 93], gonad [94, 95], abdomen [96], gut [97] and tracheal system [98]. In contrast, the vertebrate members of the *Hh* family each have different roles which depends from different expression patterns [15] (Fig. 5).

In mammals, *Desert hedgehog* (*Dhh*) expression is largely restricted to gonads, including sertoli cells of testis and granulosa cells of ovaries (Fig. 5). In testis, Dhh is the first identified morphogenetic regulator downstream of the testis determining switch *sex-determining region Y* (*SRY*) gene, facilitating testis cord formation by acting upon peritubular myoid cells and, at the same time, inducing fetal Leydig cell differentiation [99]. On the other hand, it works in synergy with *Indian hedgehog* (*Ihh*) to regulate theca cells and ovary development [17, 100, 101]. The Hh signalling pathway is inactive in the fetal ovary based on the absence of *Ptc* and *Gli1* expression [102, 103], preventing the ectopic appearance of fetal Leydig cells [104], but Hedgehog ligands are detected after birth [105]. *Dhh* is also expressed at a reduced extent in Schwann cells, in peripheral nerves, during the maturation step of mesenchymal cells in the Perypheral Nervous System (PNS) development, being responsible for perineurium development. In fact, in the absence of Dhh signalling, the perineurium is disorganized and is permeable to macromolecules and inflammatory cells [83, 106, 107].

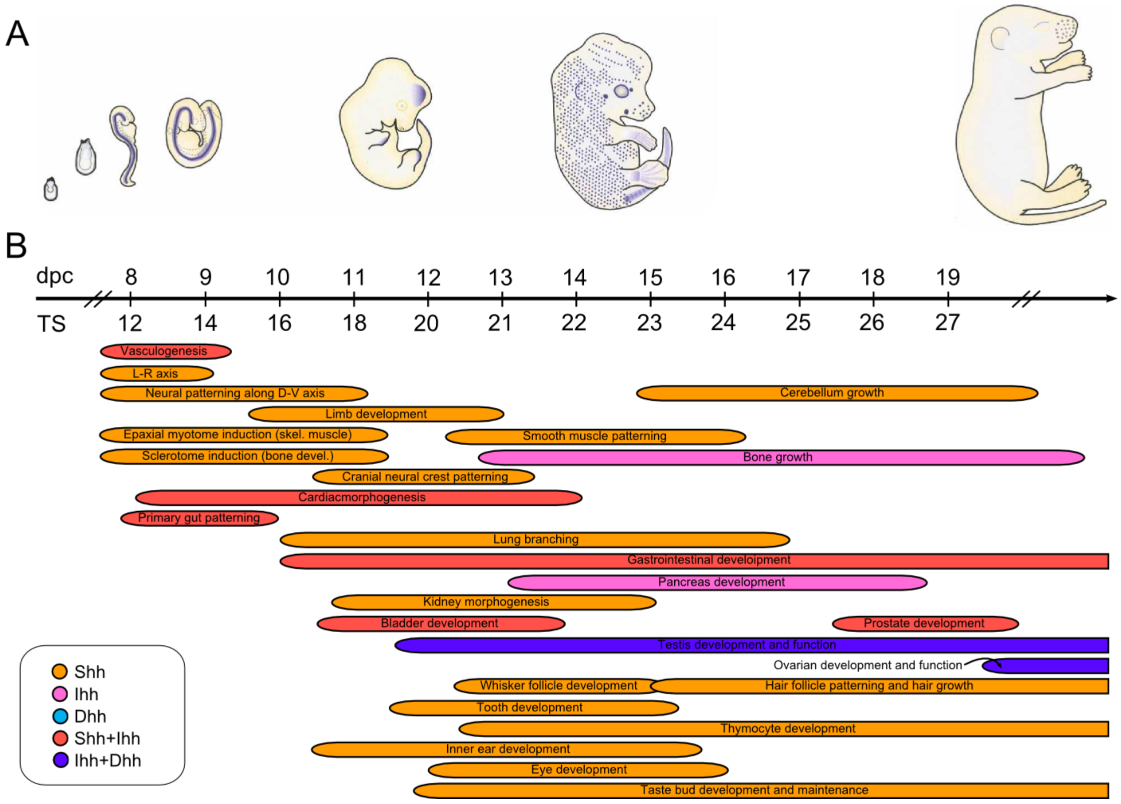


**Figure 5. Mouse *Hh* and *Ptc* genes expression pattern.** (A) The embryo cartoon shows aspects of expression of the Hh target gene *patched* (*Ptc*) (blue) during mouse embryonic development. (B) Bars show approximate embryonic stages when Sonic hedgehog (Shh), Indian hedgehog (Ihh) and Desert hedgehog (Dhh) (color code in bottom left) control developmental processes in the indicated tissues or cell types. The approximate embryonic stage by days postcoitum (dpc), and Theiler stage (TS), is presented. *Shh* is the most broadly expressed Hh signaling molecule, being expressed in all major developmental stages and tissues and cells types. Ihh is mainly expressed on bone tissues while Dhh is confined to gonads, mainly in combination with Ihh. Figure adapted from [15].

Mutations on the mammal *Dhh* gene were related to demyalinating neuropathies and it was also observed that some of those mutations can led to abnormal sex differentiation. In particular, this gene has been identified as critical in the development of Gonadal Dysgenesis with Minifascicular Neuropathy [83, 108]. Demyelinating neuropathies are a diverse and complex group of disorders associated with primary alterations of myelin sheath. Therefore, lack of *Dhh* expression leads to abnormal PNS development, with disorganized and permeable perineurium [108] and disrupts the differentiation of male gonads and spermatogenesis, a pathology known as Gonadal Dysgenesis. This leads to peripheral nerve abnormalities, such as perineural cells, which form minifascicles around small groups of nerve fibers [83]. However, regarding the activity of Dhh on ovary development, there is no evidence of pathology associated with Dhh signalling. In fact, loss of Dhh signalling has not been reported to influence folliculogenesis [101].

*Indian hedgehog* (*Ihh*) is also specifically expressed in a limited number of tissues, including primitive endoderm [109], prehypertrophic chondrocytes in the growth plates of bones [110, 111] and osteoblasts under the regulation of Transforming Growth Factor-β (TGF-β) [112] (Fig 5). Approximately 50% of embryos lacking Ihh signalling die during early embryogenesis due to poor development of yolk-sac vasculature and surviving embryos display cortical bone defects as well as aberrant chondrocytes development in the long bones [111, 113]. In fact, *Ihh* mutations are implicated in several human diseases, mainly related with skeletal abnormalities such as Acrocapitofemoral Dysplasia [114]. Skeletal dysplasias are a clinically diverse and genetically heterogeneous group of connective tissue disorders affecting skeletal morphogenesis and development. An example of Acrocapitofemoral Dysplasia's phenotype is characterized by short stature of variable degree with short limbs and brachydactyly, relatively large head, narrow thorax with pectus deformities and normal intelligence [82, 114].

Inversely, *Sonic hedgehog* (*Shh*) is the most broadly expressed mammalian Hh signalling molecule, probably retaining most of the ancestral *Hh* functions (Fig. 5). During early vertebrate embryogenesis, *Shh* expressed in midline tissues such as the node, notochord and floor plate, controls patterning of the left and dorso-ventral axes of the embryo [115-118] and *Shh* expressed in the zone of polarizing activity (ZPA) of the limb bud is also critically involved in patterning the distal elements of the limbs [11, 12, 119, 120]. Later in development, during organogenesis, *Shh* is also expressed, affecting the development of most epithelial tissues [15]. Therefore, deletion of *Shh* leads to cyclopia, and defects in ventral neural tube, somite, and foregut patterning and later defects include, but are not limited to, several distal limb malformation, absence of vertebrae and most of the ribs and failure of lung branching [121-124]. In fact, *Shh* had been identified as the first Holoprosencephaly-causing gene both in human and mouse [121, 125], the most common developmental defect of the forebrain and the face. Holoprosencephaly phenotypes are variable, ranging from a single cerebral ventricle and cyclopia to clinically unaffected patients [126, 127].

### 1.2.4. Evolution of The Hedgehog Gene Family

New classes of Hint-containing proteins with various types of activity have been discovered in bacteria and eukaryotes [128-131]. Genes containing the Intein are present in all three kingdoms of life but *Hog* genes are only known presently in eukaryotes [129]. *Hog* genes were found initially solely in metazoans, but recently, they have been found also in many different branches of protists, which indicates that they must be of ancient origin and have emerged early in eukaryotes evolution [129, 131-133]. Interestingly, many of these Hog proteins have secreted domains upstream of the Hog domain, which in most cases shows conservation only with related *Hog* genes within the same phylum [18, 129]. However, the Hedge domain seems to be of more recent origin. It has been found in Cnidaria in a large extracellular protein called Hedgling, who lacks a Hog domain, and also in sponges in the absence of a Hog domain [18, 129, 134]. Even though, at present no *Hh* gene has been found in sponges but they are present in cnidarian [18]. In this way, the Hedge domains could have evolved from a secreted amino-terminal domain already associated with a Hog domain (and proteins such as Hedgeling could have evolved from Hh from the split of the Hog domain), or it could have evolved from an extracellular protein that have then fused with a Hog protein, giving rise to *Hh* [18, 129].

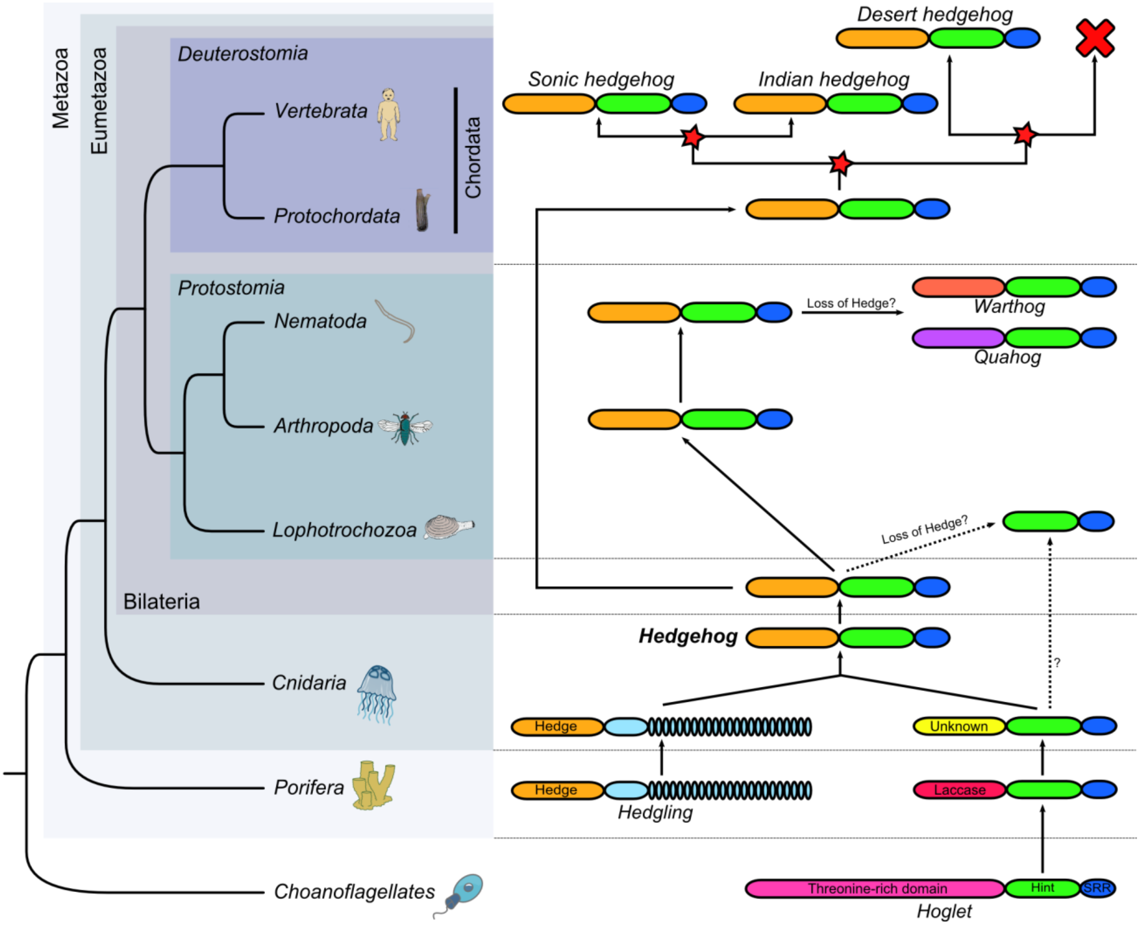


**Figure 6. A model for *Hh* evolution.** The presence of *Hint*-containing genes outside the Metazoa, such as the *Hoglet* gene identified in a freshwater choanoflagellate, suggests that evolutionary precursors of *Hh* signaling existed prior to the metazoan radiation and the lack of true *Hh* in the sponge genome suggests that the origins of the metazoan *Hh* ligand may have occurred following the divergence of sponges with Eumetazoa. The identification of both *Hint/Hog* genes and *Hh* genes in cnidarians argues that the evolution of an *Hh* gene in the cnidarians-bilaterian ancestor occurred by a domain-capturing event of an N-terminal signaling (Hedge) domain and a Hint/Hog domain-containing gene (Hog). *Hh* and *Hh*-related genes are found in Bilateria, however *Drosophila* and vertebrates lack *Hh*-related genes, nematodes carry both *Hh* and *Hh*-related genes and some Lophotrochozoans possess *Hint*-only genes. Therefore, the evolution of the *Hh* gene on the Protostome lineage may be diversified. The Lophotrochozoan *Hint*-only genes could have evolved parallel to the *Hh* genes from an ancestral *Hint*-containing gene or from the Bilaterian *Hh* gene by Hedge domain loss. Additionally, phylogenetic analysis suggests that nematode *Hh*-related genes are derived from an ancestral nematode true *Hh* gene, with *C. elegans* having loss its *Hh* gene. On the Deuterostome lineage, two wide-genome duplications (WGD) early on the evolution of chordates seems to be the origin of the three vertebrate *Hh* paralogs. Figure adapted from [135].

A model of *Hedgehog* gene evolution is represented in figure 6. In *Drosophila* and vertebrates, only Hh genes are present, but both *Hh* and *Hh*-related genes are found in Cnidaria and nematodes [129, 135]. Probably this occurs because these genes could have evolved in parallel: at least one *Hh* and one *Hh*-related gene existed at the origin of Eumetazoa, giving rise to the *Hh* and *Hh*-related genes in Cnidaria and nematodes and in *Drosophila* and vertebrates the *Hh*-related genes were lost [18]. Other alternative based on phylogenetic analysis [135] would be that the *Hh*-related genes in Cnidaria and nematodes were all derived independently from an *Hh* gene in each phylum, or that *Hh* related genes evolved from an *Hh* gene only in one or two phyla [18, 129]. Apart from these possibilities, two wide-genome duplications (WGD) before the emergence of chordates seems to be the origin of the *Hh* vertebrate paralogous genes: a first duplication 662 million years ago (mya) of an ancestral *Hh* gene gave rise to the *Shh*/*Ihh* and *Dhh* ancestor genes and an additional duplication event 563 mya generated *Shh*, *Ihh*, *Dhh* and a fourth gene quickly lost [6, 136, 137].

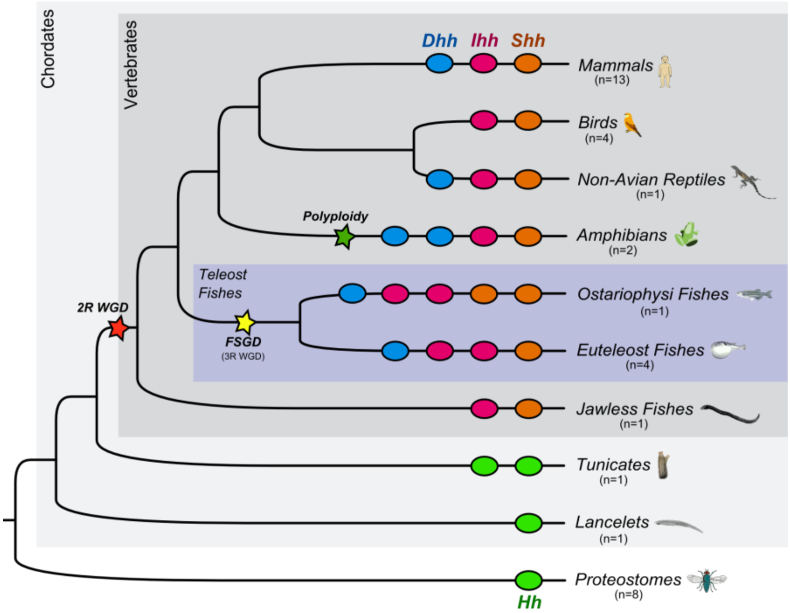


**Figure 7. Pattern of *Hh* gene presence on currently available eumetazoan genomes, according to GenBank [138] and Ensembl [139] databases.** Typically, invertebrate species possess only one *Hh* gene while vertebrate species carries at least on representative of each *Hh* vertebrate paralogs. Two rounds of wide-genome duplication (2R WGD) originated the three vertebrate *Hh* paralogs and a third fish-specific genome duplication (FSGD) and a polyploidy event on some amphibian lineages led to additional *Hh* duplicates on teleost and *Xenopus* genomes. A lineage-specific duplication is also found on the genome of the tunicate *Ciona intestinalis*. However, it is not possible to find any annotation of a *Dhh* gene on the currently available avian genome assemblies. The number of species searched and used to build the figure is described (n).

Typically, invertebrate species possess only one *Hh* gene while vertebrate species carry at least one representative of each *Hh* vertebrate paralogs (Fig. 7). Mammals have one *Hh* gene in each of the three subgroups, but due to the fish-specific genome duplication about 350 mya (FSGD) four or five *Hh* genes, *Dhh, Ihha, Ihhb, Shha* and *Shhb*, can be found in different teleost species [140-143]. A duplicated *Dhh* gene is also present on the genome of *Xenopus laevis* but not on the genome of *Xenopus tropicalis*, since the *Xenopus* species are allopolyploid, with the exception of the *tropicalis* one [137, 144]. Interestingly, southern blot analysis of genomic DNA showed that avian genomes also carry one example of *Hh* gene from each group, but none example of *Dhh*-coding sequence is found annotated on the currently available avian genome assemblies [11, 138, 139]. In addition, two *Hh* paralogs are found on the genome of the cyclostomes *Lampreta fluviatilis* and *Petromyzon marinus,* which clusters with the *Shh/Ihh* vertebrate group, suggesting that cyclostomes once had a *Dhh* gene but lost it [145] and that the *Shh, Ihh* and *Dhh* members of the *Hh* are more ancient than agnathans. However, the urochordate *Ciona intestinalis* has two *Hh* genes, *CiHh1* and *CiHh2*, that cluster with the invertebrate *Hh* group and are likely to result from a lineage-specific duplication [146].

### 1.3. Sequence Evolution After Gene Duplication

According to Ohno's classic view, the evolution of genes and genomes is typically conservative in the absence of gene duplication [147]. Tandem, regional or whole-genome duplication events produce pairs of initially similar genes, which can ultimately become scattered throughout a dynamically rearranging genome [148]. All vertebrate species, despite their generally diploid state, carry large numbers of duplicated genes, a result of two rounds of WGD that occurred early at the origin of the vertebrate lineage (the 2R hypothesis) [149-151], and represents the leading force for *Hh* gene family diversification in vertebrates [137].

Duplication of genetic material is generally accepted as an important precursor of functional divergence [147, 152-154]. No matter how duplicated genes arise, if they are duplicated in their entirety (including regulatory elements) then they can show inter-gene redundancy and have different fates [155-158] (Fig. 8). The most likely fate for these duplicated gene pairs is that one of them will degenerate to a pseudogene or be lost from the genome due to the vagaries of chromosomal remodeling, locus deletion or point mutation, a process known as non-functionalization [159]. Gene loss through these processes is permissible because only one of the duplicates is required to maintain the function provided by the single, ancestral gene, leaving one gene under purifying selection and the other gene free to accumulate evolutionary neutral or nearly neutral loss-of-function mutations in the coding region [160]. A less frequently expected outcome is that a population acquires a new, advantageous allele as the result of alterations in coding or regulatory sequences, exposing the formerly redundant gene to new and distinct selective constraints. Mutations that lead to such neo-functionalization are assumed to be extremely rare, so the classical model predicts that few duplicates should be retained in the genome over the long term [160].

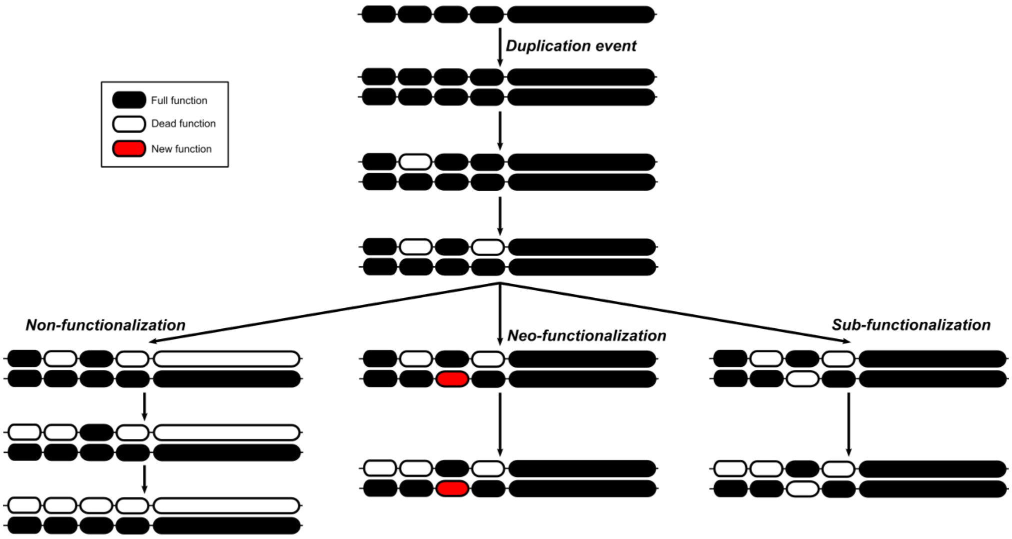


**Figure 8. Three potential fates of duplicated gene pairs with multiple regulatory regions.** The boxes denote regulatory elements with unique functions, and the large boxes denote transcribed regions. Solid boxes denote intact regions of a gene, while open boxes denote null mutations and red boxes denote the evolution of a new function. In the first two steps, one of the copies acquires null mutations in each of two regulatory regions. On the left, the next fixed mutation results in the absence of a functional protein product from the upper copy. Because this gene is now a non-functional pseudogene, the remaining regulatory regions associated with this copy eventually accumulate degenerative mutations. On the right, the lower copy acquires null mutation in a regulatory region that is intact in the upper copy. Because both copies are now essential for complete gene expression, this third mutational event permanently preserves both of the genes from future non-functionalization. The fourth regulatory region, however, may still eventually acquire a null mutation in one copy or other. In the center, a regulatory region acquires a new function that preserves that copy. If the beneficial mutation occurs at the expense of an otherwise essential function, then the duplicate copy is preserved because it retains the original function. Figure adapted from [161].

Studies indicated that duplication often results in continuing partial genetic redundancy. Expression analyses suggest that extant gene pairs might have, in many cases, partitioned the multiple functions of single ancestral genes between the descendant duplicates and population-level models and experimental evidence point out that gene multifunctionality might act to potentiate the preservation of duplicated genes [160]. A broadly applicable sub-functionalization model was proposed by Force and colleagues [161, 162] to explain the prevalence of duplicate genes that are retained in the genome. This model proposes that, after duplication, the two gene copies are required to produce the full complement of functions of the single ancestral gene (Fig. 8). A likely way for sub-functionalization to occur is through complementary changes in regulatory elements, perhaps leading to two separate expression domains that together recapitulate the more complex single expression pattern of the ancestral gene [161, 163].

Unexpectedly high numbers of duplicated genes belong to categories such as transcription factors, kinases, signaling transducers, and particular enzymes and transporters [164]. Therefore, certain types of genes must have biochemical features that allow them to be adapted easily to novel functions and other types of genes might be particularly unlikely to undergo functional innovation via duplication, because the duplication has an immediate detrimental effect [165].

### 1.3.1. Molecular Adaptation

Adaptive evolution is the process by which an allele that is beneficial to either reproduction or survival increases in frequency as a result of the individual carrying the allele having an increased fitness [166]. Adaptation by natural selection is the most important process in Biology, explaining the incredible complexity and diversity of organisms, cells, enzymes and proteins as all living structures result from the repeated fixation and elimination of genetic variants within populations [167]. The fate of a new genetic variant (mutant) present in a single individual can be driven by three main forces: mutation, natural selection and genetic drift. Although mutation is the ultimate source of all genetic variation, it is by far the weakest of these evolutionary forces, and by itself cannot rapidly change the frequency of the mutant in the population [167]. The effect of selection is to increase the frequency of a beneficial mutation until it becomes fixed in the population (positive selection) or to decrease the frequency of a deleterious mutation until it is eliminated (negative selection), not affecting the frequency of neutral mutations [167]. Therefore, the identification of genes and gene regions subjected to selection can lead to predictions regarding the putative functional important regions of genes [166].

Studies of several gene families indicated that natural selection accelerated the fixation rate of non-synonymous substitutions shortly after a duplication event, presumably to adapt those proteins to a new or modified function [2, 168-170]. However, an accelerated non-synonymous rate also could be driven by a relaxation, but not complete loss, of selective constraints. Here, duplicated proteins evolve under relaxed functional constraints for some period of time, after which functional divergence occurs when formerly neutral substitutions convey a selective advantage in a novel environment or genetic background [168]. Kimura's Neutral Theory [171] maintains that most observed molecular variation (both polymorphism within species and divergence between species) is due to random fixation of selectively neutral mutations. For protein-coding genes, the most compelling evidence for positive selection is derived from comparison of non-synonymous (amino acid replacement) and synonymous (silent) substitution rates, $d_N$ and $d_S$, respectively. The difference between these two rates, measured as the ratio $\omega = d_N/d_S$, reflects the effect of selection on the protein product of the gene [171]. Therefore, if non-synonymous mutations are deleterious, purifying selection (or negative selection) will reduce or prevent their fixation rate and $\omega$ will be less than 1, whereas if non-synonymous mutations are neutral then they will be fixed at the same rate as synonymous mutations and $\omega = 1$. Only under positive selection can non-synonymous mutations be fixed at a rate higher than that of synonymous substitutions, with $\omega > 1$ [172, 173].

Traditionally, to demonstrate adaptive evolution models of neutral evolution and purifying selection must be rejected, that means the $\omega$-value must be shown to be significantly greater than 1 [172, 173]. Models of adaptive evolution by gene duplication make predictions about patterns of genetic changes [152, 153]. After duplication, natural selection favours the fixation of mutations in one or both copies that adapt them to divergent functions. Once new or enhanced functions become established, positive selection ceases and negative selection acts to maintain the new functions. For protein-coding genes, this means non-synonymous substitutions will be accelerated following the duplication, and then slow down due to increased effects of purifying selection [174]. Statistical models of codon substitution relax the assumption of a single $\omega$-value for all branches of a phylogeny [173] and can provide a framework for constructing likelihood ratio tests of changes in selective pressure following gene duplication [175]. Other codon models allow the $\omega$-ratio to vary among amino acid sites [176, 177] and a third type of model can simultaneously account for variation in selective constraints among sites and lineages [178].

However, selection models that use $\omega$-ratios to detect selection are generally not sensitive enough to detect subtle molecular adaptations [179, 180]. One cannot conclude that positive selection has not taken place if $\omega$ is not statistically higher than 1, because even single amino acid changes can be adaptive if they are biochemically superior to extant alternatives. Inversely, it is not recommended to conclude that positive selection occurred if $\omega$ is statistically higher than 1 as non-synonymous mutations can represent different amino acids with similar biochemical properties. Therefore, using $d_N/d_S$ as the unique method to detect positive selection is too conservative to detect single adaptive amino acid changes and is, thus, extremely limited in scope [180]. In order to overcome these limitations, a few additional statistical models are emerging, including those that incorporate changes in quantitative amino acid properties [179, 181].

### 1.3.2. Functional Divergence

It has been widely accepted that following gene duplication, one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes as a result of functional redundancy or positive selection. Unless this type of functional divergence results in some new functions, over time all but one gene copy will be silenced by deleterious mutations [182]. The importance of gene function can be measured quantitatively in terms of the functional constraints of the protein sequence [171]. For instance, an amino acid residue is said to be functionally important if it is evolutionary conserved. Therefore, change of the evolutionary conservation at a particular residue may indicate the involvement of functional divergence [183, 184]. Since gene family proliferation is thought to have provided the raw materials for functional innovations, it is desirable, from sequence analysis, to identify amino acid sites that are responsible for the functional diversity [185, 186]. Because most amino acid changes are not related to functional divergence but represent neutral evolution, it is crucial to develop appropriate statistical methods to distinguish between these two possibilities [187]. Some methods measure the degree of conservation in each position on a sequence alignment and score each position for different subfamilies, with posterior visualization over the tridimensional protein structure [183, 188, 189]. However, new methods were developed, according to observed alignment patterns (amino acid configurations), characterizing two basic types of functional divergence [184-186] (Fig. 9).

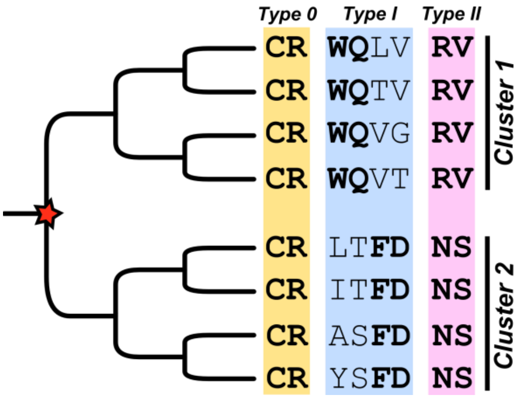


**Figure 9. Types of functional divergence after gene duplication, according to observed amino acid configurations.** Type 0 - amino acid configurations that are universally conserved through the whole gene family; Type I - amino acid configurations that are highly conserved in gene 1 but variable in gene 2, or vice versa; Type II - amino acid configurations that are very conserved in both genes but whose biochemical properties are different. Adapted from [185].

Amino acid configurations can be classified into three types (Fig. 9): Type 0 represents amino acid configurations that are universally conserved through the whole gene family, implying that these residues are important for the common function shared by all member genes; Type I represents amino acid configurations that are highly conserved in gene 1 but variable in gene 2, or vice versa, implying that these residues have experienced altered functional constraints; and Type II represents amino acid configurations that are very conserved in both genes but whose biochemical properties are different, implying that these residues may be responsible for functional specification [185]. According to these amino acid configurations it is possible to define two basic types of functional divergence after gene duplication: Type I functional divergence results in altered functional constraints (i.e., different evolutionary rates) between duplicate genes; and Type II functional divergence results in no altered functional constraints but in a radical change in amino acid properties between them (e.g., charge, hydrophobicity) [185].

One may expect that Type I (or Type II) amino acid configurations are likely to be generated by Type I (or Type II) functional divergence, which is true if the effect of Type I (or Type II) functional divergence has been shown to be statistically significant under a stochastic model [185, 186] (Fig. 9). A fundamental measure for functional divergence after gene duplication is the coefficient of functional divergence $\theta$. It can be interpreted as the decrease in rate correlation ($r$) between two duplicate genes as a result of functional divergence after gene duplication (e.g., $\theta = 1 - r$) [184-186]. On the other hand, the possibility of a site being functional divergence-related (Type I or Type II) can be measured by a posterior probability when the observed amino acid configuration is given.

## 2. OBJECTIVES

The adaptive study of the vertebrate members of the *Hh* gene family can provide valuable insights onto the evolutionary forces acting on each of the vertebrate *Hh* members after duplication, as well onto the distinct functional roles of the codified proteins. Therefore, the main goal of this study is to assess the adaptive evolution of *Hh* genes in vertebrates using a comparative genomics framework at two levels:

I. First, we studied the synteny of vertebrate *Hh* genes to retrace their evolutionary history after the two rounds of wide genome duplication and to assess the lack of a Dhh-coding sequence annotation on currently available avian genome assemblies;

II. Secondly, we evaluated signatures of positive and negative selection and functional divergence, using both a gene and protein-level approach, in order to detect evidences of functional divergence due to functional and structural constraints.

# 3. METHODS

## 3.1. Sequence Collection and Alignment

*Hh* coding sequences were retrieved from the GenBank [138] and ENSEMBL [139] databases and BLAST searches were used to recover non-annotated sequences from avian and other vertebrate genomes (Table S1). Local BLAST databases for avian genomes provided by BGI were created using the Blast+ software package [190] and blasts searches (TBLASTN and BLASTp) were performed over these avian genomes to search for *Hh* coding sequences. All putative sequences identified were confirmed by TBLASTN and BLASTp over the GenBank [138] database. We collected a total of 120 *Hh* coding sequences and reduced it to 50 by excluding the sequences which presented less than 50% represented sites (compared to the *Homo sapiens* sequences) and equally representing each vertebrate class. A codon based coding sequence alignment was performed with the 50 sequences using MUSCLE 3.3 [191], manually adjusted using MEGA 5 [192] and viewed and edited in SEAVIEW [193]. It was previously reported that the alignment of *Hh* sequences produce indels on the C-terminal/3' portion [137] and, as indels carry phylogenetic signal [194], filtering softwares were not applied. To assess the selective pressures acting over the three vertebrate *Hh* paralogs, the alignment was used to produce four different alignments: one for each paralog and a fourth with all the sequences except outgroups. Nucleotide and amino acid conservation over *Hh* sequences was assessed using MEGA 5 [192].

## 3.2. Synteny Analysis

The synteny analysis was performed using the GENOMICUS v64.01 browser [195], which makes an integration of the data available on the ENSEMBL database [139] in order to provide a better visualization of conserved synteny blocks and to reconstruct ancient genomes organization, using the *Homo sapiens* sequences as query. Genes not annotated on the GENOMICUS v64.01 browser [195] were searched on the respective species by TBLASTN and BLASTp over the GenBank [138] and ENSEMBL [139] databases and mapped localizations were annotated in order to compare it with the localization of putative syntenic genes. Local BLAST databases of the avian genomes provided by BGI were created using the Blast+ software package [190] and blasts searches (TBLASTN and BLASTp) were performed over these avian genomes to search for *Hh, LMBR1, RHEB* and *Trx/MLL2,3* coding sequences and relative locations annotated. All putative sequences identified were confirmed by TBLASTN and BLASTp over the GenBank [138] database.

Comparative *Dhh* gene synteny analysis over the reptilian group (birds and non-avian reptiles) was conducted by BLASTn of the GL343198.1 scaffold of the *Anolis carolinensis* anoCar2.0 assembly [196] over the BGI provided *F. peregrinus* and the *Gallus gallus* WUGSC2.1 [197] assemblies. The localization of the *Dhh* gene and the conserved *LMBR1L-Dhh-RHEBL1-MLL2* cluster over the *Anolis carolinensis* genome was accessed from the GENOMICUS v64.01 browser [195], the complete genome assembly was downloaded from the UCSC database [198] and the subject scaffold extracted using UGENE 1.7.2 [199]. The complete *G. gallus* WUGSC2.1 assembly was downloaded also from the UCSC database [198] and local databases of the *F. peregrinus* and *G. gallus* genomes created using the Blast+ software package [190]. BLAST searches were performed using the Blast+ software package [190] and best hits chosen for Score $> 59$, E-value $< 1 \times 10^{-25}$ and $\mathrm{ID} > 85\%$. Circular plots were created using Circos [200].

## 3.3. Experimental Detection

### 3.3.1. Sampling and Genomic DNA Extraction and Purification

Fresh blood and breast muscle from two different adult male chickens (*Gallus sp.*) were collected and fresh blood from a juvenile male peregrine falcon (*Falco peregrinus*) was kindly provided by Parque Biológico de Gaia. Tissues were collected to 15 mL falcon tubes containing 7.5 mL of 0.96% ethanol, and stored at -20°C. In order to prevent Polymerase Chain Reaction (PCR) inhibition, anticoagulant agents were not used [201-204]. Genomic DNA (gDNA) was extracted and purified from both tissue types using three different protocols: salting-out and the Purelink™ Genomic DNA mini Kit's "Blood Lysate protocol" and "Protocol Development Guidelines" (Invitrogen by life technologies, Lisbon, Portugal), in order to evaluate which tissue and protocol yield the best results. Three replicates were produced and purified gDNAs were stored at -20°C. gDNA integrity was evaluated by electrophoresis in a 1.5 % (w/v) agarose gel (Bio-Rad Laboratories Inc., California, USA) stained with ethidium bromide (Bio-Rad Laboratories Inc, California, USA) and posterior concentration quantified using Qubit® Fluorometer (Invitrogen by life technologies, Lisbon, Portugal).

For salting-out gDNA extraction and purification, 0.025 g of tissue (blood or muscle), 500 µL of Lysis Buffer [50 Mm Tris-HCl, 20 Mm EDTA and 2% sodium dodecyl sulfate (SDS)] and 10 µL of 20 mg/mL Proteinase K were added. After incubation at 55°C for 24 hours and a chill for 10 minutes, to the digested tissue 500 µL of saturated NaCl solution was added and the mixture centrifuged at 8000 rpm for 15 minutes at 4°C. One mL of 100% ethanol was added to the supernatant and both phases mixed. After overnight incubation at -20°C, phases were separated by centrifugation at 11000 rpm for 10 minutes at 4°C and the supernatant discarded. The pellet was rinsed with 500 µL of 70% ethanol and centrifuged for 5 minutes at 4°C and 11000 rpm. After draining out the ethanol, gDNA was dried and resuspended in 50 µL of molecular biology ultra-pure water.

The Purelink™ Genomic DNA mini Kit's "Blood Lysate" protocol (Invitrogen by life technologies, Lisbon, Portugal) was applied for gDNA extraction from blood tissues. A total of 0.025 g of blood was used and purified gDNA eluted in 150 µL of molecular biology ultra-pure water. On the other hand, an adapted Purelink™ Genomic DNA mini Kit's "Protocol Development Guidelines" (Invitrogen by life technologies, Lisbon, Portugal) was applied to both blood and muscle tissues. A total of 0.025 g of tissue minced with 180 µL of PureLink™ Genomic Digestion Buffer (K1823-01) and 20 µL of Proteinase K were added and mixed well. Samples were incubated at 55°C overnight and, after lysis was completed, 20 µL of RNase A was added and the mixture incubated at room temperature for 2 minutes. The lysate was centrifuged at maximum speed for 5 minutes at room temperature to remove any particulate material. The supernatant was transferred to a fresh microcentrifuge tube, 200 µL of PureLink™ Genomic Binding Buffer (K1823-02) was added to the lysate and the sample vortexed to yield a homogenous solution. 200 µL of 100% ethanol was added to the lysate and mixed well by vortexing for 5 seconds to yield a homogenous solution. gDNA was bound, washed and eluted in 150 µL of molecular biology ultra-pure water.

### 3.3.2. Primer Design

In order to experimentally detect an avian putative Dhh-coding sequence, we searched for specific regions on Hh-coding sequences that are responsible for the distinction of different *Hh* paralogs. *Hh* genes are highly variable in size, ranging from 5.000 pb up to 36.000 pb both between species and paralogs [139], but usually carry three exons with highly conserved lengths, each coding for specific regions of the Hh proteins (exon 1: Hedge N-terminal, 290-320 pb; exon 2: Hedge C-terminal, 260-270 pb; Exon 3: Hog, 600-700 pb) [137, 139]. Therefore, we divided the coding-sequence alignment in three regions, each corresponding to one exon according to the *Anolis lizard* sequence, and built three phylogenetic trees (one for each partition) using the Neighbor-Joining (NJ) method implemented in MEGA 5 [192] with 16 complete *Hh* coding sequences: *Homo sapiens Shh* (GenBank: NM_000193.2), *Homo sapiens Ihh* (GenBank: NM_002181.3), *Homo sapiens Dhh* (GenBank: NM_021044.2), *Anolis carolinensis Shh* (GenBank: XM_003221928.1), *Anolis carolinensis Ihh* (Ensembl: ENSACAG00000005172), *Anolis carolinensis Dhh* (GenBank: XM_003223232.1), *Gallus gallus Shh* (GenBank:

NM_204821.1), *Gallus gallus Ihh* (NM_204957.1), *Xenopus laevis Shh* (GenBank: NM_001088313.1), *Xenopus laevis Ihh* (GenBank: NM_001085793.1), *Xenopus laevis Dhha* (GenBank: NM_001085791.1), *Xenopus laevis Dhhb* (GenBank: NM_001085792.1), *Gasterosteus aculeatus Shh* (Ensembl: ENSGACG00000003893), *Gasterosteus aculeatus Ihh1* (Ensembl: ENSGACG00000015562), *Gasterosteus aculeatus Dhh* (Ensembl: ENSGACG00000009063) and *Drosophila melanogaster Hh* (Ensembl Metazoa: GA18321-RA).

Comparing each tree with a tree built with the complete coding sequences (control tree), we found that exon 3 is the one responsible for the distinction between each paralog, and the division of this exon into three regions never retrieved a tree similar to the control tree. As a result, we used the previously built alignment to design primers specific for the *Dhh* third exon (according to the *Anolis carolinensis Dhh* gene), searching for conserved regions within *Dhh* orthologs that are not conserved within *Hh* paralogs. Putative oligonuclotides were analyzed with OligoAnalyzer 3.1 [205] for GC content, melting temperature and hairpin, homo- and hetero-dimer formation ability, and specifity comproved by BLAST over the GenBank and Ensembl databases [138, 139]. Due to the high GC content of the *Dhh* third exon, the best primer pair found presents a high GC content and consequently a high melting temperature: F1: 3’-WCNGGNGGCTGBTTNCCNGG-5’ ($T_M$ = 49 °C, 50 nM $Na^+$) R: 3’-GTARAGSAGSCSNGAGTACCA-5’ ($T_M$ = 51 °C, 50 nM $Na^+$). Due to the results obtained with this pair of oligonucleotide primers, a second, non-degenerated, forward oligonucleotide primer was constructed for the Anolis carolinensis Dhh third exon by the alignment of recent avian Dhh-coding sequences: F2: 3’-TAACTCGCTGGCTGTCCGCA-5’ ($T_M$ = 51 °C, 50 nM $Na^+$).

### 3.3.3. Polymerase Chain Reaction (PCR) and Sequencing

In order to determine the best PCR conditions for each set of primers, different annealing temperatures and reaction components’ concentrations were tested. Starting PCR reaction mixtures were prepared using a total volume of 20 µL per reaction, containing 1x PCR buffer, 2.5 mM $MgCl_{2,}$ 1.0 mM of each dNTP, 1 unit of Biotaq$^{TM}$ DNA Taq polymerase (Bioline, Luckenwalde, Germany), 10.9 µL molecular grade PCR $H_2O$ (AccuGENE ®, Lonza, Verviers, Belgium), 0.5 µM of both forward (F1) and reverse (R) primers (Invitrogen by life technologies, Lisbon), and finally 2 µL of gDNA template (Testing template: *Gallus sp.* gDNA; Positive control: *Falco peregrinus* gDNA; Negative control: molecular grade PCR $H_2O$). Using Biometra T-Professional standard thermocycler (Biometra, Goettingen, Germany), the following PCR cycling conditions were used: initial denaturation 2 min at 94 °C, followed by 35 cycles of 1 min at 94 °C, 30 sec at annealing temperature and 1 min extension at 72 °C, and a final extension of 10 min at 72 °C. For the F1xR primer pair, a first annealing temperature gradient between 48 and 60 °C and a second annealing temperature gradient between 47 and 50 °C were tested. For the F2xR primer pair, a first annealing temperature gradient between 47 and 57 °C and a second annealing temperature gradient between 48 and 53 °C were tested.

Further adjustments were applied to the reaction mixtures, testing the double of DNA quantity, the double of total reaction volume, $MgCl_2$ concentrations of 1.0 mM, 1.25 mM and 2.5 mM, dNTP concentrations of 0.75 mM and 0.5 mM and the presence and absence of Bovine Serum Albumin (BSA). Further adjustments were also applied to the PCR cycling conditions, testing 30 and 37 cycles, with 40 sec extensions. Amplifications were confirmed by electrophoresis in 1.5 and 2.0 % (w/v) agarose gel (Bio-Rad Laboratories Inc., California, USA) stained with ethidium bromide (Bio-Rad Laboratories Inc, California, USA).

Bands of interest were extracted and purified using a modified PureLink® Quick Gel Extraction and PCR Combo Kit (Invitrogen by life technologies, Lisbon, Portugal) Centrifugation protocol. After excising and dissolving the gel piece containing the DNA fragment of interest, it was pipetted into the center of a PureLink™ Clean-Up Spin Column inside a Wash Tube and centrifuged at 10.000 rpm for 1 min. The flow-through was discarded and 500 µL of Wash Buffer containing ethanol was added. The column was again centrifuged at 10.000 rpm for 1 min, the flow-through discarded, and a third round of centrifugation at maximum speed for 3 min applied to remove any residual Wash Buffer and ethanol. The column was incubated at 55 °C for 5 min and the Wash Tube discarded. The PureLink™ Spin Column was placed into an Elution Tube and 30 µL of molecular grade PCR $H_2O$ (AccuGENE ®, Lonza, Verviers, Belgium) added to the center of the column. After incubation for 1 min at room temperature, the column was centrifuged at 10.000 rpm for 1 min and the PureLink™ Spin Column discarded. Purified DNA was sequenced directly (Macrogen-Advancing through genomics, South Korea) and the results analyzed using the UGENE 1.7.2. [199], FinchTV 1.4 [206] and Geneious™ Pro v5.4 [207] softwares.

### 3.4. Phylogenetic Analyses

For phylogenetic analyses, the substitution model that best fit our dataset (GTR+I+G) was selected using the Akaike Information Criterion (AIC) implemented in jModelTest [208], starting with 11 substitution schemes and using the fixed BIONJ-JC base tree for likelihood calculations. The dataset was checked for saturation bias in DAMBE [209], both by plotting the rate of transitions and transversions versus the genetic distance and by applying the Xia *et al.* test [210] to measure substitution saturation. By plotting the observed number of transitions and transversions against the genetic distance, transitions and transversions should both increase linearly with the genetic distance, with transitions being higher than transversions [211]. On the other hand, the Xia *et al.* test [210] compares half of the theoretical saturation index expected when assuming full saturation ($I_{SS.C}$, critic value) with the observed saturation index ($I_{SS}$). If $I_{SS}$ is significantly lower than $I_{SS.C,}$ the data has no evidences of saturation bias and can be further used for phylogenetic analysis. The phylogeny was estimated using the Maximum Likelihood (ML) and Bayesian inference methods. The ML phylogenetic tree was constructed in PhyML 3.0 [212], with 1000 bootstrap replicates and the NNI branch search algorithm. Bayesian inference methods with Markov chain Monte Carlo (MCMC) sampling were preformed in MrBayes [213, 214], with 100000 generations, a sample frequency of 100 and burn-in set to correspond to 25% of the sampled trees. For site tests of the *Hh* vertebrate paralogs, independent phylogenies for each gene were produced.

## 3.5. Adaptive Selection Detection

### 3.5.1. Codon-Level Analysis

About 40% of the four codon alignments produced was filtered with GBLOCKS 0.91 [215, 216], applying the less stringent method, and used with the ML/Bayesian trees in the program codeml from the PAML v4.3 package [217] in order to evaluate adaptive evolution in the *Dhh, Ihh* and *Shh* coding sequences. To examine the ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) (the dN/dS or ω ratio), the branch-specific and site-specific codon substitution models of maximum likelihood analysis were used.

For branch tests, four likelihood ratio-tests (LRT) were preformed to compare the log likelihood values of a two-ratio model, where the selected post-duplication branch has a different evolutionary rate relative to other branches (model = 2, NS sites = 0), against a one-ratio model, where all branches are supposed to evolve at a same rate (model = 0, NS sites = 0) [173]. The two-ratio (unconstrained two-ratio) model, if found to fit the data better with $\omega_1 > 1$, was tested against another null (constrained two-ratio) model where the $\omega_1$ value for the branch of interest was constrained to $\omega_1 < 1$ fixing $\omega_1 = 1$. The LRT between these two nested two-ratio models allows the detection of the prevalence of positive selection or relaxed selective constraints [173]. Hypothesis decision was performed assuming that LRT approximately follows the chi-square $2\Delta lnL$ approximation ($\mathrm{P} < 0.05$), the double of the difference between the alternative and null model log likelihood [177]. LRT degrees of freedom are calculated as the difference of the number of parameters between the nested models. Individual two-ratio models were created using as foreground branch each one of the branches to test: the branch leading to the *Dhh* group, the branch leading to the *Ihh/Shh* group and the branches leading to the *Shh* and *Ihh* groups.

However, this lineage-base analysis assume that all amino acid sites are under the same selective pressure and it is a very conservative test of adaptive evolution, as many sites can be evolving at a different rate [218]. Thus, in order to detect signatures of adaptive evolution over the *Dhh*, *Ihh* and *Shh* codon sequences, three smaller phylogenetic trees were built for each group and each topology used for site analysis with PAML v4.03 [217]. Two LRTs were preformed to compare the log likelihood values of two nested models, a model that does not allow and a model that allows sites to be under positive selection [177]. First, the M0 (uniform selective pressure among sites; model = 0, NS sites = 0) and M3 (variable selective pressure among sites; model = 0, NS sites = 3) models were compared; and finally the M7 (beta distributed variable selective pressure; model = 0, NS sites = 7) and M8 (beta plus positive selection; model = 0, NS sites = 8) models. The identification of sites under positive selection was performed by Bayes Empirical Bayes (BEB) analysis [219].

As the BEB method does not detect negatively selected residues and PAML is not able to access purifying selection [217, 219], we used the Single Likelihood Ancestor Counting (SLAC) and the Fixed Effects Likelihood (FEL) methods [220], implemented in the Datamonkey web server [221, 222], in order to detect signatures of purifying selection over the data. SLAC is a modified and improved derivative of the Suzuki-Gojobori counting approach that maps changes in the phylogeny to estimate selection on a site-by-site basis and it calculates the number of non-synonymous and synonymous substitutions that have occurred at each site using ML reconstructions of ancestral sequences [220, 221]. On the other hand, the FEL model estimates the ratio of non-synonymous to synonymous substitutions not assuming *a priori* distribution of rates across sites substitution on a site-by-site analysis [220].

Since the *Dhh* and *Shh* avian sequences, as well the turkey *Ihh* sequence, are incomplete, these sequences were removed from the analysis, in order to improve the calculations and reduce the number of ambiguous sites.

### 3.5.2. Amino Acid-Level Analysis

We analyzed destabilizing selection over our data, as selection models that use ω ratios to detect selection on protein-coding genes are generally not sensitive enough to detect subtle molecular adaptations in conserved protein-coding genes. ω ratios models can fail on the detection of positively and negatively selected sites as they do not allow the possibility that adaptation may come in the form of very few amino acid changes and do not provide information on the chemical and structural variations caused by these amino acid replacements [179, 180, 223, 224]. Thus, a statistical approach that looks for deviations of the observed amino acid properties relative to the expectation under neutrality is necessary.

In order to detect destabilizing selection signatures over *Dhh, Shh* and *Ihh* coding sequences, the three codon alignments and ML/Bayesian trees used for site-selection analysis where analyzed with the method implemented in TreeSAAP [225], finding which sites and significant physicochemical properties can be under positive and negative selection over the three analyzed lineages. TreeSAAP [225] compares the observed distribution of physicochemical changes inferred from the phylogenetic tree with an expected distribution based on the assumption of completely random amino acid replacement expected under the condition of selective neutrality. The evaluation of the magnitude of property change at non-synonymous residues and their location on a protein tridimensional structure may provide important information into the structural and functional consequences of the substitutions [179, 180].

Eight magnitude categories (1 to 8) represent one-step nucleotide changes in a codon and rank the correspondent variation in a property scale of the coded amino acid: categories 1 to 3 indicate stabilizing substitutions (small variations that tend to maintain the overall biochemistry of the protein) while categories 6 to 8 represent destabilizing substitutions (variations that result in radical structural and functional shifts in local regions of the protein). By accounting for the property changes across the data set, a set of relative frequencies changes for each category is obtained allowing to test the null hypothesis under the assumption of neutral conditions: (1) positive selection is detected when the number of inferred amino acid replacements significantly exceed the number expected by chance alone, resulting in positive Z-scores; (2) negative selection is detected when the expected number of amino acid replacement significantly exceeds those that are inferred, resulting in negative Z-scores [179, 180]. To detect both strong negative and positive selective pressures, only changes corresponding to categories 7 and 8 at the $P \leq 0.05$ (Z-score $> |1.64|$) and $P \leq 0.001$ (Z-score $> |3.09|$) levels were considered, due to the strong purifying signatures over our data. A total of 31 amino acid properties [180] were evaluated for each paralog and, to verify which specific regions were affected by negative and positive destabilizing selection, we performed a sliding window analysis using the properties which were significant for the signal. Sliding windows of 10 amino acid length with a sliding step of one codon were selected to show the best signal-to-noise ratio and to identify regions in the vertebrate *Hh* proteins that differ significantly from a nearly neutral model [226]. In addition, we identified the total number of changes per site assuming it as the sum of those occurring in each branch of the phylogeny [223].

### 3.6. Functional Divergence Analysis

The detection of functional divergence was carried out with DIVERGE 2.0 [227], using the Gu2001 method [185] for Type I functional divergence and the Gu *et al.* method [186] for Type II functional divergence. Type I functional divergence represents amino acid residues that are universally conserved through one subfamily but highly variable in another, implying that these residues have experienced altered functional constraints after duplication [185]. On the other hand, Type II functional divergence represents amino acid configurations that are much conserved in each subfamily but whose biochemical properties are very different, implying that these residues may be responsible for functional specification [186].

The coefficient of Type I and Type II functional divergence ($\hat{\theta}_I \text{ and } \hat{\theta}_{II}$) between each pair of *Hh* paralogs was estimated. A $\hat{\theta}$ parameter significantly greater than zero means that either altered selective constraints or a radical shift of amino acid physicochemical properties after gene duplication were likely to have occurred. LRT calculations for the null hypothesis (i.e., the absence of functional divergence) were performed to assess the significance of the $\hat{\theta}$ parameter. In order to detect which residues are more likely to be responsible for functional divergence, the posterior probability [$P(S_1|X)$] for the functional divergence for each position in the alignment was calculated. The cut-off value for the posterior probability was first set to $P(S_1|X) > 0.5$, which corresponds to a posterior odd ratio $R(S_1|S_0) = P(S_1|X)/P(S_0|X) > 1$ and to a meaningful evidence [228]. A more stringent cutoff was selected based on the Harold Jeffreys scale for interpretation of $R(S_1|S_0)$, selecting $P(S_1|X) \geq 0.91$ as it corresponds to $R(S_1|S_0) \geq 10$ (strong evidence) [229].

### 3.7. Protein Structural Modeling and Manipulation

Only the tridimensional structures of the two separated Hedgehog domains are currently available on the Protein Data Bank (PDB) [24, 230]: the human and murine ShhN, IhhN and DhhN regions and the *Drosophila melanogaster* HhN and HhC domains. Thus, we used the PDB: 3HO5 (Human ShhN), PDB: 2WFR (Human DhhN) and the PDB: 3K7G (Human IhhN) files in order to represent the Hedge domain of the human Hh proteins and modeled the tridimensional structure of the human ShhC, IhhC and DhhC domains using I-TASSER [231], a platform for protein tridimensional structure and function prediction implemented on the I-TASSER server [232] that combines *ab initio* and comparative modeling approaches to generate a high quality tridimensional model and has been ranked as the best method for automated protein structure prediction in the last CASP experiments [233-238].

The I-TASSER platform measures the quality of the generated model using two different scoring functions: (1) the C-score is a confidence score for estimating the quality of the predicted models and it is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations [239]; (2) the TM-score is a scale for measuring the structural similarity between two structures and is used to measure the accuracy of structure modeling when the native structure is known in order to test if the result topology is not random [240]. As in these cases the native structure is not known, the TM-score is calculated based on the C-score [239]. To accurately infer the correct topology, the model should have a C-score above -1.5, varying from [-5;2], and TM-score above 0.5 [239, 240] (Table 1).Visualization and manipulation of the generated models, as well as root-mean-square (RMSD) deviation values determination, were assessed with PyMol [241].

**Table 1.** Quality scores for modelled ShhC, IhhC and DhhC protein domains, determined using I-TASSER [239, 240].

| | *C-Score* | *TM-Score* |
|---|---|---|
| **DhhC** | -1.44 | 0.54±0.15 |
| **IhhC** | -2.01 | 0.47±0.15 |
| **ShhC** | -2.65 | 0.41±0.15 |

# 4. RESULTS

## 4.1. Evolution at The Genomic Level

As it was expected to find a *Dhh* gene on avian genomes [136, 137], we searched for the synteny of this gene on the major groups of vertebrates using the GENOMICUS v64.01 browser [195] and compared it with the synteny of the other two vertebrate members of the *Hh* gene family (Fig. 10). We observed that the *Dhh* gene forms a conserved cluster with the *LMBR1L, RHEBL1* and *MLL2* genes in all the tetrapods available on the database, with the exception of birds. Teleost fishes present a similar cluster composed by the *LMBR1L*, *Dhh* and *MLL2* genes, where the *Dhh* and *MLL2* genes are adjacent to each other and the *RHEBL1* gene is found separated from these genes.

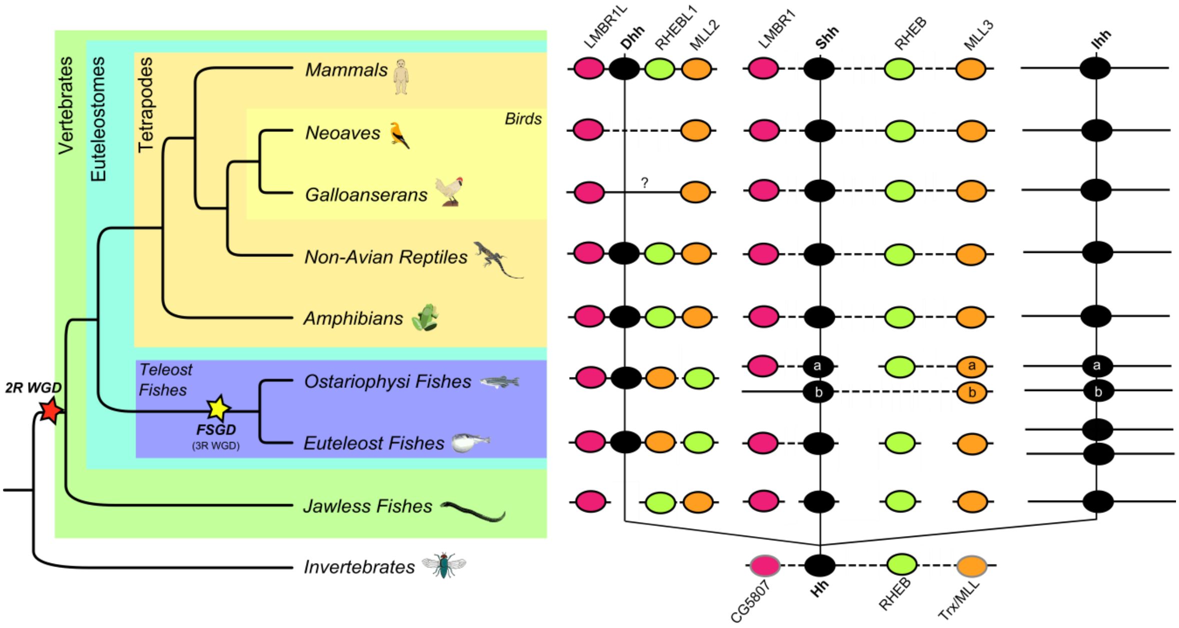


**Figure 10. Illustrative representation of the presence of *Hh* and syntenic related genes in vertebrates according to Genomicus 64.01 [195] and the GenBank [138] and ENSEMBL [139] databases.** The close synteny of the mammal *Dhh* gene was used as reference as the *Dhh* member of the *Hh* gene family in vertebrates is the most ancient one. A doted line between two genes is equivalent to a gap in the alignment, i.e. the two genes are neighbors in this species but not in the reference species, where their orthologs are separated by one or more genes. On the other hand, a large white space represents that the genes are found on the subject genome but are located on different chromosomes/scaffolds. A question mark (?) represents that the syntenic relationship is not known. Genes outlined by a black line where found using Genomicus 64.01 [139] and genes outlined by a grey line where found by blast searches over the GenBank [138] and ENSEMBL [139] databases. The abcense of a gene represents that that gene is not anottated on Genomicus 64.01 and was not found by blast searches.

Interestingly, we observed that paralogs of the *LMBR1L, RHEBL1* and *MLL2* genes are found on the same chromosome/scaffold of the *Shh* gene on the genome of all tetrapods, but none near the *Ihh* gene (Fig. 10). The *LMBR1, RHEB* and *MLL3* genes are located linear to the *Shh* on the genomes of most tetrapods in the same order of that found for the *LMBR1L-Dhh-RHEBL1-MLL2* cluster but are divided by other several genes, forming a larger cluster that is present at least on tetrapods [195]. Similarly to the previously observed, teleost fishes also present a similar cluster (Fig. 10), with the *Shh* and *LMBR1* genes found together on the same chromosome/scaffold. However, on this case, *RHEB* and *MLL3* are also separated from this cluster. *Danio rerio* (zebra fish), the only representative of an ostariophysi fish available on the server, carries a duplication of the *Shh* gene: *Shha* and *Shhb* [140]; and we noticed that the *Shha* and *LMBR1* genes are found on the same chromosome separated from the *RHEB* and *MLL3* genes. The *MLL3* gene is also duplicated on the genome of teleost fishes [195, 242] and, on the case of *D. rerio*, the *MLL3a* gene is found on the same chromosome of *RHEB* and *MLL3b* on the same chromosome of *Shhb.* However, on the genome of euteleost fishes, the *RHEB* and *MLL3* duplicates are found separated (Fig. 10).

Searching for these genes on the genome of *Petromyzon marinus* (sea lamprey), the only representative of jawless fishes available on the server, orthologs of all these genes are found annotated (with the exception of the *Dhh* gene, as expected [145]). However, it was not possible to study synteny as the currently available lamprey genome assembly (WUGSC v3.0) is not fully complete and each of the subject genes is found on different small scaffolds (Fig. 10). On the other hand, we observed that one *RHEB* gene is annotated on the genome of *Drosophila melanogaster* and that this gene locates on the same chromosome of the *Hh* gene. When the *LMBR1* and *MLL2/3* genes where searched on the *D. melanogaster* genome, it was not possible to find any result. We performed blast searches (TBLASTN and BLASTp) on the GenBank [138] and ENSEMBL [139] databases to determine if *LMBR1* and *MLL* genes are present on this invertebrate genome and we found that the *CG5807* (the *D. melanogaster* homolog of the *LMBR1* genes [138]) and *Trx* (the *D. melanogaster* homolog of the *MLL* genes [243, 244]) genes are found on the same chromosome, linearly to the *Hh* and *RHEB* genes and on the same order of that found for the clusters described above (Fig. 10) but separated by larger gene gaps.

### 4.1.1. *Dhh* Gene Synteny on Birds

Despite the described results for tetrapods, the conserved *LMBR1L-Dhh-RHEBL1-MLL2* cluster was not found on the genome of the current three avian species available on the GENOMICUS v64.01 browser [195] (Fig. 10). Only the *LMBR1L* and a *MLL2* gene can be found on the genome of the Neoave *Taeniopygia gutatta*, located on the same chromosome (Un_random) separated by a large gap of genes, and on the genomes of the Galloanserae *Gallus gallus* and *Meleagris gallopavo*, only the *MLL2* is annotated (Fig. 10). As this can be due to a lack of gene annotation, we performed blast searches (TBLASTN and BLASTp) to determine if the absent genes are actually present on the four avian genome assemblies available to date (WUGSC2.1 [197], TGC Turkey_2.01 [245], Anas platyrhynchos 1.0 and WUGSC3.2.4 [246]). In any case, it was possible to identify these genes.

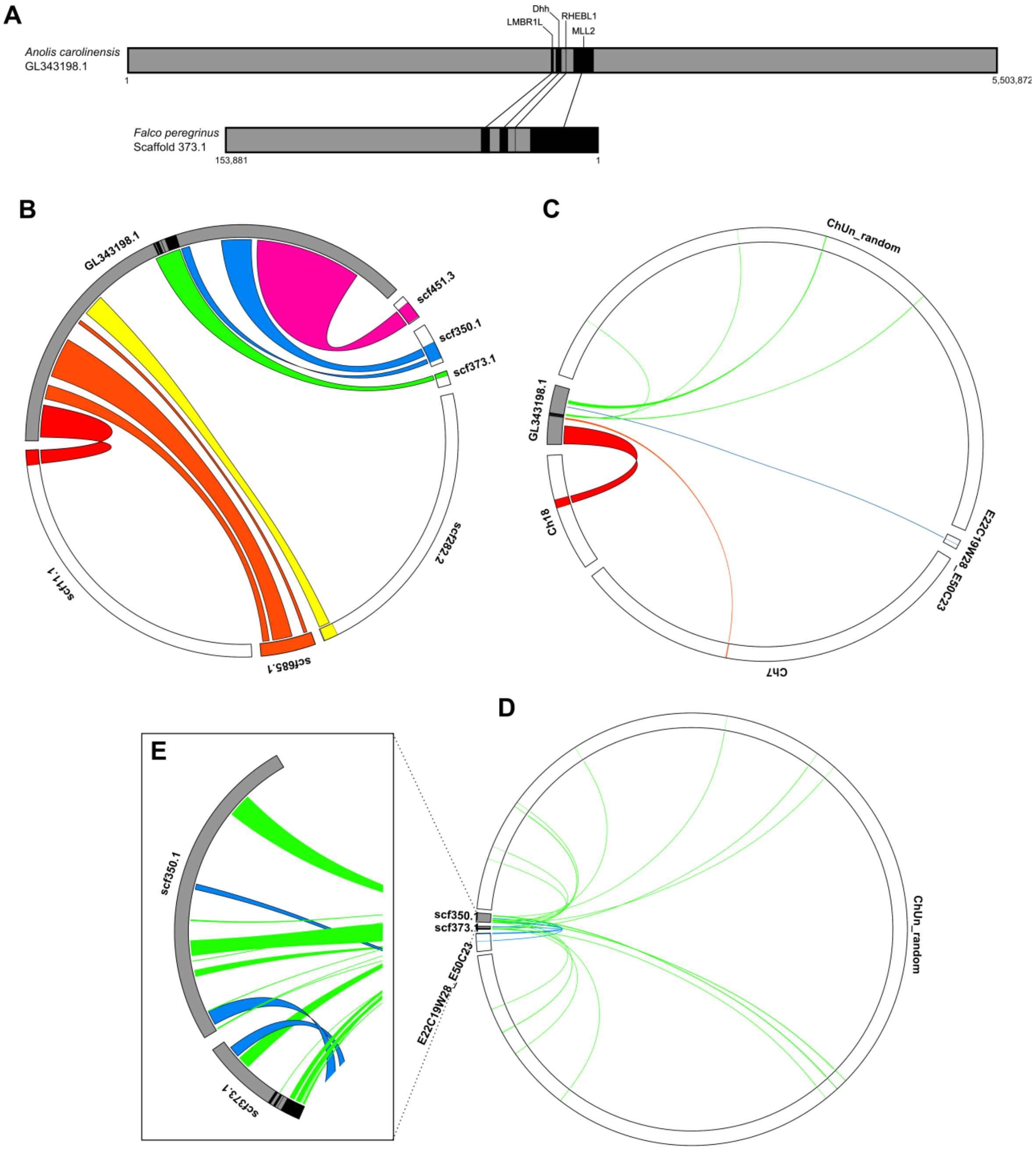


**Figure 11. Homology between the *Anolis carolinensis Dhh* gene synteny and the *Falcon peregrinus* and *Gallus gallus* genomes.** (a) The tetrapod *LMBR1L-Dhh-RHEBL1-MLL2* gene cluster is found on the scaffold GL343198.1 scaffold of the *A. carolinensis* assembly (anoCar 2.0 [196]), and a similar cluster is also found on the 373.1 scaffold of the *F. peregrinus* genome assembly. (b) 6 main *F. peregrinus* scaffolds shows great homology for specific regions of the lizard GL343198.1 scaffold, (c) on the *G. gallus* genome homology is found on 2 macrochromosomes, a linkage group and the Un_random chromosome. (d) The 350.1 and 373.1 *F. peregrinus* scaffolds have high homology with several random regions of the *G. gallus* Un_random chromosome. (e) Hits for the *F. peregrinus* cluster are found on *G. gallus* genome mainly for the *MLL2* gene.

As the BGI Bird Phylogenomic Project provided us privileged access to their recently sequenced avian genomes, we performed BLAST (TBLASTN and BLASTp) searches to determine if *Hh*, *LMBR1, RHEB* and *MLL2/3* genes are present on other avian genomes, mainly Neoaves. The best results were obtained from the *Falco peregrinus* (peregrine falcon) and *Melopsittacus ondulatus* (budgerigar) genomes, as their assembly is more complete and we were able to find homologues sequences for all the subject genes. On the case of *F. peregrinus*, we identified the 373.1 scaffold as carrying the conserved tetrapod *LMBR1L-Dhh-RHEBL1-MLL2* cluster (Fig. 11A). However, although we have found all the genes that compose this cluster, on the *M. ondulatus* genome all of them are separated in small scaffolds and it was not possible to study synteny. The *MLL3, RHEB*, *Shh*, *LMBR1* and *Ihh* genes were also found on both genomes with a similar organization of that found for the other tetrapods, however, on both cases, the *LMBR1-Shh* and the *RHEB-MLL3* groups were separated on different small scaffolds.

As these two species bring evidences of the presence of *Hh*, *LMBR1, RHEB* and *MLL2/3* genes in Neoaves, a question still remains: why it is not possible to find evidences of some of these genes on Galloanserans? We were not able to access other Galloanserae genomes than the currently available *G. gallus* (WUGSC2.1 [197]), *M. gallopavo* (TGC Turkey_2.01 [245]) and *A. platyrhynchos* (Anas platyrhynchos 1.0) assemblies. As the *G. gallus* genome has an overall high quality [138, 139, 198] and the lizard *Anolis carolinensis* is the tetrapod most closely related to birds whose genome have been sequenced to date [196], we compared the synteny of the *Dhh* gene in *A. carolinensis*, *F. peregrinus* and *G. gallus*. The tetrapod *LMBR1L-Dhh-RHEBL1-MLL2* cluster is found on the *A. carolinensis* GL343198.1 scaffold (Fig. 11A) [138, 139, 198] and we used this complete scaffold as query on a BLASTn over two local databases of the *G. gallus* and *F. peregrinus* genomes to find where on these genomes we can find sequences similar to the ones found on the GL343198.1 lizard scaffold, confirming the results by aligning the best hits scaffolds/chromosomes with the lizard GL343198.1 scaffold using Mauve 2.3.1. [247].

We found that there are six main *F. peregrinus* scaffolds that shows great homology for specific regions of the lizard GL343198.1 scaffold (Fig. 11B) while on the *G. gallus* genome we found homology on two macrochromosomes, a linkage group and the Un_random chromosome (Fig. 11C). Although the correspondences on the subject genomes were easily found for the regions on the GL343198.1 scaffold outside the region where the *Dhh* close synteny is found, it was more difficult to make an accurate correspondence within the *Dhh* and syntenic region (Fig. 11B and C). This can be explained by the fact that upstream the *LMBR1L* gene on the lizard scaffold we find three genes members of the Tubulin-α family [139, 195], a highly conserved and numerous family of genes coding for an important structural family of proteins [248, 249]. However, we were able to find hits beneath this region: on the 373.1 scaffold for the falcon assembly, without dispersed hits as expected, and on the Un_random chromosome for the chicken assembly, with highly dispersed hits (Fig. 11B and C).

On the case of the *F. peregrinus* assembly, we found that the 350.1 scaffold shares homology with a region closely located downstream the *Dhh* and syntenic region (Fig. 11B), and on the case of the *G. gallus* assembly this region shares high homology with part of the E22C19w28_E50C23 linkage group (Fig. 11C). So that, we chose the 350.1 and 373.1 *F. peregrinus* scaffolds as query on a BLASTn over the local database of the *G. gallus* genome and noticed that the 373.1 *F. peregrinus* scaffold has high homology with several random regions of the *G. gallus* Un_random chromosome (Fig. 11D), but its 5' extremity has a highly homology with the 3' extremity of the *G. gallus* E22C19w28_E50C23 linkage group. Similarly, the *F. peregrinus* 350.1 scaffold carries regions with high homology with random positions on the *G. gallus* Un_random chromosome but also two specific regions with homology for two regions of the *G. gallus* E22C19w28_E50C23 linkage group, one of them closely located with the 373.1 scaffold hit. This may suggest that the *F. peregrinus* 373.1 and 350.1 hits may assemble on each other, however when the assembly and alignment of both scaffolds wass performed it was not possible to build a sequence as no contig was found. On the other hand, when we look to the region of the 373.1 scaffold where the *Dhh* and syntenic genes are found on the falcon genome (Fig. 11E), we find hits on the *G. gallus* genome only for the regions encompassing the *MLL2* gene, that are dispersedly located on the Un_random chromosome.

### 4.1.2. Detection of a Dhh Coding Sequence on Avian Genomes

Due to the lack of a Dhh-coding sequence annotation on the public available avian genome assemblies, we experimentally searched for putative coding sequences on the genome of *Gallus sp,* using *Falco peregrinus* as a positive control. As avian red blood cells are nucleated and blood is a simple tissue to collect [203], we collected fresh blood from two male chicken right after killing and Parque Biológico de Gaia provided us access to fresh peregrine falcon blood, and we used these tissues as a source of genomic DNA (gDNA) for the polymerase-chain reaction (PCR). Blood has as a disadvantage to coagulate right after blood vessel injury and the clot formed could be seen as a limitation requiring the use of anticoagulant agents (e.g., citrate or EDTA [201-204]) on the moment of collection. However, the use of anticoagulant agents is only necessary if we want to use all the blood as a source of material to study and they seem to influence the efficiency of the PCR process [201-204]. Therefore, since for this study the gDNA extraction is the major goal, clot formation is not an important issue and anticoagulant agents were not applied. However, we noted that it was required an increased digestion time due to the formation of a tissue gel. Therefore, to overcome any problem that blood tissue could bring and also to compare its yield, we additionally collected muscle from the breast of both male chicken sources. Comparing the gDNA yield from both chicken blood and muscle tissues and the three DNA extraction and purification methods (Fig. 12), a better yield was obtained using blood as a source of gDNA and the PureLink™ Protocol Development Guidelines as the extraction method. It is noteworthy that this protocol was the best for both tissues, but much more effective for fresh blood tissues. Therefore, this was the elected method for falcon (*Falco peregrinus*) gDNA extraction and the gDNA purified used for posterior tests and analysis.

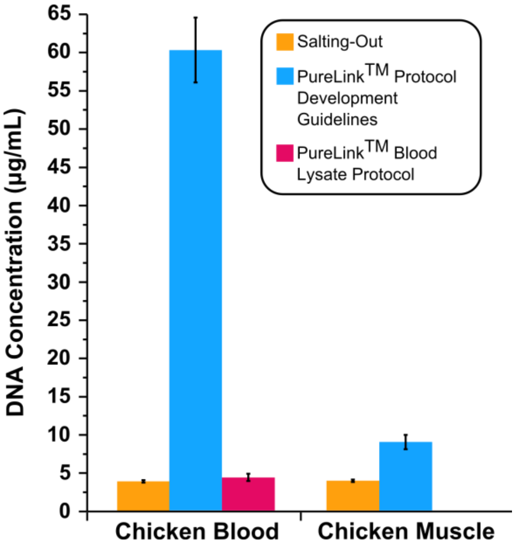


**Figure 12. Yields from chicken fresh blood and breast muscle using salting-out, PureLink™ Genomic DNA mini Kit's Protocol Development Guidelines and Blood Lysate Protocol.** Error bars represent standard deviation (three replicates). In both cases, the PureLink™ Protocol Development Guidelines provided the best results, but it was much more effective for blood tissues.

In order to define the best PCR conditions for each of the oligonucleotide primers pairs, we performed several gradients and tested varied PCR cycling conditions. For the degenerated oligonucleotide primers pair (F1xR), resolved bands were only observed for chicken samples (Fig. 13a) in a reaction with a final volume of 40 µL, template gDNA volume of 8 µL per 40 µL of reaction and in the presence of Bovine Serum Albumin (BSA), at an annealing temperature of 48 °C. For each conditions tested, it was never possible to find resolved bands for falcon samples (Fig. 13b). Despite the primer set being designed to specifically amplify an incomplete putative Dhh-coding sequence, it was expected to observe a maximum of three resolved bands due to its degenerancy. However, when it was possible to observe resolved bands, a minimum of four bands were always detected: one intense band at 850-650 pb, two light bands at 650-500 pb and one clear band at 400-500 pb. According to the expected band sizes for *Hh* amplification determined in accordance to the nucleotide multiple sequence alignment, *Dhh* (490 pb) and *Ihh* (465 pb) fragments would fit within the smaller band and *Shh* (500 pb) into one of the less intense bands. Therefore, we purified these four bands, and direct sequencing of purified products resulted into highly ambiguous sequences that did not matched to any sequence already annotated on GenBank [138] and Ensembl [139] databases.

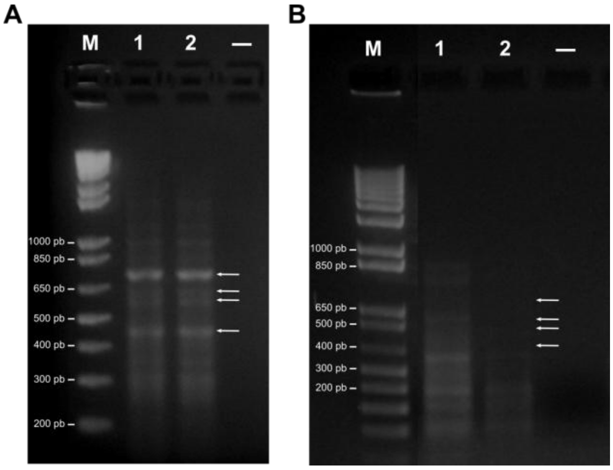


**Figure 13. Agarose gel (1.5% w/v) electrophoresis of avian fresh blood PCR products, using the F1xR primer set at best PCR conditions.** (a) Using *Gallus gallus* fresh blood as a source of gDNA, a minimum of four distinct bands (white arrows) are observed at 650-850, 500-650 and 400-500 pb. (b) However, when *Falco peregrinus* gDNA is analyzed, it is not possible to observe the four bands.

Interestingly, while this work was being preformed a new draft of the *G. gallus* genome assembly (Gallus_gallus-4.0) was released and incomplete predicted Dhh-coding (GenBank: XM_003643524) and LMBR1L-coding (GenBank: XM_003643389) annotations were added in different Un_random chromosome contigs. As the Dhh-coding sequence corresponds to the region that we intended to amplify, we used it to design a new non-degenerated oligonucleotide primer pair. This primer set was tested for both gDNA samples and for several reaction mixtures and PCR cycling conditions. However, in any case it was possible to observe amplification. Despite these results, the opened question about the absence or presence of a Dhh-coding sequence over avian genomes is already uncovered, as suggested by the presence of a partial Dhh-coding sequence on the new draft of the *G. gallus* genome and on the other studied avian genome assemblies.

## 4.2. Evolution at The Gene and Protein Level

At the coding sequence level, the three vertebrate *Hh* paralogs sequences used in this study share a high similarity, with a mean of 0.57 substitutions per site between *Shh* and *Ihh* and 0.67 between *Ihh* and *Dhh* and between *Dhh* and *Shh* (Fig. 14). This relation is also found at the protein level, where the Shh and Ihh proteins share 64.1% of their protein sequences, while Ihh and Dhh share 59.91% and Dhh and Shh 60.9%. Within each group: all the *Shh* sequences present a mean of 0.37 substitutions per site, revealing 78.3% of similarity between Shh proteins; *Ihh* sequences present a mean of 0.49 substitutions per site corresponding to a 70.5% of protein similarity; and *Dhh* sequences differ with a mean of 0.58 substitutions per site and share 59.91% of their protein sequence (Fig. 14). The same analysis was not performed for *Hh* coding sequences as the represented groups are highly divergent.

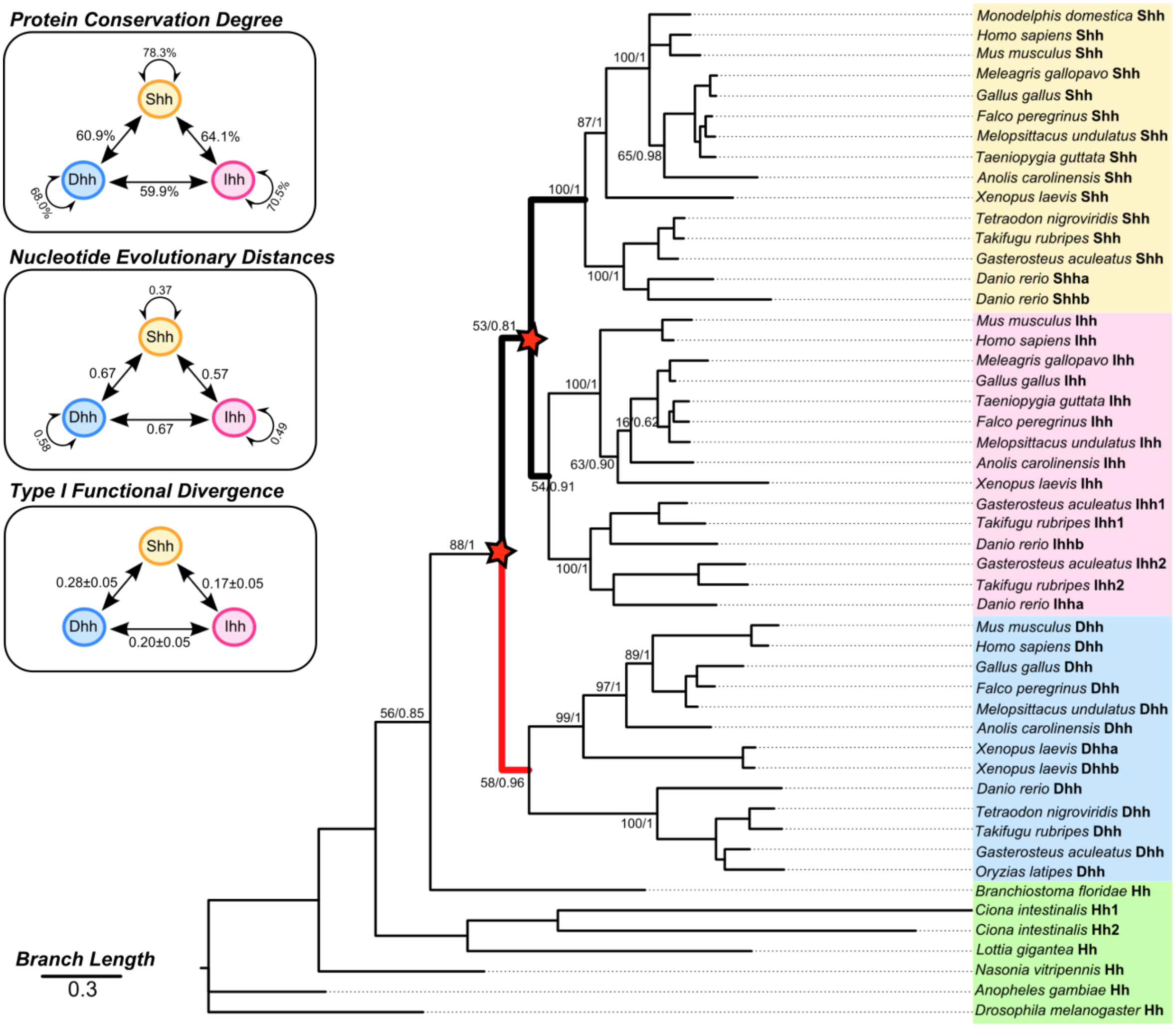


**Figure 14. Phylogenetic relationship of *Hh* coding sequences.** The phylogenetic tree was constructed using Maximum Likelihood (PhyML [212]) and Bayesian inference (MrBayes [213, 214]) algorithms, with supporting values as branch labels (ML/Bayesian). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The post-duplication branches tested with the branch model implemented in PAML [217] are represented in bold and the faster evolving ones are coloured red. The degree of similarity between Hh proteins and the evolutionary distances between *Hh* coding sequences was inferred using MEGA 5 [192] and the Type I functional divergence coefficient values ($\theta_I$) between Hh proteins were inferred using DIVERGE 2.0 [227].

Although the saturation plot suggests a lower extent of substitution saturation, no statistically significant evidence of saturation was found for our dataset (Fig. 15 and Table 2). Therefore, the phylogenetic analyses of the 50 *Hh* coding sequences showed similar overall topologies with both Bayesian and Maximum Likelihood (ML) methods. In agreement with previous works [6, 136, 137], the two phylogenetic methods used retrieved the (Hh,(Dhh,(Ihh,Shh))) topology (Fig. 14), which is compatible with the conservation and distances quickly retrieved from the multiple sequences alignment. The similarities and distances determined shows that the three vertebrate Hh paralog groups are highly conserved but that after duplications the *Shh* group must have been under more constrained evolution, while Ihh and Dhh must be evolving under increasingly relaxed constraints. However, further adaptive evolutionary analyses are necessary to understand the forces influencing this gene family evolution and thus we used the built phylogenetic tree for the detection of selection signatures and functional divergence in the vertebrate members of the Hedgehog family.

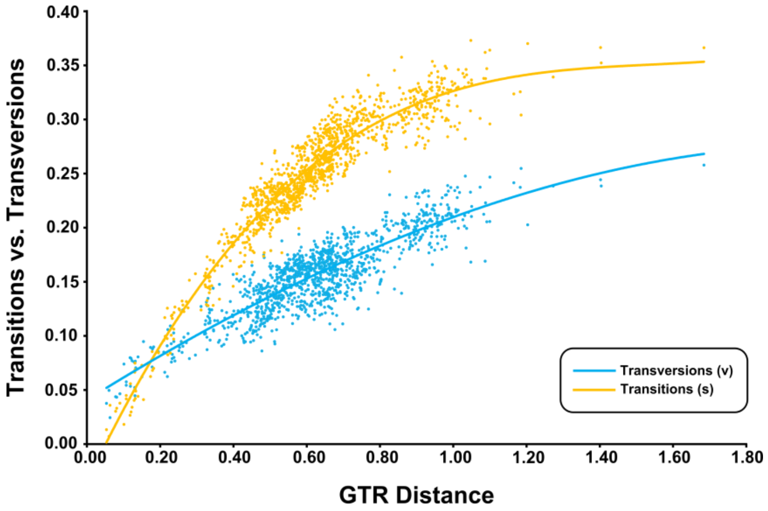


**Figure 15. Nucleotide saturation plot for vertebrate *Hh* paralogs coding sequences.** Representation of transitions (s) and transversions (v) accumulated in the three codon positions versus the genetic distance retrieved by the nucleotide substitutions model, GTR.

**Table 2.** Test of substitution saturation by Xia et al. using DAMBE [209, 210]. Analysis performed on fully resolved sites only, testing whether the observed Iss is significantly lower than Iss.c.. IssSym is Iss.c. assuming a symmetrical topology; IssAsym is Iss.c. assuming an asymmetrical topology.

| *NumOTU* | *Iss* | *Iss.cSym* | *T* | *df* | *p-value* | *Iss.cAsym* | *T* | *df* | *p-value* |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 0,577 | 0,797 | 8,475 | 515,000 | 0,0000 | 0,763 | 7,172 | 515,000 | 0,0000 |
| 8 | 0,600 | 0,753 | 5,529 | 515,000 | 0,0000 | 0,642 | 1,509 | 515,000 | 0,1320 |
| 16 | 0,602 | 0,723 | 4,290 | 515,000 | 0,0000 | 0,514 | 3,149 | 515,000 | 0,0017 |
| 32 | 0,610 | 0,704 | 3,326 | 515,000 | 0,0009 | 0,378 | 8,189 | 515,000 | 0,0000 |

### 4.2.1. Selective Constraints at The Codon-Level After Duplication

To test for different evolutionary rates upon duplication, we started by accessing positive selection on post-duplication branches, using the branch models implemented in PAML v4.03 [217]. The likelihood ratio test (LRT) between the alternate and null model likelihoods shows that only for the *Dhh* branch the two-ratio model fits the data better (Table 3), meaning that this branch is evolving at a different rate than the other three. Therefore, the two-ratio model for the *Dhh* branch was further compared with a constrained two-ratio model, in order to check for the prevalence of positive selection [173]. In this case, the null hypothesis was not rejected, favoring this branch not to be under positive selection but under relaxed selective constraints (Table 3). However, this lineage-base analysis assumes that all amino acid sites are under the same selective pressure and, as many sites can be evolving at a different rate, it is a very conservative test of adaptive evolution [218]. Thus, we used the site models implemented in PAML in order to detect signatures of adaptive evolution over the *Dhh, Ihh* and *Shh* coding sequences (Table 4) and, as a result, the Dhh, Shh and Ihh proteins had a ω value of 0.114, 0.080 and 0.058, with no positively selected residues. In all cases, the M3 and M7 nested models were accepted, which mean that each codon on *Dhh, Ihh* and *Shh* sequences are under variable selective pressures but do not show evidences of positive selection.

**Table 3.** Likelihood parameter estimates under lineage-specific model of post-duplication branches of Hh vertebrate paralogs, branch calculated with PAML v4.3 [217].

| | ***Model*** | ***$\omega_0$*** | ***$\omega_1$*** | ***Lnl*** | ***Models compared*** | ***LRT (2Δl)*** | ***p-value*** | ***df*** |
|---|---|---|---|---|---|---|---|---|
| **A** | One-ratio (M0) | 0.0610 | NA | -29435.14 | | | | |
| **B** | Dhh two-ratio (unconstrained) | 0.0610 | 999 | -29432.95 | A and B | 4.38 | 0.04 | 1 |
| **C** | Dhh two-ratio (constrained) | 0.0610 | 1 | -29433.13 | C and B | 0.36 | 0.55 | 1 |
| **D** | Ihh/Shh two-ratio | 0.0612 | 921 | -29434.62 | A and D | 1.05 | 0.31 | 1 |
| **E** | Ihh two-ratio | 0.0610 | 0.8302 | -29434.96 | A and E | 0.37 | 0.54 | 1 |
| **F** | Shh two-ratio | 0.0610 | 0.1115 | -29434.85 | A and F | 0.59 | 0.44 | 1 |

The detection of positively selected residues in PAML v4.03 [217] is accessed by a Bayes Empirical Bayes (BEB) [219] analysis. However, this method does not detect negatively selected residues and PAML is not able to access purifying selection [217, 219]. As our data do not show evidences of positive selection, we used the Single Likelihood Ancestor Counting (SLAC) and the Fixed Effects Likelihood (FEL) methods [220], implemented in the Datamonkey web server [221, 222], to test for evidences of purifying selection and to detect which residues are responsible for these evidences. In agreement with our previous results, no evidences of positive selection was found for the three *Hh* paralogs with both Single Likelihood Ancestor Counting (SLAC) and Fixed Effects Likelihood (FEL) methods [220] (Table S2). With a significance threshold of 0.05 ($\mathrm{P} < 0.05$), the SLAC method showed, for the Dhh, Ihh and Shh proteins, 28%, 39% and 38% negatively selected residues and none residue under positive selection. The FEL method, being less conservative and more powerful than SLAC [221], detected 45%, 54% and 55% negatively selected residues for each paralog, and none under positive selection. However, with both methods, there were found codons with $dN/dS > 1$ (Fig 16 and Table S3), but statistically they were not significant. When we used a significance threshold of 0.10, these codons were also not detected as positively selected and, as expected, the number of negatively selected codons increased (Table S2). Analyzing the $dN/dS$ values distribution over the *Hh* codon sequences, those codons with stronger purifying signatures codifiy mainly for residues located over the Hedge/signaling domain, with a clear definition between each of the main four regions found on Hh proteins (Fig. 16). As expected from the overall $\omega$ values for each paralog, the $dN/dS$ values for each codon are lower for Shh and higher for Dhh.

**Table 4.** Likelihood parameter estimates under site-specific models of Hh vertebrate paralogs, branch calculated with PAML v4.3 [217]. An asterisk (*) marks the accepted model.

| *Gene* | *Model* | *Parameters* | *Lnl* | *Models Compared* | *LRT (2Δl)* | *p-value* | *df* |
|---|---|---|---|---|---|---|---|
| *Dhh* | M0 | ω0 = 0.08479 | -7360.011 | M0 *vs* M3* | 190.185 | 0.000 | 4 |
| | M3 | ω0 = 0.00492 ω1 = 0.08337<br>ω2 = 0.31612<br>p0 = 0.39785 p1 = 0.34084 p2 = 0.26131 | -7162.160 | | | | |
| | M7 | p = 0.41513 q = 3.11107<br>ω = 0.114 | -7163.063 | M7* *vs* M8 | 0.006 | 0.997 | 2 |
| | M8 | p0 = 0.99999 p =0.41486<br>q = 3.10767<br>(p1 = 0.00001) ω1 = 2.90725 | -7163.066 | | | | |
| *Ihh* | M0 | ω0 = 0.06236 | -8206.454 | M0 *vs* M3* | 740.516 | 0.000 | 4 |
| | M3 | ω0 = 0.00520 ω1 = 0.11825<br>ω2 = 0.34903<br>p0 = 0.59836 p1 = 0.27289 p2 = 0.12874 | -7836.196 | | | | |
| | M7 | p = 0.25686 q = 2.77341<br>ω = 0.080 | -7838.824 | M7* *vs* M8 | 0.002 | 0.999 | 2 |
| | M8 | p0 = 0.99999 p =0.25687 q = 2.77361<br>(p1 = 0.00001) ω1 = 1.00000 | -7838.825 | | | | |
| *Shh* | M0 | ω0 = 0.04658 | -7648.861 | M0 *vs* M3* | 612.130 | 0.000 | 4 |
| | M3 | ω0 = 0.00188 ω1 = 0.06952<br>ω2 = 0.26493<br>p0 = 0.60204 p1 = 0.25287 p2 = 0.14509 | -7342.796 | | | | |
| | M7 | p = 0.20836 q = 3.15101<br>ω = 0.058 | -7342.669 | M7* *vs* M8 | 0.007 | 0.996 | 2 |
| | M8 | p0 = 0.99999 p = 0.20836 q = 3.15102<br>(p1 = 0.00001) ω1 = 5.18630 | -7342.672 | | | | |

### 4.2.2. Selective Constraints at The Amino Acid-Level After Duplication

Some concerns have been raised over the trustworthiness of the site selection models that use $\omega$ ratios to detect subtle molecular adaptations, as they can fail to detect positively and negatively selected sites that can be evolving under biochemical constrains [179, 180, 223, 224]. Therefore, we used TreeSAAP [225] to detect evidences of positive and negative selection over destabilizing substitutions in order to infer about the biochemical forces acting on the evolution and diversification of vertebrate *Hh* proteins. We started by assessing which destabilizing properties are under negative and positive selection in each of the three vertebrate *Hh* paralogs and we found that 24 out of 31 biochemical properties are under negative selection, from which six are under strong purifying selection (Fig. 17). These negatively selected properties are both classified as chemical or structural properties, highlighting the importance of the chemical and structural features of *Hh* proteins for the correct activation of the *Hh* signaling pathway. Interestingly, one property was found to be under strong positive selection in all paralogs:

the amino acid isoeletric point, a chemical property that correlates with the pH at which the amino acid surface carries no net electrical charge; suggesting that it may provide adaptive features to the vertebrate *Hh* paralogs. A sliding window analysis for this property Z-scores (Fig. 18) shows that it is strongly positively selected ($\mathrm{P} < 0.001$) on the Hog domain but also in a smaller extent on the Hedge domain ($\mathrm{P} < 0.05$).

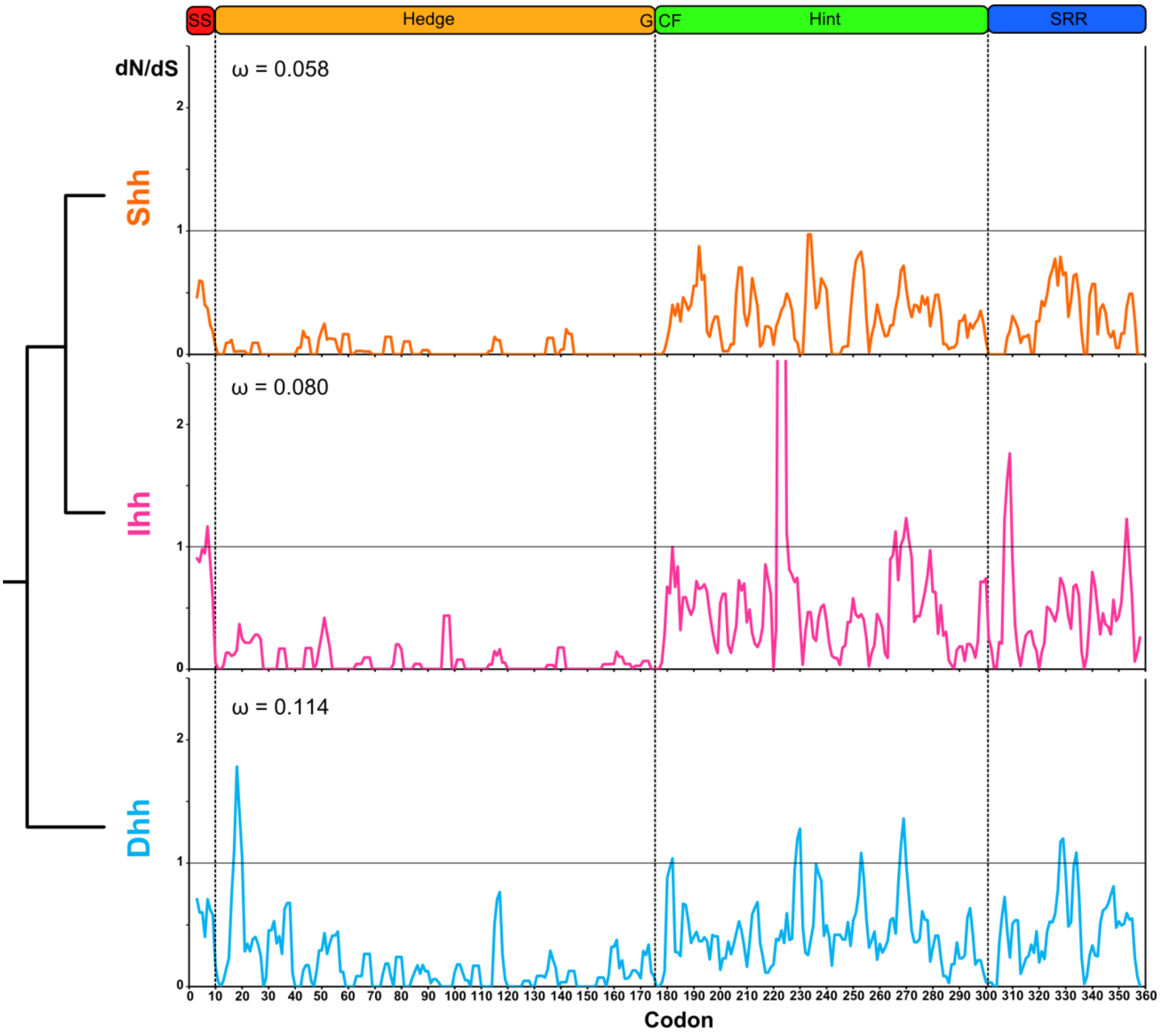
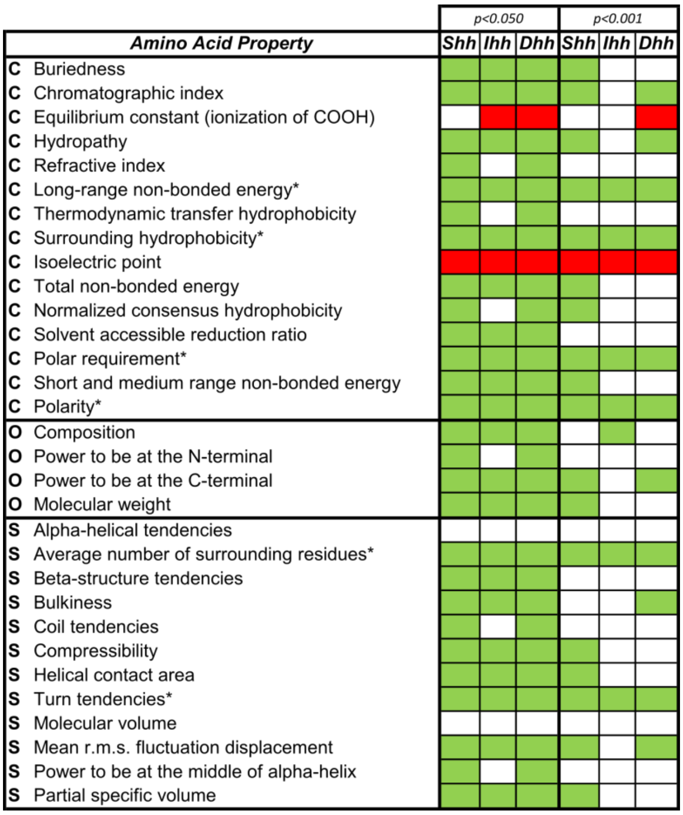


**Figure 16. Differences in the selection pattern of the three vertebrate *Hh* paralogs.** Sliding window analysis of the $dN/dS$ ratio applying the SLAC and FEL methods [220] for the three vertebrate *Hh* paralogs, represented as a mobile mean with a period of 3. The phylogenetic relationship between each group and the mean omega value ($\omega_0$) for each branch calculated with PAML v4.3 [217] are shown. The Hh proteins domains are displayed as annotated for the Hh, Dhh, Ihh and Shh proteins on the GenBank [138] and UniProt [250] databases.

The amino acid isoelectric point is positively selected in all vertebrate *Hh* paralogs, but within different regions for each paralog (Fig. 18). Over the Hedge domain, a region comprising the 33 and 55 alignment codon positions is under positive selection ($\mathrm{P} < 0.05$) for the three paralogs. However, only for the Ihh group the region within positions 33 and 44 is positively selected while the region comprised by the 44 and 55 positions is under positive selection only for the *Shh* and *Dhh* paralogs. In addition, two other regions over the Hedge domain are under positive selection for the amino acid isoelectric point property but only for the *Dhh* paralog: one within positions 62 and 84 and other within 126 and 142. Regarding the Hog domain, 7 regions are found under positive selection for this property: 5 on the Hint module and 2 on the SRR. From these 7 regions, only two over the Hint module are not common to the three paralogs: a region between positions 240 and 260 is common to Shh and Dhh and a region within positions 285 and 296 is unique to Shh.



| | | p<0.050 | | | p<0.001 | | |
|---|---|---|---|---|---|---|---|
| | ***Amino Acid Property*** | ***Shh*** | ***Ihh*** | ***Dhh*** | ***Shh*** | ***Ihh*** | ***Dhh*** |
| **C** | Buriedness | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **C** | Chromatographic index | 🟩 | 🟩 | 🟩 | 🟩 | | 🟩 |
| **C** | Equilibrium constant (ionization of COOH) | | 🟥 | 🟥 | | | 🟥 |
| **C** | Hydropathy | 🟩 | 🟩 | 🟩 | 🟩 | | 🟩 |
| **C** | Refractive index | 🟩 | | 🟩 | | | |
| **C** | Long-range non-bonded energy* | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| **C** | Thermodynamic transfer hydrophobicity | 🟩 | | 🟩 | | | |
| **C** | Surrounding hydrophobicity* | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| **C** | Isoelectric point | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| **C** | Total non-bonded energy | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **C** | Normalized consensus hydrophobicity | 🟩 | | 🟩 | 🟩 | | |
| **C** | Solvent accessible reduction ratio | 🟩 | 🟩 | 🟩 | | | |
| **C** | Polar requirement* | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| **C** | Short and medium range non-bonded energy | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **C** | Polarity* | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| **O** | Composition | 🟩 | 🟩 | 🟩 | | 🟩 | |
| **O** | Power to be at the N-terminal | 🟩 | | 🟩 | | | |
| **O** | Power to be at the C-terminal | 🟩 | 🟩 | 🟩 | 🟩 | | 🟩 |
| **O** | Molecular weight | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **S** | Alpha-helical tendencies | | | | | | |
| **S** | Average number of surrounding residues* | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| **S** | Beta-structure tendencies | 🟩 | 🟩 | 🟩 | | | |
| **S** | Bulkiness | 🟩 | 🟩 | 🟩 | | | 🟩 |
| **S** | Coil tendencies | 🟩 | | 🟩 | | | |
| **S** | Compressibility | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **S** | Helical contact area | 🟩 | 🟩 | 🟩 | 🟩 | | |
| **S** | Turn tendencies* | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| **S** | Molecular volume | | | | | | |
| **S** | Mean r.m.s. fluctuation displacement | 🟩 | 🟩 | 🟩 | 🟩 | | 🟩 |
| **S** | Power to be at the middle of alpha-helix | 🟩 | | 🟩 | | | |
| **S** | Partial specific volume | 🟩 | 🟩 | 🟩 | 🟩 | | |

**Figure 17. Amino acid properties under positive (red) and negative (green) selection in vertebrate *Hh* coding sequences.** Two different significance levels are shown: $P \leq 0.05$ (Z-score > |1.64|) to detect significant selective signatures and $P \leq 0.001$ (Z-score > |3.09|) to detect strong selective signatures. Amino acid properties are classified as chemical (C), structural (S) and other (O), according to da Fonseca *et al.* [223].

Different patterns of amino acid properties selection are also observed within paralogs at different significance thresholds (Fig. 17). At a significance of 0.05, the same number of negatively selected properties is found for both *Shh* and *Dhh* paralogs, but a reduced number is found for *Ihh*. In addition, the amino acid equilibrium constant (ionization of COOH) is found to be positively selected, but only on the *Ihh* and *Dhh* paralogs. When we reduce the significance threshold to a value of 0.001, we find that, despite the common properties described above, other properties are under strong negative selection within different paralogs. As expected from the codon-level analysis, a higher number of amino acid properties are under strong negative selection for the Shh proteins. It was not expected to find a higher number of negatively selected properties for Dhh than Ihh, as the latest show stronger purifying signatures at the codon level. However, a second property is found under strong positive selection for Dhh, which is not observed for the other two paralogs. Despite these differences, the majority of these strongly negatively selected properties comprise chemical properties.

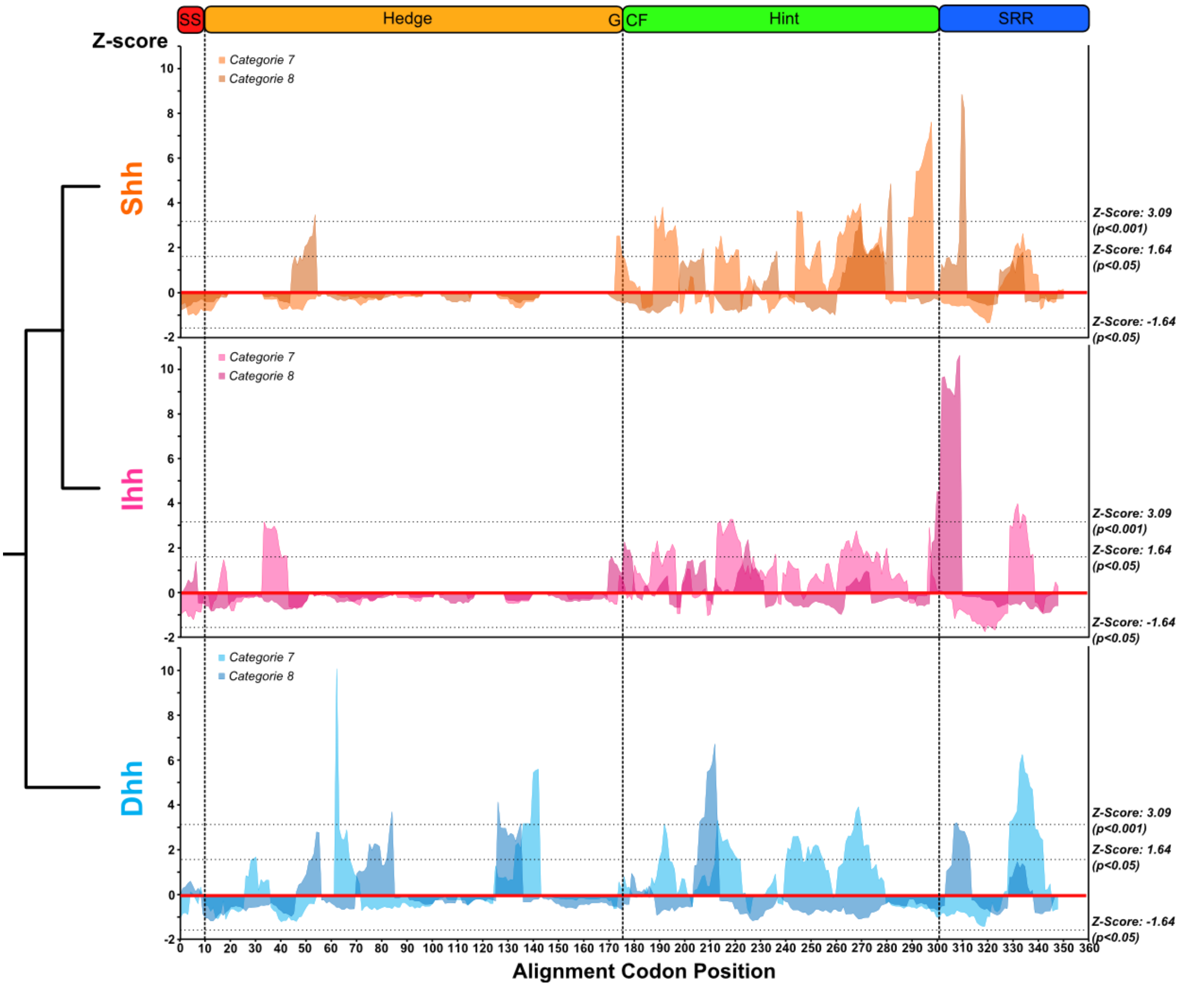


**Figure 18. Differences on the amino acid isoelectric point property selection pattern for the three vertebrate *Hh* paralogs.** Sliding window analysis for the Z-scores calculated for categories 7 and 8 using TreeSAAP [225] for the three vertebrate *Hh* paralogs, showing the phylogenetic relationship between each group.

At the amino acid level, we found 20 strongly positively selected $(P < 0.001)$ positions in the Shh group for at least one amino acid property, while for Ihh and Dhh there were found 27 and 32 sites, respectively (Table S4). The majority of these are located on the Hog domain, as expected and a great number of them comprise sites that at a codon level showed $\omega > 1$ values but were not statistically detected as under positive selection (Table S3 and Table S4). However, some of the positively selected sites detected by TreeSAAP were previously identified as under negative selection with FEL (Table S2 and Table S4). Analyzing the codon and amino acid alignment on these positions, these correspond to variable sites, with a great rate of non-synonymous substitutions, surrounded by highly conserved positions, suggesting that FEL overestimated the dS value for these positions due to their highly negatively selected environment. When we apply the empirical threshold of at least three properties with at least three properties showing signatures of positive selection, the number of positively selected residues decreased to 8 in Dhh, 3 in Ihh and 1 in Shh. These are only found on the Hog domain and none of them was previously identified as under negative selection at the codon level but 3 on Dhh and 1 in Ihh showed ω values above 1 (Table S4). Interestingly, the Shh residue 385 (numbered according to the *Homo sapiens* sequence) showed positive selection in 7 amino acid properties and corresponds to the only positively selected residue identified over this paralog. None residue was found as equivalent to this one over the alignment of the Dhh and Ihh paralogs, but positions surrounding this position share the same signatures of positive selection within both paralogs: residue 358 for Dhh and residues 372 and 373 for Ihh (Table S4). In addition, despite being located apart from these residues on the Dhh protein sequence, the Dhh residue 396 also shows 7 amino acid properties under positive selection.

### 4.2.3. Functional Divergence of Hh proteins After Duplication

After gene duplication, one gene copy maintains the original function while the other copies accumulate changes toward functional diversification and different *Hh* paralogs show different selection signatures. We tested our data for the prevalence of Type I and Type II functional divergence, using DIVERGE 2.0 [184-186, 227]. Only the results for Type I functional divergence are presented (Type II divergence was not statistically significant). Given the topology presented on figure 14, the ML estimate of the coefficient ($\hat{\theta}_I$) of Type I functional divergence between Shh and Ihh, Shh and Dhh and Ihh and Dhh was 0.17±0.05, 0.28±0.05 and 0.20±0.05 (Fig. 14 and Table 5). In all cases, the null hypothesis was rejected, meaning that the three *Hh* paralogs showed signatures of Type I functional divergence.

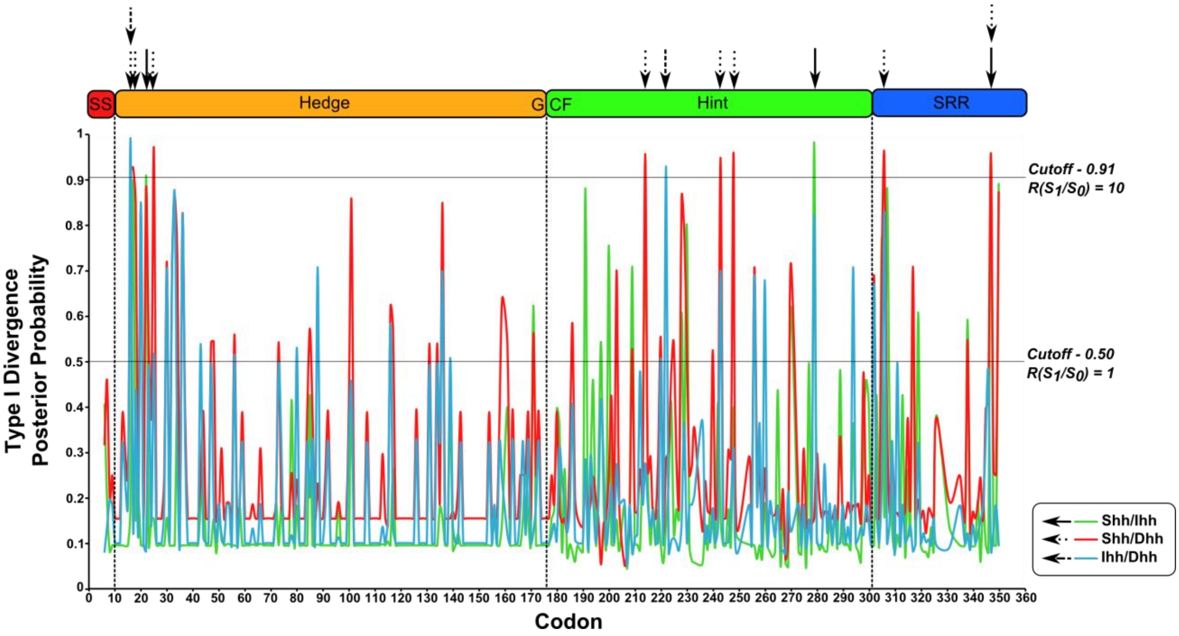


**Figure 19. Type I functional divergence over the vertebrate Hh paralogs.** Posterior probability for predicting critical amino acid residues for the functional divergence between the three vertebrate members of the Hh family. The arrows point to the residues with P($S_1$|$S_0$) > 0.91 and their position on the Hh proteins primary structure.

**Table 5.** Estimates of the coefficient of functional divergence Type I ($\hat{\theta}_I$) calculated with DIVERGE 2.0 [227] for each pair of vertebrate Hh paralog proteins.

| | *Shh/Ihh* | *Shh/Dhh* | *Ihh/Dhh* |
|---|---|---|---|
| **AlphaML** | 0.324153 | 0.526252 | 0.520681 |
| **ThetaML** | 0.172818 | 0.276325 | 0.197076 |
| **SE Theta** | 0.045282 | 0.054251 | 0.051544 |
| **LRT Theta** | 17.592.342 | 30.154.471 | 15.880.700 |

The site-specific profile based on the posterior analysis for scoring amino acid residues that are likely to be involved in Type I functional divergence between vertebrate *Hh* paralogs is presented on figure 19, and it shows that the higher posterior probabilities are found within the Hog domain. Between Shh and Ihh, 18 out of 358 sites are above a posterior probability of 0.5, while this number rises to 46 for Shh/Dhh and to 19 Ihh/Dhh (Fig. 19). Using the cutoff value of 0.91 (corresponding to a posterior odd ratio $R(S_1|S_0) >$ 10), we identified 3 sites that significantly counts for the type I functional divergence between Shh and Ihh, 8 between Shh and Dhh and 2 between Ihh and Dhh (Table 6 and Fig. 19). These predicted functional sites are not equally distributed throughout the respective protein, but clustered on the N-terminal region of the Hedge domain and on the Hint and SRR regions. Interestingly, different clusters are found for different pairs of paralogs (Fig. 19) and those over the Hog domain fall in regions positively selected for the amino acid isoelectric point property (Fig. 18 and Fig. 19).

**Table 6.** Amino acid residues with a Type I functional divergence posterior probability P(S1|X)≥0.91 for each pair of vertebrate Hh paralog proteins, calculated with DIVERGE 2.0 [227]. The site position in each alignment is listed, as well the correspondent position on the human protein sequence. Homo sapiens First, refers to the first member of the pair, and Homo sapiens Second, to the second member of the pair. An asterisk (*) marks negatively selected residues presented on Table S2.

| *Pairs* | *Position on the Alignment* | *Homo sapiens First* | *Homo sapiens Second* | *Posterior Probability* |
|---|---|---|---|---|
| *Ihh/Shh* | 279 | 324 | 340* | 0.98 |
| | 347 | 396 | 445 | 0.96 |
| | 22 | 45 | 40* | 0.91 |
| *Dhh/Shh* | 25 | 44 | 43* | 0.97 |
| | 248 | 275 | 274* | 0.96 |
| | 347 | 394 | 445 | 0.96 |
| | 214 | 238 | 237 | 0.96 |
| | 306 | 347 | 367* | 0.96 |
| | 243 | 270 | 269* | 0.95 |
| | 17 | 36 | 35 | 0.93 |
| | 16 | 35 | 34* | 0.92 |
| *Ihh/Dhh* | 16 | 39* | 35 | 0.99 |
| | 222 | 250 | 246* | 0.93 |

We observed that the 3 sites that are responsible for Type I functional divergence between Shh and Ihh, with a posterior probability above 0.91, are highly conserved in Shh proteins but highly variable in Ihh proteins. The same is observed for the Shh/Dhh pair but not for the Ihh/Dhh pair (Fig. 20). In addition, 2 of the type I functionally divergent sites between Shh and Ihh and 5 of the found between Shh and Dhh correspond to negatively selected residues (Table S4 and Table 6). Thus, these residues count for the functional divergence of Shh proteins relatively to the other two paralogs.

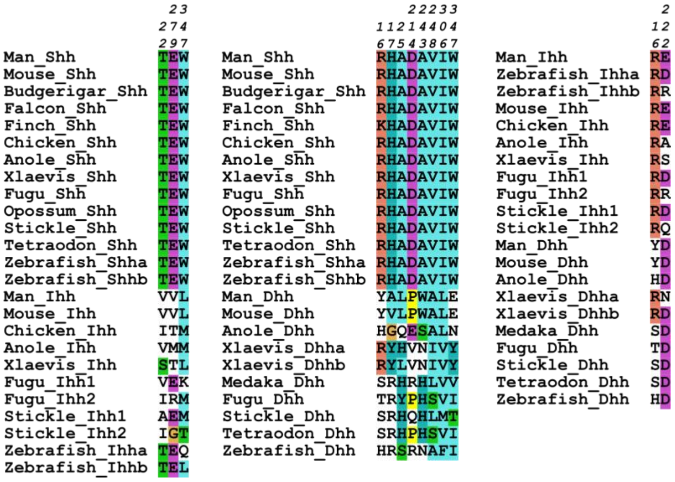


**Figure 20. Amino acid configurations of the sites with a Type I functional divergence posterior probability P($S_1$|X) ≥ 0.91 for each pair of vertebrate Hh paralog proteins.** The sites which are responsible for type I functional divergence between Shh and Ihh and Shh and Dhh are highly conserved in Shh proteins but highly variable in Ihh and Dhh proteins. However, within the Ihh and Dhh pair, this is not observed. These residues must count for the functional divergence of Shh proteins relatively to the other two paralogs.

### 4.2.4. Structural Analysis of Selected Domains

To further relate the spatial position of the selected regions and divergent sites on the tridimensional structure of the three vertebrate Hh paralog proteins, it was necessary to assess them to the tridimensional structure of the Shh, Ihh and Dhh proteins. We started by mapping the negatively selected sites identified at the codon level ($\mathrm{P} < 0.05$) on the tridimensional structure of the Hedge domain for each paralog and we observed that those are found both on the interior and on the surface of the HhN peptide (Fig. 21), suggesting strong constraints acting in order to keep the tridimensional structure of the signaling peptide unchanged. As the number of negatively selected sites decrease (from Shh to Dhh, passing trough Ihh), we observed that this reduction occurs in sites that are exposed on the peptide surface, keeping the interior of the peptide and the ion binding sites highly negatively selected.

Mapping the regions identified as under positive selection for the amino acid isoelectric point property on the Hedge peptide, we observed that they are located on specific regions on the surface of the signaling peptides, specific for each paralog (Fig. 21). On the case of the Shh and Ihh signaling peptides, the two regions identified comprise a large surface loop that defines a surface area of the Hedge signaling peptides close to the binding site, but both define different parts of this loop, forming two distinct regions that rise on highly negatively selected areas and may provide different adaptive features to these two lineages. On the case of the Dhh signaling peptide, the same loop is under positive selection, grouping the two regions that define Shh and Ihh, but we also observed that the other two areas previously identified as under positive selection for this property (Fig. 18), despite being apart on the primary structure of the Dhh signaling peptide, are folded in order to expand the region formed by the positively selected loop around the binding site (Fig. 21). Interestingly, the sites identified under positive selection for at least one amino acid property are located away from this region.

In regard to the Hog domain, its tridimensional structure was predicted for each of the vertebrate Hh paralogs using as template the *Drosophila melanogaster* HhC peptide. Due to the high divergence found between vertebrate Hog sequences, the three models obtained are similar but not superimpose (RMSD: ShhC/DhhC – 2.14 Å; ShhC/IhhC – 2.20Å; IhhC/DhhC – 1.70Å). Interestingly, the catalytic site is located within a deep pocket on the peptide interior in all cases (Fig. 21). When the negatively selected sites identified at the codon level were mapped on these models, we found that they comprise residues that are most probably located on the interior of the Hog domain and are all arranged around the catalytic site (Fig. 21). Mapping the positively selected regions for the amino acid isoelectric point property we detected that they are mainly located on the surface of the Hog domain, as well as those residues detected as positively selected for at least one amino acid property (Fig. 21). We further mapped the position of Dhh residues 358 and 396, Ihh residues 372 and 373 and Shh residue 385 on the tridimensional structure of the Hog domain (Fig. 21) and we found that they are located on the surface around the catalytic site but not on the same spatial organization between different paralogs. These results suggest that, inversely to what is observed for the Hedge domain, purifying constrains only act over the Hog domain in order to maintain the catalytic site intact, allowing the tridimensional structure of this domain to change under relaxed chemical and structural constrains.

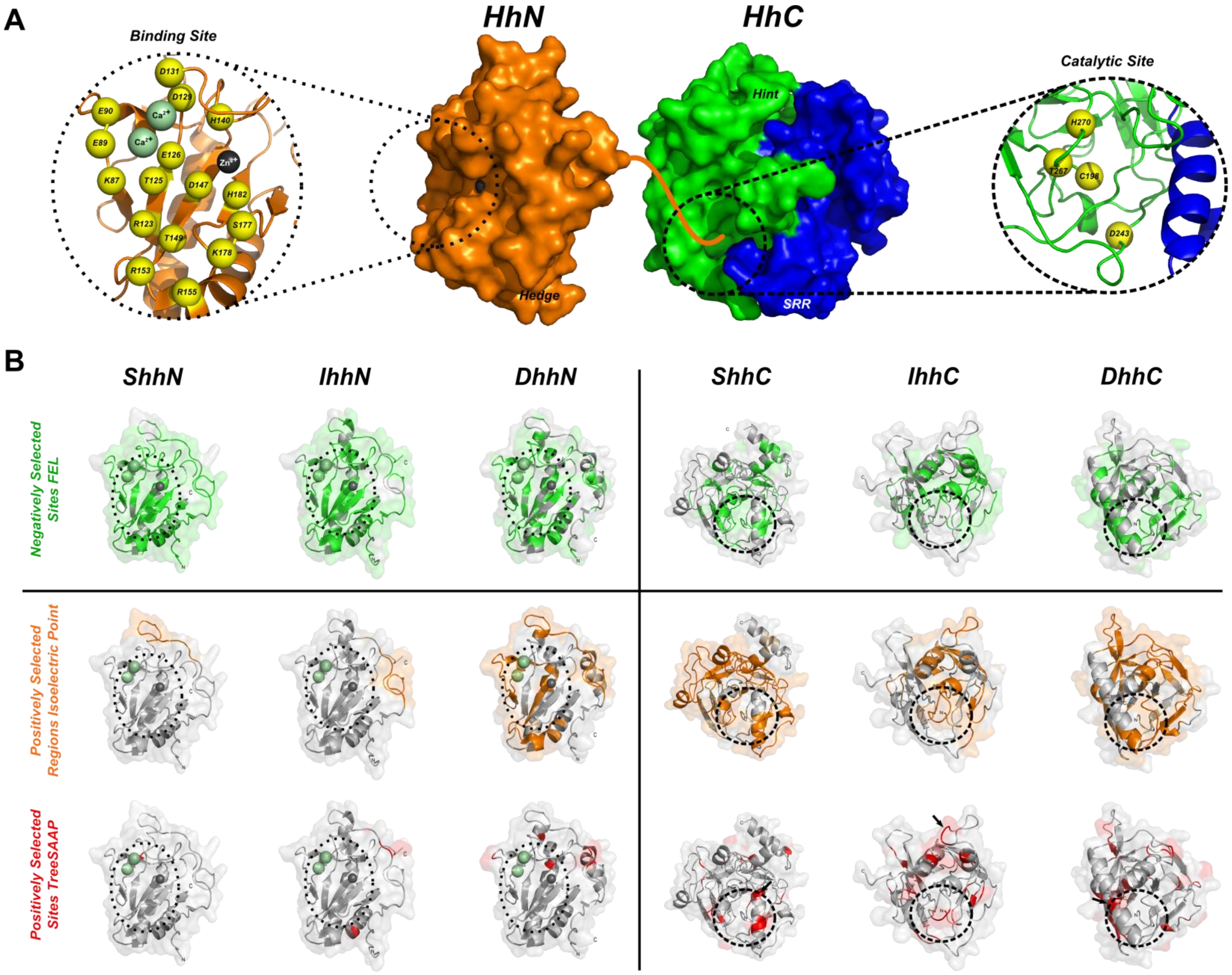


**Figure 21. Tridimensional arrangement of negatively and positively regions over the Hedgehog proteins.** (a) Tridimensional representation of Hedge (PDB: 3HO5) and Hog (Shh modelled by homology) domains, coloured according to key identified regions. A straight orange line denotes how both domains may be linked in the pro-protein. Key residues important for binding and forming the catalytic site are represented in yellow spheres, numbered according to the human Shh protein [28, 29, 250]. (b) Tridimensional arrangement of negatively and positively regions over the Hedge (HhN) and Hog (HhC) domains. Protein represented in grey cartoon with transparent surface. Negatively selected sites (green) identified with FEL, positively selected regions for the amino acid isoelectric point property (orange) and positively selected sites (red) identified with TreeSAAP are shown for each paralog domain. Arrows marks those residues surrounding the 324 codon alignment position. A dashed circle denotes the position of the zinc/calcium binding site and the catalytic site.

# 5. DISCUSSION

The three vertebrate members of the *Hh* family are due to two rounds of genome duplications in the vertebrate ancestor [137] and our synteny analyses suggested that the synteny of each of the three vertebrate *Hh* paralogs is very conserved and must have evolved independently after duplications. This result is in agreement with the recent finding of an ancestral linkage group shared between the amphioux *Hh* and mouse *Hh* genes [251]. Therefore, we hypothesize that before the first round of whole genome duplication, the ancestral *Hh* synteny was composed by a conserved cluster of genes, encompassing at least the ancestral *LMBR, Hh*, *RHEBL* and *MLL* genes. These genes were present on the ancestral of vertebrates in this same order but separated by a great number of genes, forming a high dimension cluster. After the first duplication, this cluster duplicated to form two clusters: one containing the ancestor of *Shh* and *Ihh* and other containing the ancestor of *Dhh*. Probably before the second round of whole genome duplication, these two clusters evolved independently, suffering rearrangements that reduced the size of both clusters but retained the ancestral duplicated *LMBR, Hh*, *RHEBL* and *MLL* genes. After the second round of duplication, a total of four duplicated clusters were produced, each containing a duplicated version of the ancestral conserved cluster. Further rearrangements before the emergence of vertebrates may lead to the loss of a duplicated *Dhh* gene and to the creation of the currently observed synteny for each of the vertebrate *Hh* paralogs, which remained conserved until today on the tetrapod lineage but suffered further arrangements at least on the teleostei lineage.

As evidences of these clusters are present on the genome of all the studied vertebrate genomes, including jawless fishes, this data supports that the vertebrate members of the *Hh* gene family arose 600-500 mya as suggested by Kumar *et al.* [137]. Despite it was not possible to study synteny of the *LMBR, Hh*, *RHEBL* and *MLL* genes in the studied jawless fishes available to date, the presence of the LMBR1L, RHEBL1 and MLL2 on these genomes supports the hypothesis of *Dhh* gene loss on the cyclostome lineage suggested by Kano *et al.* [145] and is in agreement with the hypothesis that cyclostomes diverged after the two whole genome duplications characteristic of vertebrates [252]. When we analyzed the teleostei lineage, a different but conserved pattern is found within the two main classes analyzed (ostariophysi and euteleost fishes), which suggests that further rearrangements occurred on this lineage before the third whole genome duplication specific of the teleost lineage. Only in tetrapods the three conserved syntenic clusters seem to have not suffered further rearrangements, assuming the same gene composition and order in all classes.

Birds could be an exception to the previous statement as evidences of genes encompassing the *Dhh* conserved synteny lack from some avian sequenced genomes. However, our comparative genomics analyses using the lizard physical position of this conserved cluster showed that regions neighboring these genes are found on several galliform and neoave genomes and randomly assigned into the Un_Random chromosome of those avian genomes whose karyotype is already mapped. In fact, the avian karyotype is composed of seven to nine pairs of macrochromosmes and 30 to 32 pair of microchromosomes [253]. Microchromosomes are very small chromosomes that range in size between 3.5 and 23 Mb [254], remarkably gene rich, have a high recombination rate and present a high content on CpG islands [253], which make them difficult to be cloned, sequenced and identified by cytogenetic approaches. Thus, many microchromosomes still remain absent from the current avian assemblies and the contigs that could not be assigned to a chromosome are arranged in a virtual chromosome, the Un_Random chromosome [255, 256]. It was previously demonstrated that those sequences can be assigned to small microchromosomes but also that many chicken cDNA and EST sequences are absent from the current assembly, including from the Un_Random fraction, suggesting the total absence of large amounts of the corresponding DNA sequences on the chicken genome assembly [257]. Therefore, our hypothesis is that this conserved cluster may be present on avian genomes, both galliform and neoave, probably physically located on a microchromosome. In the case of the *Gallus gallus* cluster, one good possibility could be the microchromosome 21, as Trukhina and Smirnov [258] shown that microsatellites from the linkage group E50C23 are located on this chicken microchromosome. However, when we experimentally search for a partial Dhh-coding sequence on *Gallus sp.* genome, we were not able to retrieve any conclusive result, probably due to the quality of the primer sets used and to the high GC content of the target sequences to amplify. Further experimental tests would be necessary to assess this hypothesis. However, a new draft of the *G. gallus* genome assembly was released during this work and evidences of a Dhh-coding and a LMBR1L-coding sequence were found, annotated over different Un_random chromosome contigs. This new data supports our hypothesis that the avian lineage carries a *Dhh* copy.

Our major findings in the context of the *Hh* gene family evolution, showed strong variable purifying selection and Type I functional divergence acting over the three vertebrate *Hh* branches. It is a fact that the three vertebrate *Hh* paralogous genes codify for highly conserved proteins (Fig. 14) involved in key developmental processes, yet, the two main domains that comprise these proteins are differently conserved: the Hedge domain is more conserved than the Hog domain. This suggests that these three proteins may act similarly on their physiological role, deactivating the transmembrane receptor Patched in a conserved manner, and that the differences found on their roles depends from their different expression patterns [5, 15]. However, *Shh* coding sequences seem to be more conserved than *Ihh* and *Dhh* and we showed that each of these vertebrate members are evolving at different selective rates.

At the codon-level, any evidence of positive selection was identified for the three paralogs but the *Dhh* branch seems to be evolving under more relaxed constraints than the other two duplicated branches. In fact, each vertebrate member of the *Hh* family shows very low ω values (around 0.1) (Fig. 16), with the *Shh* members showing the lower value and the *Dhh* members the higher value. This suggests that the *Shh* coding sequences are evolving at more purifying constraints than the other two vertebrate *Hh* paralogs. Actually, the Shh protein is responsible for more complex traits than the other two members (reviewed in [15]), being a central player in the development and patterning of the nervous and skeletal systems and the most broadly expressed *Hh* member. On the other hand, *Ihh* is specifically expressed in a narrowed number of tissues, mainly in the primitive endoderm and in prehypertrophic chondrocytes, and *Dhh* is confined to the gonads, has a role on the development of the peripheral nervous system and is always expressed in combination with *Ihh* [15]. Therefore, mutations affecting the fitness of the Shh protein should be more deleterious than those affecting Ihh, and the latest should be more deleterious than those over Dhh. It was previously reported that stronger purifying constraints act on the evolution of genes expressed early on the embryonic development process, as mutations occurring in genes expressed early in development will on average have more deleterious fitness consequences than mutations in genes expressed later [259, 260]. In agreement, *Shh* is in fact the first member of the vertebrate *Hh* family to be expressed during vertebrate embryonic development, followed by *Ihh* and *Dhh* [15].

An average percentage of 50% of the vertebrate *Hh* proteins residues are evolving under strong purifying selection and those most significantly negatively selected are located on the N-terminal domain, with a clear definition between the signaling peptide, the Hedge region, the Hint module and the SRR region. Interestingly, most of these residues were previously reported as disease-causing mutation-spots (Table S2). These results come in agreement with two previous works on the *Hh* gene family: Kumar *et al.* [137] showed that, comparing *Shh, Dhh* and *Drosophila Hh* genes, the Hedge domain coding region shows a lower evolutionary rate than the Hog coding region; and Gunbin *et al.* [261] showed that the invertebrate members of the *Hh* gene family also share the same pattern of selective signatures over these two main domains. However, these two works also suggested that positive selection occurs in the Hint coding region and our codon-level analyses, besides showing that different vertebrate *Hh* paralogs and their distinct regions show different evolutionary rates, did not identified any significantly positively selected residue over the three vertebrate *Hh* paralogs. On the other hand, our protein-level analysis came in agreement with these two previous works and added evidences of strong positive selection over the Hog domain but also of more relaxed positive selection over the Hedge domain, both only for the amino acid isoelectric point. As found at the codon-level, the Shh members show a smaller extent of positive selection and Dhh is the member with higher signatures of positive selection. We found that strong purifying selection is acting on the interior of the two domains that comprise these proteins, both at a chemical and structural level and within both domains, suggesting strong constraints occurring in order to keep the core role of the *Hh* proteins intact. However, these constraints are stronger over the Hedge domain rather than the Hog domain.

In fact, the signaling activity of the Hedge domain requires a highly conserved tridimensional structure in order to conservatively be recognized by its receptors, which is observed on the highly conserved structure of the HhN peptide among Hh paralogs (Fig. 21) [28] and explains why strong negative selection is found both on the interior and on the surface of the signaling peptide. We found that the zinc binding site, located on a conserved cleft that resembles the zinc hydrolases catalytic site [29] and is involved on protein interactions with receptor proteins (reviewed in [24]), is also under strong negative selection but surrounded by a surface region with signatures of positive selection for the amino acid isoelectric point property. This chemical property is associated with the pH value at which the amino acid surface carries no net electrical charge and is directly related with the surface charge of protein regions.

Comprising mainly surface loop regions, the positively selected areas may provide adaptive features to the signaling peptide, most probably on the interaction with different receptor proteins. As different areas of these regions show different patterns of positive selection within vertebrate Hh paralogs, with Shh and Ihh showing the smaller extent of positive selection, these regions must provide the ability of different *Hh* proteins to bind different protein receptors. The Dhh signaling domain shows evidences of more relaxed negative selection and a larger extent of positive selection surrounding the zinc binding domain, which was expected from the codon-level analysis and also from its narrowed activity on the development of gonads. Interestingly, it was previously reported that genes involved in reproduction and sexual differentiation show higher rates of divergence and positive selection than other genes in the genome, providing reproductive adaptations [262, 263] and we can hypothesize a link between the physiological signaling role of Dhh on gonadal development and its relaxed evolution, providing adaptive signaling features during the embryonic development of gonads.

On the case of the Hog domain, negative selection is acting only on the interior of the HhC peptide, mainly on the surroundings of the catalytic site, while positive selection is found on the peptide surface. The models built for the *Homo sapiens* sequences show that the tridimensional structure of this domain is highly variable among paralogs and our results suggests that this constraints act only to keep the catalytic site unchanged and within a pocket on the interior of the Hog domain, assuring that the catalytic activity of this domain is retained. It is this domain that provides most of the functional divergence observed between vertebrate *Hh* paralogs but, as the sites responsible for this feature are found within areas under positive selection on the surface of the peptide, this divergence may be responsible for their structural features and not for their chemical activity.

## 6. MAIN CONCLUSIONS

In this work, it was aimed to assess the evolutionary history of the *Hh* gene family in vertebrates into two levels – at the genomic and the coding-sequence levels – and we showed that:

- The synteny of the vertebrate members of the *Hh* gene family is highly conserved and had an ancestral origin;
- Our results support the hypothesis of Dhh gene loss on the cyclostomes lineage and are in agreement with the hypothesis that cyclostomes diverged after the two wide-genome duplications characteristic of the vertebrate lineage;
- Avian genomes must carry at least one exemplar of a Dhh-coding sequence, most probably over a microchromosome, which is highly difficult to detect and sequencing;
- Strong variable purifying selection and Type I functional divergence is acting over the three vertebrate *Hh* branches;
- At the amino acid level, positive selection is acting over both main Hh domains, mainly for the isoelectric point chemical property;
- For all vertebrate *Hh* paralogs, stronger purifying constraints are acting mainly over the Hedge/signaling domain, and positive selection is acting mainly over the Hog domain;
- The zinc and calcium binding sites are highly negatively selected and surrounded by a positively selected loop, which can act as an adaptive feature of Hh proteins and probably allow different protein-protein interactions;
- The *Dhh* branch seems to be evolving under more relaxed constraints than the other two duplicated branches, which may be related with its role on gonad embryonic development;
- The *Shh* branch seems to be evolving under more purifying selection constraints, which may be related with its role on the embryonic development of critical systems and tissues and its early expression pattern.

## 7. FUTURE PERSPECTIVES

The work presented has allowed a detailed characterization of the adaptive and genomic evolution of the *Hh* gene family in vertebrates. However, it has also raised some new questions. As an example, the study could be expanded to the study of *Hh* and *Hh*-related genes in invertebrates, both at the genomic, gene and protein levels. Comparing the synteny of *Hh*-related genes with the one found for *Hh* genes could bring new insights into the evolutionary origin of the bilaterian *Hh* gene. In addition, assessing selection for both domains separately could provide a better understanding of how adaptive selection is acting over these proteins. Another approach to better relate the evidences of adaptive evolution and functional divergence with the structural and functional roles of vertebrates *Hh* proteins would be also interesting. For example, the study of the mechanisms of sub-functionalization and selection acting over *Hh* regulatory sequences could allow us to assess functional divergence not at the protein level, but at the expression level. It would be also exciting to have access to fully resolved *Hh* proteins' tridimensional structures, for both the complete proteins and the vertebrate Hog domains. This could bring insights into the possible interactions formed between both domains and uncover the functional role of some residues found to be negatively and positively selected and responsible for Hh proteins functional divergence.

Also, in order to solve the experimental problems faced, the identification and sequencing of an avian Dhh-coding sequence could be assessed by different techniques, mainly searching over *Gallus gallus* genomic libraries with Dhh specific probes, cloning, western-blotting and mRNA purification and cDNA sequencing. The last approach, despite bringing evidences over the presence or absence of a Dhh-coding sequence, would also give insights into the expression patterns of this gene on avian species. However, blood cells would probably be not the best tissue to test, and it would require mainly avian gonadal tissues.

# 8. SCIENTIFIC FORMATION

Fortnight groupmeetings and workshop events promoted by the LEGE laboratory on different thematics and the acquired knowledge were relevant for the purposes in scope of this master:

“Gene Protein Evolution in Biotechnology” Workshop ─ ShareBiotech. (Pereira, S., Branco, R., Jorge, M., Santos, M.), supervised by A. Antunes, December 16 of 2011, Interdisciplinary Centre of Marine and Environmental Research, Porto, Portugal.

**Pereira, J.** Evolutionary Genomics and Adaptive Evolution of the Hedgehog Gene Family (*Shh, Ihh* and *Dhh*) in Vertebrates. Interdisciplinary Centre of Marine and Environmental Research (CIIMAR) ─ Laboratory of Ecotoxicology, Genomics and Evolution (LEGE), April 13 of 2012, Porto, Portugal. *(LEGE Groupmeeting)*

## 8.1. Publications and Communications

### 8.1.1. Papers

The results attained during this period resulted in:

**Pereira J**, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. (2012). Evolutionary Genomics and Adaptive Evolution of the Hedgehog Gene Family (*Shh, Ihh* and *Dhh*) in Vertebrates. (Submitted for publication)

### 8.1.2. Communications in Conferences

The results attained during this period were accepted in following communication panels:

**Pereira J**, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. Evolutionary Genomics of the Hedgehog Gene Family in Metazoans. IJUP’12: 5th Meeting of Young Researchers from the University of Porto, Porto, Portugal, February 23 of 2012 (Oral Presentation).

**Pereira J**, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. Evolutionary Genomics of the Hedgehog Gene Family in Metazoans: Identification of the Desert Hedgehog Gene on Avian Genomes. SMBE2012: Annual Meeting of the Society for Molecular Biology and Evolution, Dublin, Ireland, June 23-26 of 2012 (Poster Presentation).

## 9. REFERENCES

1. Shubin, N., *Your Inner Fish: A Journey Into The 3.5-Billion-Year History of The Human Body*. 2008, New York: Pantheon Books.
2. Zhang, J., Y.P. Zhang, and H.F. Rosenberg, *Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey.* Nat Genet, 2002. **30**(4): p. 411-5.
3. Antunes, A. and M.J. Ramos, *Gathering computational genomics and proteomics to unravel adaptive evolution.* Evol Bioinform Online, 2007. **3**: p. 207-9.
4. Nichols, S.A., et al., *Early evolution of animal cell signaling and adhesion genes.* Proc Natl Acad Sci U S A, 2006. **103**(33): p. 12451-6.
5. Ingham, P.W., Y. Nakano, and C. Seger, *Mechanisms and functions of Hedgehog signalling across the metazoa.* Nat Rev Genet, 2011. **12**(6): p. 393-406.
6. Ingham, P.W. and A.P. McMahon, *Hedgehog signaling in animal development: paradigms and principles.* Genes Dev, 2001. **15**(23): p. 3059-87.
7. Nüsslein-Volhard, C. and E. Wieschaus, *Mutations affecting segment number and polarity in Drosophila.* Nature, 1980. **287**: p. 795-1596.
8. Shimeld, S.M., *The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog.* Dev Genes Evol, 1999. **209**(1): p. 40-7.
9. Echelard, Y., et al., *Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity.* Cell, 1993. **75**(7): p. 1417-30.
10. Krauss, S., J.P. Concordet, and P.W. Ingham, *A functionally conserved homolog of the Drosophila segment polarity gene hh is expressed in tissues with polarizing activity in zebrafish embryos.* Cell, 1993. **75**(7): p. 1431-44.
11. Riddle, R.D., et al., *Sonic hedgehog mediates the polarizing activity of the ZPA.* Cell, 1993. **75**(7): p. 1401-16.
12. Chang, D.T., et al., *Products, genetic linkage and limb patterning activity of a murine hedgehog gene.* Development, 1994. **120**(11): p. 3339-53.
13. Roelink, H., et al., *Floor plate and motor neuron induction by vhh-1, a vertebrate homolog of hedgehog expressed by the notochord.* Cell, 1994. **76**(4): p. 761-75.
14. Hammerschmidt, M., A. Brook, and A.P. McMahon, *The world according to hedgehog.* Trends Genet, 1997. **13**(1): p. 14-21.
15. Varjosalo, M. and J. Taipale, *Hedgehog: functions and mechanisms.* 2008.
16. Nagase, T., et al., *Hedgehog signaling: a biophysical or biomechanical modulator in embryonic development?* Ann N Y Acad Sci, 2007. **1101**: p. 412-38.

17. Yao, H.H., W. Whoriskey, and B. Capel, *Desert Hedgehog/Patched 1 signaling specifies fetal Leydig cell fate in testis organogenesis.* Genes Dev, 2002. **16**(11): p. 1433-40.
18. Burglin, T.R., *The Hedgehog protein family.* Genome Biol, 2008. **9**(11): p. 241.
19. Lee, J.J., et al., *Autoproteolysis in hedgehog protein biogenesis.* Science, 1994. **266**(5190): p. 1528-37.
20. Porter, J.A., et al., *The product of hedgehog autoproteolytic cleavage active in local and long-range signalling.* Nature, 1995. **374**(6520): p. 363-6.
21. Beachy, P.A., et al., *Multiple roles of cholesterol in hedgehog protein biogenesis and signaling.* Cold Spring Harb Symp Quant Biol, 1997. **62**: p. 191-204.
22. Koonin, E.V., *A protein splice-junction motif in hedgehog family proteins.* Trends Biochem Sci, 1995. **20**(4): p. 141-2.
23. Hall, T.M., et al., *Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins.* Cell, 1997. **91**(1): p. 85-97.
24. Beachy, P., et al., *Interactions between Hedgehog proteins and their binding partners come into view.* Genes & development, 2010. **24**: p. 2001-2013.
25. Hall, T., et al., *A potential catalytic site revealed by the 1.7-A crystal structure of the amino-terminal signalling domain of Sonic hedgehog.* Nature, 1995. **378**: p. 212-218.
26. Day, E.S., et al., *Zinc-dependent structural stability of human Sonic hedgehog.* Biochemistry, 1999. **38**(45): p. 14868-80.
27. Fuse, N., et al., *Sonic hedgehog protein signals not as a hydrolytic enzyme but as an apparent ligand for patched.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**: p. 10992-11001.
28. McLellan, J., et al., *The mode of Hedgehog binding to Ihog homologues is not conserved across different phyla.* Nature, 2008. **455**: p. 979-1062.
29. Bosanac, I., et al., *The structure of SHH in complex with HHIP reveals a recognition role for the Shh pseudo active site in signaling.* Nature structural & molecular biology, 2009. **16**: p. 691-698.
30. McLellan, J.S., et al., *Structure of a heparin-dependent complex of Hedgehog and Ihog.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17208-13.
31. Chamoun, Z., et al., *Skinny hedgehog, an acyltransferase required for palmitoylation and activity of the hedgehog signal.* Science, 2001. **293**(5537): p. 2080-4.

32. Chen, M.-H., et al., *Palmitoylation is required for the production of a soluble multimeric Hedgehog protein complex and long-range signaling in vertebrates.* Genes & development, 2004. **18**: p. 641-700.

33. Peters, C., et al., *The cholesterol membrane anchor of the Hedgehog protein confers stable membrane association to lipid-modified proteins.* Proc Natl Acad Sci U S A, 2004. **101**(23): p. 8531-6.

34. Lewis, P.M., et al., *Cholesterol modification of sonic hedgehog is required for long-range signaling activity and effective modulation of signaling by Ptc1.* Cell, 2001. **105**(5): p. 599-612.

35. Gallet, A., et al., *Cholesterol modification is necessary for controlled planar long-range activity of Hedgehog in Drosophila epithelia.* Development (Cambridge, England), 2006. **133**: p. 407-425.

36. Zeng, X., et al., *A freely diffusible form of Sonic hedgehog mediates long-range signalling.* Nature, 2001. **411**(6838): p. 716-20.

37. Goetz, J.A., et al., *A highly conserved amino-terminal region of sonic hedgehog is required for the formation of its freely diffusible multimeric form.* J Biol Chem, 2006. **281**(7): p. 4087-93.

38. Eugster, C., et al., *Lipoprotein-heparan sulfate interactions in the Hh pathway.* Dev Cell, 2007. **13**(1): p. 57-71.

39. Ingham, P.W., *Transducing Hedgehog: the story so far.* EMBO J, 1998. **17**(13): p. 3505-11.

40. Yam, P.T., et al., *Sonic hedgehog guides axons through a noncanonical, Src-family-kinase-dependent signaling pathway.* Neuron, 2009. **62**(3): p. 349-62.

41. Chen, X., et al., *Processing and turnover of the Hedgehog protein in the endoplasmic reticulum.* J Cell Biol, 2011. **192**(5): p. 825-38.

42. Burke, R., et al., *Dispatched, a novel sterol-sensing domain protein dedicated to the release of cholesterol-modified hedgehog from signaling cells.* Cell, 1999. **99**: p. 803-818.

43. Ayers, K., et al., *The long-range activity of Hedgehog is regulated in the apical extracellular space by the glypican Dally and the hydrolase Notum.* Developmental cell, 2010. **18**: p. 605-625.

44. Panáková, D., et al., *Lipoprotein particles are required for Hedgehog and Wingless signalling.* Nature, 2005. **435**: p. 58-123.

45. Ingham, P.W., A.M. Taylor, and Y. Nakano, *Role of the Drosophila patched gene in positional signalling.* Nature, 1991. **353**(6340): p. 184-7.

46. Rohatgi, R., L. Milenkovic, and M.P. Scott, *Patched1 regulates hedgehog signaling at the primary cilium.* Science, 2007. **317**(5836): p. 372-6.

47. Hausmann, G., C. von Mering, and K. Basler, *The hedgehog signaling pathway: where did it come from?* PLoS Biol, 2009. **7**(6): p. e1000146.

48. Hooper, J. and M. Scott, *The Drosophila patched gene encodes a putative membrane protein required for segmental patterning.* Cell, 1989. **59**: p. 751-816.

49. Nakano, Y., et al., *A protein with several possible membrane-spanning domains encoded by the Drosophila segment polarity gene patched.* Nature, 1989. **341**: p. 508-521.

50. Marigo, V., et al., *Biochemical evidence that Patched is the Hedgehog receptor.* Nature, 1996. **384**(6605): p. 176-179.

51. Zheng, X., et al., *Genetic and biochemical definition of the Hedgehog receptor.* Genes Dev, 2010. **24**(1): p. 57-71.

52. Tenzen, T., et al., *The cell surface membrane proteins Cdo and Boc are components and targets of the Hedgehog signaling pathway and feedback network in mice.* Dev Cell, 2006. **10**(5): p. 647-56.

53. Filmus, J., M. Capurro, and J. Rast, *Glypicans.* Genome Biol, 2008. **9**(5): p. 224.

54. Lum, L., et al., *Identification of Hedgehog pathway components by RNAi in Drosophila cultured cells.* Science, 2003. **299**(5615): p. 2039-45.

55. Beckett, K., X. Franch-Marro, and J.-P. Vincent, *Glypican-mediated endocytosis of Hedgehog has opposite effects in flies and mice.* Trends in Cell Biology, 2008. **18**(8): p. 360-363.

56. Capurro, M.I., et al., *Glypican-3 inhibits Hedgehog signaling during development by competing with patched for Hedgehog binding.* Dev Cell, 2008. **14**(5): p. 700-11.

57. Allen, B.L., T. Tenzen, and A.P. McMahon, *The Hedgehog-binding proteins Gas1 and Cdo cooperate to positively regulate Shh signaling during mouse development.* Genes Dev, 2007. **21**(10): p. 1244-57.

58. Chuang, P.T. and A.P. McMahon, *Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein.* Nature, 1999. **397**(6720): p. 617-21.

59. Martinelli, D.C. and C.M. Fan, *Gas1 extends the range of Hedgehog action by facilitating its signaling.* Genes Dev, 2007. **21**(10): p. 1231-43.

60. McCarthy, R.A., et al., *Megalin Functions as an Endocytic Sonic Hedgehog Receptor.* J. Biol. Chem., 2002. **277**(28): p. 25660-25667.

61. Pons, S. and E. Marti, *Sonic hedgehog synergizes with the extracellular matrix protein vitronectin to induce spinal motor neuron differentiation.* Development, 2000. **127**(2): p. 333-42.

62. Park, Y., et al., *Drosophila perlecan modulates FGF and hedgehog signals to activate neural stem cell division.* Dev Biol, 2003. **253**(2): p. 247-57.

63. Tsai, M.T., et al., *Isolation and characterization of a secreted, cell-surface glycoprotein SCUBE2 from humans.* Biochem J, 2009. **422**(1): p. 119-28.

64. Glise, B., et al., *Shifted, the Drosophila ortholog of Wnt inhibitory factor-1, controls the distribution and movement of Hedgehog.* Dev Cell, 2005. **8**(2): p. 255-66.

65. Gorfinkiel, N., et al., *The Drosophila ortholog of the human Wnt inhibitor factor Shifted controls the diffusion of lipid-modified Hedgehog.* Dev Cell, 2005. **8**(2): p. 241-53.

66. Carstea, E.D., et al., *Niemann-Pick C1 disease gene: homology to mediators of cholesterol homeostasis.* Science, 1997. **277**(5323): p. 228-31.

67. Denef, N., et al., *Hedgehog induces opposite changes in turnover and subcellular localization of patched and smoothened.* Cell, 2000. **102**: p. 521-552.

68. Incardona, J.P., et al., *Receptor-mediated endocytosis of soluble and membrane-tethered Sonic hedgehog by Patched-1.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12044-9.

69. Taipale, J., et al., *Patched acts catalytically to suppress the activity of Smoothened.* Nature, 2002. **418**(6900): p. 892-7.

70. Alcedo, J., et al., *The Drosophila smoothened gene encodes a seven-pass membrane protein, a putative receptor for the hedgehog signal.* Cell, 1996. **86**(2): p. 221-32.

71. van den Heuvel, M. and P.W. Ingham, *smoothened encodes a receptor-like serpentine protein required for hedgehog signalling.* Nature, 1996. **382**(6591): p. 547-51.

72. Bijlsma, M., et al., *Repression of smoothened by patched-dependent (pro-)vitamin D3 secretion.* PLoS biology, 2006. **4**.

73. Dwyer, J.R., et al., *Oxysterols are novel activators of the hedgehog signaling pathway in pluripotent mesenchymal cells.* J Biol Chem, 2007. **282**(12): p. 8959-68.

74. Eaton, S., *Multiple roles for lipids in the Hedgehog signalling pathway.* Nat Rev Mol Cell Biol, 2008. **9**(6): p. 437-45.

75. Chen, J.K., et al., *Inhibition of Hedgehog signaling by direct binding of cyclopamine to Smoothened.* Genes Dev, 2002. **16**(21): p. 2743-8.

76. Bidet, M., et al., *The Hedgehog Receptor Patched Is Involved in Cholesterol Transport.* PLoS One, 2011. **6**(9): p. e23834.

77. Yavari, A., et al., *Role of lipid metabolism in smoothened derepression in hedgehog signaling.* Dev Cell, 2010. **19**(1): p. 54-65.

78. Hooper, J.E. and M.P. Scott, *Communicating with Hedgehogs.* Nat Rev Mol Cell Biol, 2005. **6**(4): p. 306-17.

79. Wu, X., H. Chen, and X. Wang, *Can lung cancer stem cells be targeted for therapies?* Cancer Treatment Reviews, (0).

80. Odent, S., et al., *Expression of the Sonic hedgehog (SHH ) gene during early human development and phenotypic expression of new mutations causing holoprosencephaly.* Hum Mol Genet, 1999. **8**(9): p. 1683-9.

81. Nanni, L., et al., *The mutational spectrum of the sonic hedgehog gene in holoprosencephaly: SHH mutations cause a significant proportion of autosomal dominant holoprosencephaly.* Hum Mol Genet, 1999. **8**(13): p. 2479-88.

82. Mortier, G.R., et al., *Acrocapitofemoral dysplasia: an autosomal recessive skeletal dysplasia with cone shaped epiphyses in the hands and hips.* J Med Genet, 2003. **40**(3): p. 201-7.

83. Umehara, F., et al., *A novel mutation of desert hedgehog in a patient with 46,XY partial gonadal dysgenesis accompanied by minifascicular neuropathy.* American journal of human genetics, 2000. **67**(5): p. 1302-1307.

84. Rubin, L. and F. de Sauvage, *Targeting the Hedgehog pathway in cancer.* Nature reviews. Drug discovery, 2006. **5**(12): p. 1026-1059.

85. Taipale, J. and P. Beachy, *The Hedgehog and Wnt signalling pathways in cancer.* Nature, 2001. **411**(6835): p. 349-403.

86. Mohler, J., *Requirements for hedgehog, a segmental polarity gene, in patterning larval and adult cuticle of Drosophila.* Genetics, 1988. **120**(4): p. 1061-72.

87. Basler, K. and G. Struhl, *Compartment boundaries and the control of Drosophila limb pattern by hedgehog protein.* Nature, 1994. **368**(6468): p. 208-14.

88. Diaz-Benjumea, F.J., B. Cohen, and S.M. Cohen, *Cell interaction between compartments establishes the proximal-distal axis of Drosophila legs.* Nature, 1994. **372**(6502): p. 175-9.

89. Heberlein, U., et al., *Growth and differentiation in the Drosophila eye coordinated by hedgehog.* Nature, 1995. **373**(6516): p. 709-11.

90. Dominguez, M., *Dual role for Hedgehog in the regulation of the proneural gene atonal during ommatidia development.* Development, 1999. **126**(11): p. 2345-53.

91. Deshpande, G., et al., *Hedgehog signaling in germ cell migration.* Cell, 2001. **106**(6): p. 759-69.

92. Huang, Z. and S. Kunes, *Hedgehog, transmitted along retinal axons, triggers neurogenesis in the developing visual centers of the Drosophila brain.* Cell, 1996. **86**(3): p. 411-22.

93. Huang, Z. and S. Kunes, *Signals transmitted along retinal axons in Drosophila: Hedgehog signal reception and the cell circuitry of lamina cartridge assembly.* Development, 1998. **125**(19): p. 3753-64.

94. Forbes, A.J., et al., *hedgehog is required for the proliferation and specification of ovarian somatic cells prior to egg chamber formation in Drosophila.* Development, 1996. **122**(4): p. 1125-35.

95. Zhang, Y. and D. Kalderon, *Regulation of cell proliferation and patterning in Drosophila oogenesis by Hedgehog signaling.* Development, 2000. **127**(10): p. 2165-76.

96. Struhl, G., D.A. Barbash, and P.A. Lawrence, *Hedgehog organises the pattern and polarity of epidermal cells in the Drosophila abdomen.* Development, 1997. **124**(11): p. 2143-54.

97. Pankratz, M.J. and M. Hoch, *Control of epithelial morphogenesis by cell signaling and integrin molecules in the Drosophila foregut.* Development, 1995. **121**(6): p. 1885-98.

98. Glazer, L. and B.Z. Shilo, *Hedgehog signaling patterns the tracheal branches.* Development, 2001. **128**(9): p. 1599-606.

99. Franco, H.L. and H.H. Yao, *Sex and hedgehog: roles of genes in the hedgehog signaling pathway in mammalian sexual differentiation.* Chromosome Res, 2012. **20**(1): p. 247-58.

100. Bitgood, M.J., L. Shen, and A.P. McMahon, *Sertoli cell signaling by Desert hedgehog regulates the male germline.* Curr Biol, 1996. **6**(3): p. 298-304.

101. Wijgerde, M., et al., *Hedgehog signaling in mouse ovary: Indian hedgehog and desert hedgehog from granulosa cells induce target gene expression in developing theca cells.* Endocrinology, 2005. **146**(8): p. 3558-66.

102. Barsoum, I. and H.H. Yao, *Redundant and differential roles of transcription factors Gli1 and Gli2 in the development of mouse fetal Leydig cells.* Biol Reprod, 2011. **84**(5): p. 894-9.

103. Bitgood, M.J. and A.P. McMahon, *Hedgehog and Bmp Genes Are Coexpressed at Many Diverse Sites of Cell–Cell Interaction in the Mouse Embryo.* Developmental Biology, 1995. **172**(1): p. 126-138.

104. Barsoum, I.B., et al., *Activation of the Hedgehog pathway in the mouse fetal ovary leads to ectopic appearance of fetal Leydig cells and female pseudohermaphroditism.* Developmental Biology, 2009. **329**(1): p. 96-103.

105. Russell, M.C., et al., *The Hedgehog Signaling Pathway in the Mouse Ovary.* Biology of Reproduction, 2007. **77**(2): p. 226-236.

106. Kucenas, S., et al., *CNS-derived glia ensheath peripheral nerves and mediate motor root development.* Nat Neurosci, 2008. **11**(2): p. 143-51.

107. Parmantier, E., et al., *Schwann cell-derived Desert hedgehog controls the development of peripheral nerve sheaths.* Neuron, 1999. **23**(4): p. 713-24.

108. Scherer, S.S. and L. Wrabetz, *Molecular mechanisms of inherited demyelinating neuropathies.* Glia, 2008. **56**(14): p. 1578-89.

109. Dyer, M.A., et al., *Indian hedgehog activates hematopoiesis and vasculogenesis and can respecify prospective neurectodermal cell fate in the mouse embryo.* Development, 2001. **128**(10): p. 1717-30.

110. Vortkamp, A., et al., *Regulation of rate of cartilage differentiation by Indian hedgehog and PTH-related protein.* Science, 1996. **273**(5275): p. 613-22.

111. St-Jacques, B., M. Hammerschmidt, and A.P. McMahon, *Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation.* Genes Dev, 1999. **13**(16): p. 2072-86.

112. Murakami, S., A. Nifuji, and M. Noda, *Expression of Indian hedgehog in osteoblasts and its posttranscriptional regulation by transforming growth factor-beta.* Endocrinology, 1997. **138**(5): p. 1972-8.

113. Colnot, C., et al., *Indian hedgehog synchronizes skeletal angiogenesis and perichondrial maturation with cartilage development.* Development, 2005. **132**(5): p. 1057-67.

114. Hellemans, J., et al., *Homozygous mutations in IHH cause acrocapitofemoral dysplasia, an autosomal recessive disorder with cone-shaped epiphyses in hands and hips.* Am J Hum Genet, 2003. **72**(4): p. 1040-6.

115. Sampath, K., et al., *Functional differences among Xenopus nodal-related genes in left-right axis determination.* Development, 1997. **124**(17): p. 3293-3302.

116. Pagan-Westphal, S.M. and C.J. Tabin, *The transfer of left-right positional information during chick embryogenesis.* Cell, 1998. **93**(1): p. 25-35.

117. Schilling, T.F., J.P. Concordet, and P.W. Ingham, *Regulation of left-right asymmetries in the zebrafish by Shh and BMP4.* Dev Biol, 1999. **210**(2): p. 277-87.

118. Watanabe, Y. and H. Nakamura, *Control of chick tectum territory along dorsoventral axis by Sonic hedgehog.* Development, 2000. **127**(5): p. 1131-40.

119. Johnson, R.L., et al., *Ectopic expression of Sonic hedgehog alters dorsal-ventral patterning of somites.* Cell, 1994. **79**(7): p. 1165-73.

120. Marti, E., et al., *Distribution of Sonic hedgehog peptides in the developing chick and mouse embryo.* Development, 1995. **121**(8): p. 2537-47.

121. Chiang, C., et al., *Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function.* Nature, 1996. **383**(6599): p. 407-13.

122. Litingtung, Y., et al., *Shh and Gli3 are dispensable for limb skeleton formation but regulate digit number and identity.* Nature, 2002. **418**(6901): p. 979-83.

123. Litingtung, Y., et al., *Sonic hedgehog is essential to foregut development.* Nat Genet, 1998. **20**(1): p. 58-61.

124. Pepicelli, C.V., P.M. Lewis, and A.P. McMahon, *Sonic hedgehog regulates branching morphogenesis in the mammalian lung.* Curr Biol, 1998. **8**(19): p. 1083-6.

125. Roessler, E., et al., *The mutational spectrum of holoprosencephaly-associated changes within the SHH gene in humans predicts loss-of-function through either key structural alterations of the ligand or its altered synthesis.* Hum Mutat, 2009. **30**(10): p. E921-35.

126. Matsunaga, E. and K. Shiota, *Holoprosencephaly in human embryos: epidemiologic studies of 150 cases.* Teratology, 1977. **16**(3): p. 261-72.

127. Roach, E., et al., *Holoprosencephaly: birth data, benetic and demographic analyses of 30 families.* Birth Defects Orig Artic Ser, 1975. **11**(2): p. 294-313.

128. Amitai, G., et al., *Distribution and function of new bacterial intein-like protein domains.* Mol Microbiol, 2003. **47**(1): p. 61-73.

129. Burglin, T.R., *Evolution of hedgehog and hedgehog-related genes, their origin from Hog proteins in ancestral eukaryotes and discovery of a novel Hint motif.* BMC Genomics, 2008. **9**: p. 127.

130. Dassa, B., I. Yanai, and S. Pietrokovski, *New type of polyubiquitin-like genes with intein-like autoprocessing domains.* Trends Genet, 2004. **20**(11): p. 538-42.

131. Dassa, B. and S. Pietrokovski, *Origin and Evolution of Inteins and Other Hint Domains Homing Endonucleases and Inteins*, M. Belfort, et al., Editors. 2005, Springer Berlin Heidelberg. p. 211-231.

132. Requena, N., et al., *Early developmentally regulated genes in the arbuscular mycorrhizal fungus Glomus mosseae: identification of GmGIN1, a novel gene with homology to the C-terminus of metazoan hedgehog proteins.* Plant and Soil, 2002. **244**(1): p. 129-139.

133. Snell, E.A., et al., *An unusual choanoflagellate protein released by Hedgehog autocatalytic processing.* Proc Biol Sci, 2006. **273**(1585): p. 401-7.

134. Adamska, M., et al., *The evolutionary origin of hedgehog proteins.* Curr Biol, 2007. **17**(19): p. R836-7.

135. Matus, D.Q., et al., *The Hedgehog gene family of the cnidarian, Nematostella vectensis, and implications for understanding metazoan Hedgehog pathway evolution.* Dev Biol, 2008. **313**(2): p. 501-18.

136. Zardoya, R., E. Abouheif, and A. Meyer, *Evolution and orthology of hedgehog genes.* Trends Genet, 1996. **12**(12): p. 496-7.

137. Kumar, S., K. Balczarek, and Z. Lai, *Evolution of the hedgehog gene family.* Genetics, 1996. **142**(3): p. 965-1037.

138. Benson, D.A., et al., *GenBank.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D32-D37.

139. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Research, 2012. **40**(D1): p. D84-D90.

140. Avaron, F., et al., *Characterization of two new zebrafish members of the hedgehog family: atypical expression of a zebrafish indian hedgehog gene in skeletal elements of both endochondral and dermal origins.* Dev Dyn, 2006. **235**(2): p. 478-89.

141. Hadzhiev, Y., et al., *Functional diversification of sonic hedgehog paralog enhancers identified by phylogenomic reconstruction.* Genome Biol, 2007. **8**(6): p. R106.

142. Bingham, S., et al., *Sonic hedgehog and tiggy-winkle hedgehog cooperatively induce zebrafish branchiomotor neurons.* Genesis, 2001. **30**(3): p. 170-4.

143. Meyer, A. and Y. Van de Peer, *From 2R to 3R: evidence for a fish-specific genome duplication (FSGD).* Bioessays, 2005. **27**(9): p. 937-45.

144. Ekker, S.C., et al., *Distinct expression and shared activities of members of the hedgehog gene family of Xenopus laevis.* Development, 1995. **121**(8): p. 2337-47.

145. Kano, S., et al., *Two lamprey Hedgehog genes share non-coding regulatory sequences and expression patterns with gnathostome Hedgehogs.* PLoS One, 2010. **5**(10): p. e13332.

146. Takatori, N., Y. Satou, and N. Satoh, *Expression of hedgehog genes in Ciona intestinalis embryos.* Mech Dev, 2002. **116**(1-2): p. 235-8.

147. Ohno, S., *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag., 1970.

148. Song, K., et al., *Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution.* Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7719-23.

149. Sidow, A., *Gen(om)e duplications in the evolution of early vertebrates.* Curr Opin Genet Dev, 1996. **6**(6): p. 715-22.

150. Meyer, A. and M. Schartl, *Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions.* Curr Opin Cell Biol, 1999. **11**(6): p. 699-704.

151. Wolfe, K.H., *Yesterday's polyploids and the mystery of diploidization.* Nat Rev Genet, 2001. **2**(5): p. 333-41.

152. Ohta, T., *Further simulation studies on evolution by gene duplication.* Evolution, 1988: p. 375-761.

153. Ohta, T., *Multigene and supergene families.* Oxf. Surv. Evol. Biol, 1988. **5**: p. 41-106.

154. Hughes, A.L., *Adaptive evolution of genes and genomes*. 1999: Oxford University Press, USA.

155. Tautz, D., *Redundancies, development and the flow of information.* Bioessays, 1992. **14**(4): p. 263-6.

156. Pickett, F.B. and D.R. Meeks-Wagner, *Seeing double: appreciating genetic redundancy.* Plant Cell, 1995. **7**(9): p. 1347-56.

157. Thomas, J.H., *Thinking about genetic redundancy.* Trends in genetics : TIG, 1993. **9**(11): p. 395-399.

158. Evans, B.J., *Genome evolution and speciation genetics of clawed frogs (Xenopus and Silurana).* Front Biosci, 2008. **13**: p. 4687-706.

159. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.

160. Prince, V.E. and F.B. Pickett, *Splitting pairs: the diverging fates of duplicated genes.* Nat Rev Genet, 2002. **3**(11): p. 827-37.

161. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.

162. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**(1): p. 459-73.

163. Force, A., W.A. Cresko, and F.B. Pickett, *Informational accretion, gene duplication, and the mechanisms of genetic module parcellation.*, in *Modularity in Development and Evolution*, G. Schlosser and G. Wagner, Editors. 2002, University of Chicago Press: Illinois.

164. Taylor, J.S. and J. Raes, *Duplication and divergence: the evolution of new genes and old ideas.* Annu Rev Genet, 2004. **38**: p. 615-43.

165. Papp, B., C. Pal, and L.D. Hurst, *Dosage sensitivity and the evolution of gene families in yeast.* Nature, 2003. **424**(6945): p. 194-7.

166. Swanson, W.J., *Adaptive evolution of genes and gene families.* Curr Opin Genet Dev, 2003. **13**(6): p. 617-22.

167. Pybus, O.G. and B. Shapiro, *Natural Selection and Adaptation of Molecular Sequences*, in *The Phylogenetic Handbook: A Pratical Approach to Phylogenetic Analysis and Hypothesis Testing*, P. Lemey, M. Salemi, and A.-M. Vandamme, Editors. 2009, Cambridge University Press: Cambridge.

168. Zhang, J., H.F. Rosenberg, and M. Nei, *Positive Darwinian selection after gene duplication in primate ribonuclease genes.* Proc Natl Acad Sci U S A, 1998. **95**(7): p. 3708-13.

169. Schmidt, T.R., M. Goodman, and L.I. Grossman, *Molecular evolution of the COX7A gene family in primates.* Mol Biol Evol, 1999. **16**(5): p. 619-26.

170. Duda, T.F., Jr. and S.R. Palumbi, *Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus.* Proc Natl Acad Sci U S A, 1999. **96**(12): p. 6820-3.

171. Kimura, M., *The Neutral Theory of Molecular Evolution*. 1985: Cambridge University Press.

172. Hughes, A.L. and M. Nei, *Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection.* Nature, 1988. **335**(6186): p. 167-170.

173. Yang, Z., *Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.* Molecular Biology and Evolution, 1998. **15**: p. 568-641.

174. Bielawski, J. and Z. Yang, *Maximum likelihood methods for detecting adaptive evolution after gene duplication.* Journal of structural and functional genomics, 2003. **3**: p. 201-213.

175. Bielawski, J.P. and Z. Yang, *Positive and negative selection in the DAZ gene family.* Mol Biol Evol, 2001. **18**(4): p. 523-9.

176. Nielsen, R. and Z. Yang, *Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.* Genetics, 1998. **148**(3): p. 929-36.

177. Yang, Z., et al., *Codon-substitution models for heterogeneous selection pressure at amino acid sites.* Genetics, 2000. **155**: p. 431-480.

178. Yang, Z. and R. Nielsen, *Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages.* Molecular Biology and Evolution, 2002. **19**(6): p. 908-917.

179. McClellan, D. and K. McCracken, *Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains.* Molecular Biology and Evolution, 2001. **18**: p. 917-942.

180. McClellan, D., et al., *Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins.* Molecular Biology and Evolution, 2005. **22**: p. 437-492.

181. Xia, X. and W.-H. Li, *What Amino Acid Properties Affect Protein Evolution?* Journal of molecular evolution, 1998. **47**(5): p. 557-564.

182. Li, W.H., *Evolution of Duplicated Genes*, in *Evolution of Genes and Proteins*, M. Nei and R.K. Koehn, Editors. 1983, Sinauer Associates: Sunderland, Massachusetts.

183. Lichtarge, O., H.R. Bourne, and F.E. Cohen, *An evolutionary trace method defines binding surfaces common to protein families.* J Mol Biol, 1996. **257**(2): p. 342-58.

184. Gu, X., *Statistical methods for testing functional divergence after gene duplication.* Molecular Biology and Evolution, 1999. **16**: p. 1664-1738.

185. Gu, X., *Maximum-likelihood approach for gene family evolution under functional divergence.* Molecular Biology and Evolution, 2001. **18**: p. 453-517.

186. Gu, X., *A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences.* Molecular Biology and Evolution, 2006. **23**: p. 1937-1982.

187. Golding, G.B. and A.M. Dean, *The structural basis of molecular adaptation.* Mol Biol Evol, 1998. **15**(4): p. 355-69.

188. Casari, G., C. Sander, and A. Valencia, *A method to predict functional residues in proteins.* Nat Struct Biol, 1995. **2**(2): p. 171-8.

189. Landgraf, R., D. Fischer, and D. Eisenberg, *Analysis of heregulin symmetry by weighted evolutionary tracing.* Protein Eng, 1999. **12**(11): p. 943-51.

190. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

191. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

192. Tamura, K., et al., *MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.* Molecular Biology and Evolution, 2011. **28**(10): p. 2731-2739.

193. Gouy, M., S. Guindon, and O. Gascuel, *SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.* Mol Biol Evol, 2010. **27**(2): p. 221-4.

194. Dessimoz, C. and M. Gil, *Phylogenetic assessment of alignments reveals neglected tree signal in gaps.* Genome biology, 2010. **11**.

195. Muffato, M., et al., *Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes.* Bioinformatics, 2010. **26**(8): p. 1119-21.

196. Alföldi, J., et al., *The genome of the green anole lizard and a comparative analysis with birds and mammals.* Nature, 2011. **477**(7366): p. 587-678.

197. Hillier, L.D.W., et al., *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-1411.

198. Fujita, P.A., et al., *The UCSC Genome Browser database: update 2011.* Nucleic Acids Research, 2010.

199. Okonechnikov, K., et al., *Unipro UGENE: a unified bioinformatics toolkit.* Bioinformatics, 2012.

200. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Research, 2009. **19**(9): p. 1639-1684.

201. Bickle, T.A., V. Pirrotta, and R. Imber, *A simple, general procedure for purifying restriction endonucleases.* Nucleic Acids Res, 1977. **4**(8): p. 2561-72.

202. Winterbourne, D. and J. Salisbury, *Heparan sulphate is a potent inhibitor of DNA synthesis in vitro.* Biochemical and biophysical research communications, 1981. **101**(1): p. 30-37.

203. Doenecke, D., *Differential response of avian red blood cell nucleosomes to heparin.* Biochem Int, 1984. **9**(1): p. 129-36.

204. Khosravinia, H. and K.P. Ramesha, *Influence of EDTA and magnesium on DNA extraction from blood samples and specificity of polymerase chain reaction.* African Journal of Biotechnology, 2007. **6**(3): p. 184-187.

205. IDT. *Oligo Analyzer 3.1*. 2012; http://eu.idtdna.com/analyzer/applications/oligoanalyzer/default.aspx].

206. PerkinElmer, *FinchTV 1.4*. 2012. p. Available from http://www.geospiza.com/Products/finchtv.shtml.

207. Drummond, A.J., et al., *Geneious v5.4*. 2011. p. Available from http://www.geneious.com/.

208. Posada, D., *jModelTest: Phylogenetic Model Averaging.* Molecular Biology and Evolution, 2008. **25**(7): p. 1253-1256.

209. Xia, X. and Z. Xie, *DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution.* Journal of Heredity, 2001. **92**(4): p. 371-373.

210. Xia, X., et al., *An index of substitution saturation and its application.* Molecular phylogenetics and evolution, 2003. **26**: p. 1-8.

211. Strimmer, K. and A. Von Haeseler, *Genetic Distances and Nucleotide Subsitution Models*, in *The Phylogenetic Handbook: a Pratucal Approach to Phylogenetic Analysis and Hypothesis Testing*, P. Lemey, M. Salemi, and A.-M. Vandamme, Editors. 2009, Cambridge University Press: Cambridge.

212. Guindon, S., et al., *Estimating maximum likelihood phylogenies with PhyML.* Methods in molecular biology (Clifton, N.J.), 2009. **537**: p. 113-150.

213. Huelsenbeck, J. and F. Ronquist, *MRBAYES: Bayesian inference of phylogenetic trees.* Bioinformatics (Oxford, England), 2001. **17**(8): p. 754-759.

214. Ronquist, F. and J. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics (Oxford, England), 2003. **19**(12): p. 1572-1576.

215. Castresana, J., *Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis.* Molecular Biology and Evolution, 2000. **17**(4): p. 540-552.

216. Talavera, G. and J. Castresana, *Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments.* Systematic Biology, 2007. **56**(4): p. 564-577.

217. Yang, Z., *PAML 4: Phylogenetic Analysis by Maximum Likelihood.* Molecular Biology and Evolution, 2007. **24**(8): p. 1586-1591.

218. Yang, Z., *Inference of selection from multiple species alignments.* Current opinion in genetics & development, 2002. **12**: p. 688-782.

219. Yang, Z., W. Wong, and R. Nielsen, *Bayes empirical bayes inference of amino acid sites under positive selection.* Molecular Biology and Evolution, 2005. **22**: p. 1107-1125.

220. Kosakovsky Pond, S. and S. Frost, *Not so different after all: a comparison of methods for detecting amino acid sites under selection.* Molecular Biology and Evolution, 2005. **22**: p. 1208-1230.

221. Pond, S. and S. Frost, *Datamonkey: rapid detection of selective pressure on individual sites of codon alignments.* Bioinformatics (Oxford, England), 2005. **21**: p. 2531-2534.

222. Delport, W., et al., *Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology.* Bioinformatics (Oxford, England), 2010. **26**: p. 2455-2462.

223. da Fonseca, R., et al., *The adaptive evolution of the mammalian mitochondrial genome.* BMC Genomics, 2008. **9**: p. 119.

224. da Fonseca, R., et al., *Structural divergence and adaptive evolution in mammalian cytochromes P450 2C.* Gene, 2007. **387**: p. 58-124.

225. Woolley, S., et al., *TreeSAAP: selection on amino acid properties using phylogenetic trees.* Bioinformatics (Oxford, England), 2003. **19**: p. 671-673.

226. Porter, M., et al., *Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins.* Molecular Biology and Evolution, 2007. **24**: p. 253-321.

227. Gu, X. and K. Vander Velden, *DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family.* Bioinformatics (Oxford, England), 2002. **18**: p. 500-501.

228. Wang, Y. and X. Gu, *Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction.* Genetics, 2001. **158**: p. 1311-1331.

229. Jeffreys, H., ed. *Theory of Probability*. 3rd ed., ed. O.U. Press. 1998.

230. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.

231. Roy, A., et al., *A protocol for computer-based protein structure and function prediction.* J Vis Exp, 2011(57): p. e3259.

232. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.

233. Battey, J.N., et al., *Automated server predictions in CASP7.* Proteins, 2007. **69 Suppl 8**: p. 68-82.

234. Zhang, Y., *Template-based modeling and free modeling by I-TASSER in CASP7.* Proteins, 2007. **69 Suppl 8**: p. 108-17.

235. Zhang, Y., *I-TASSER: fully automated protein structure prediction in CASP8.* Proteins, 2009. **77 Suppl 9**: p. 100-13.

236. Cozzetto, D., et al., *Evaluation of template-based models in CASP8 with standard measures.* Proteins, 2009. **77 Suppl 9**: p. 18-28.

237. Wu, S., A. Szilagyi, and Y. Zhang, *Improving protein structure prediction using multiple sequence-based contact predictions.* Structure, 2011. **19**(8): p. 1182-91.

238. Kinch, L., et al., *CASP9 assessment of free modeling target predictions.* Proteins, 2011. **79 Suppl 10**: p. 59-73.

239. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nature protocols, 2010. **5**: p. 725-763.

240. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.

241. *The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC*.

242. Sun, X.-J., et al., *Genome-Wide Survey and Developmental Expression Mapping of Zebrafish SET Domain-Containing Genes.* PLoS One, 2008. **3**(1): p. e1499.

243. Djabali, M., et al., *A trithorax-like gene is interrupted by chromosome 11q23 translocations in acute leukaemias.* Nature genetics, 1992. **2**(2): p. 113-121.

244. Krivtsov, A. and S. Armstrong, *MLL translocations, histone modifications and leukaemia stem-cell development.* Nature reviews. Cancer, 2007. **7**(11): p. 823-856.

245. Dalloul, R., et al., *Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis.* PLoS biology, 2010. **8**(9).

246. Warren, W., et al., *The genome of a songbird.* Nature, 2010. **464**(7289): p. 757-819.

247. Darling, A.C.E., et al., *Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements.* Genome Research, 2004. **14**(7): p. 1394-1403.

248. Dutcher, S.K., *The tubulin fraternity: alpha to eta.* Curr Opin Cell Biol, 2001. **13**(1): p. 49-54.

249. McKean, P.G., S. Vaughan, and K. Gull, *The extended tubulin superfamily.* J Cell Sci, 2001. **114**(Pt 15): p. 2723-33.

250. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database : the journal of biological databases and curation, 2011. **2011**.

251. Irimia, M., et al., *Comparative genomics of the Hedgehog loci in chordates and the origins of Shh regulatory novelties.* Sci. Rep., 2012. **2**.

252. Kuraku, S., A. Meyer, and S. Kuratani, *Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after?* Molecular Biology and Evolution, 2009. **26**: p. 47-106.

253. Burt, D.W., *Origin and evolution of avian microchromosomes.* Cytogenet Genome Res, 2002. **96**(1-4): p. 97-112.

254. Ladjali-Mohammedi, K., et al., *International system for standardized avian karyotypes (ISSAK): standardized banded karyotypes of the domestic fowl (Gallus domesticus).* Cytogenet Cell Genet, 1999. **86**(3-4): p. 271-6.

255. *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.

256. Masabanda, J.S., et al., *Molecular cytogenetic definition of the chicken genome: the first complete avian karyotype.* Genetics, 2004. **166**(3): p. 1367-73.

257. Morisson, M., et al., *The chicken RH map: current state of progress and microchromosome mapping.* Cytogenet Genome Res, 2007. **117**(1-4): p. 14-21.

258. Trukhina, A.V. and A.F. Smirnov, *Microsatellites from the linkage groups E26C13 and E50C23 are located on the Gallus gallus domesticus microchromosomes 20 and 21.* Genetika, 2010. **46**(4): p. 509-16.

259. Davis, J., O. Brandman, and D. Petrov, *Protein evolution in the context of Drosophila development.* Journal of molecular evolution, 2005. **60**: p. 774-859.

260. Roux, J. and M. Robinson-Rechavi, *Developmental constraints on vertebrate genome evolution.* PLoS Genet, 2008. **4**(12): p. e1000311.

261. Gunbin, K., D. Afonnikov, and N. Kolchanov, *The evolution of the Hh-signaling pathway genes: a computer-assisted study.* In silico biology, 2007. **7**: p. 333-387.

262. Swanson, W.J. and V.D. Vacquier, *The rapid evolution of reproductive proteins.* Nat Rev Genet, 2002. **3**(2): p. 137-144.

263. Sobrinho, I. and R. de Brito, *Evidence for positive selection in the gene fruitless in Anastrepha fruit flies.* BMC evolutionary biology, 2010. **10**(1): p. 293.

# *SUPPLEMENTARY MATERIAL*

**Table S1**. List of Hh-coding sequences collected from currently available genomes. (*Continues*)

| Species | | | Gene | Database | Ref/ID | CDS State | | |
|---|---|---|---|---|---|---|---|---|
| ***Common name*** | ***Scientific name*** | ***Class*** | | | | ***State*** | ***Present*** | ***Missing*** |
| Alpaca | *Vicugna pacos* | Mammalia | Dhh | Ensembl | ENSVPAT00000000165 | Incomplete | HhC complete | HhN |
| Nine-Banded Armadillo | *Dasypus novemcinctus* | Mammalia | Dhh | Ensembl | ENSDNOG00000024276 | Incomplete | HhC complete | HhN N-region |
| | | | Shh | Ensembl | ENSDNOG00000011281 | Complete | | HhC and HHN incomplete |
| Hamadryas baboon | *Papio hamadryas* | Mammalia | Ihh | Ensembl (BLASTN) | ENSP00000295731_1 | Incomplete | All domains complete | N and C-terminal |
| Northern greater galago | *Otolemur garnettii* | Mammalia | Dhh | Ensembl | ENSOGAG00000013485 | Incomplete | HhC complete | HhN N-region |
| | | | Ihh | Ensembl | ENSOGAG00000014108 | Complete | All domains complete | na |
| | | | Shh | Ensembl | ENSOGAG00000000175 | Incomplete | HhC complete | HhN N-region |
| Cat | *Felis catus* | Mammalia | Dhh | Ensembl | ENSFCAG00000000097 | Incomplete | All domains complete | HhN N-terminal |
| | | | Shh | NCBI | HQ437701.1 | Incomplete | HhC complete | HhN |
| Chimpanzee | *Pan troglodytes* | Mammalia | Ihh | NCBI | XM_526034.3 | Complete | All domains complete | na |
| | | | Shh | NCBI | XM_001147185.2 | Complete | All domains complete | na |
| Cow | *Bos taurus* | Mammalia | Dhh | NCBI | XM_001788869.1 | Complete | All domains complete | na |
| | | | Ihh | NCBI | NM_001076870.2 | Complete | All domains complete | na |
| | | | Shh | NCBI | XM_614193.3 | Complete | All domains complete | na |
| Dog | *Canis familiaris* | Mammalia | Dhh | Ensembl | ENSCAFG00000008694 | Incomplete | All domains complete | Termination |
| | | | Ihh | NCBI | XM_545653.3 | Complete | All domains complete | na |
| | | | Shh | NCBI | XM_845357.1 | Complete | All domains complete | na |
| Dolphin | *Tursiops truncatus* | Mammalia | Dhh | Ensembl | ENSTTRG00000004783 | Complete | All domains complete | na |
| | | | Ihh | Ensembl | ENSTTRG00000010019 | Complete | All domains complete | na |
| | | | Shh | Ensembl | ENSTTRG00000013721 | Complete | All domains complete | na |
| African Bush Elephant | *Loxodonta africana* | Mammalia | Dhh | NCBI | XM_003405761.1 | Complete | All domains complete | na |
| | | | Ihh | NCBI | XM_003405893.1 | Complete | All domains complete | na |
| | | | Shh | NCBI | XM_003420730.1 | Complete | All domains complete | na |
| Human | *Homo sapiens* | Mammalia | Dhh | NCBI | NM_021044.2 | Complete | All domains complete | na |
| | | | Ihh | NCBI | NM_002181.3 | Complete | All domains complete | na |
| | | | Shh | NCBI | NM_000193.2 | Complete | All domains complete | na |

**Table S1**. List of Hh-coding sequences collected from currently available genomes. (*Continues*)

| Species | | | Gene | Database | Ref/ID | CDS State | | |
|---|---|---|---|---|---|---|---|---|
| ***Common name*** | ***Scientific name*** | ***Class*** | | | | ***State*** | ***Present*** | ***Missing*** |
| Large flying fox (megabat) | *Pteropus vampyrus* | Mammalia | Dhh | Ensembl | ENSPVAG00000004638 | Complete | All domains complete | na |
| | | | Ihh | Ensembl | ENSPVAG00000010782 | Incomplete | HhC complete | HhN N-region |
| | | | Shh | Ensembl | ENSPVAG00000000225 | Complete | HhC complete | HhN incomplete |
| Little brown bat (microbat) | *Myotis lucifugus* | Mammalia | Dhh | Ensembl | ENSMLUG00000008353 | Incomplete | HhN complete | HhC C-terminal |
| | | | Ihh | Ensembl | ENSMLUG00000011503 | Incomplete | All domains complete | na |
| | | | Shh | Ensembl | ENSMLUG00000025004 | Incomplete | HhN incomplete | HhN C-terminal and HhC |
| Mouse | *Mus musculus* | Mammalia | Dhh | NCBI | NM_007857.4 | Complete | All domains complete | na |
| | | | Ihh | NCBI | NM_010544.2 | Complete | All domains complete | na |
| | | | Shh | NCBI | NM_009170.3 | Complete | All domains complete | na |
| Gray short-tailed opossum | *Monodelphis domestica* | Mammalia | Ihh | NCBI | XM_001364284.1 | Complete | All domains complete | na |
| | | | Shh | NCBI | NM_001198553.1 | Complete | All domains complete | na |
| Platypus | *Ornithorhynchus anatinus* | Mammalia | Dhh | Ensembl | ENSOANG00000011798 | Incomplete | | HhN and HhC C-terminal |
| | | | Ihh | NCBI | XM_001514393.2 | Incomplete | HhC complete | HhN N-terminal |
| Tammar Wallaby | *Macropus eugenii* | Mammalia | Dhh | Ensembl | ENSMEUG00000014181 | Complete | All domains complete | na |
| | | | Ihh | Ensembl | ENSMEUG00000003555 | Complete | All domains complete | na |
| | | | Shh | Ensembl | ENSMEUG00000001695 | Complete | HhC complete | HhN incomplete |
| Anole lizard | *Anolis carolinensis* | Reptilia | Dhh | NCBI | XM_003223232.1 | Complete | All domains complete | na |
| | | | Ihh | Ensembl | ENSACAG00000005172 | Complete | All domains complete | na |
| | | | Shh | NCBI | XM_003221928.1 | Complete | All domains complete | na |
| African Rock Python | *Python sebae* | Reptilia | Shh | NCBI | EU555185.1 | Incomplete | All domains complete | N- and C-terminal |
| Chicken (Red Junglefowl) | *Gallus gallus* | Aves | Ihh | NCBI | NM_204957.1 | Complete | All domains complete | na |
| | | | Shh | NCBI | NM_204821.1 | Complete | All domains complete | na |
| | | | Dhh | NCBI | NW_003770744.1 | Incomplete | HhC complete | HhN |
| Wild duck | *Anas platyrhynchos* | Aves | Ihh | Ensembl | ENSAPLG00000012391 | Incomplete | | HhC and HHN incomplete |
| | | | Shh | Ensembl | ENSAPLG00000007226 | Incomplete | All domains complete | Termination |
| Turkey | *Meleagris gallopavo* | Aves | Ihh | Ensembl | ENSMGAG00000011370 | Incomplete | HhC complete | HhN N-terminal |
| | | | Shh | NCBI | XM_003206957.1 | Incomplete | HhC complete | HhN |
| Zebra finch | *Taeniopygia guttata* | Aves | Ihh | NCBI | XM_002192246.1 | Incomplete | HhC complete | HhN incomplete |
| | | | Shh | NCBI | XM_002190708.1 | Complete | All domains complete | na |

**Table S1**. List of Hh-coding sequences collected from currently available genomes. (*Continues*)

| **Species** | | | **Gene** | **Database** | **Ref/ID** | **CDS State** | | |
|---|---|---|---|---|---|---|---|---|
| ***Common name*** | ***Scientific name*** | ***Class*** | | | | ***State*** | ***Present*** | ***Missing*** |
| Western clawed frog | *Xenopus tropicalis* | Amphibia | Dhh | NCBI | NM_001097169.1 | Complete | All domains complete | na |
| | | | Ihh | NCBI | XM_002933942.1 | Complete | All domains complete | na |
| | | | Shh | NCBI | XM_002932498.1 | Complete | All domains complete | na |
| African clawed frog | *Xenopus laevis* | Amphibia | Dhha | NCBI | NM_001085791.1 | Complete | All domains complete | na |
| | | | Dhhb | NCBI | NM_001085792.1 | Complete | All domains complete | na |
| | | | Ihh | NCBI | NM_001085793.1 | Complete | All domains complete | na |
| | | | Shh | NCBI | NM_001088313.1 | Complete | All domains complete | na |
| Fugu | *Takifugu rubripes* | Actinopterygii | Dhh | Ensembl | ENSTRUG00000012191 | Complete | All domains complete | na |
| | | | Ihh1 | Ensembl | ENSTRUG00000012233 | Incomplete | All domains complete | Termination |
| | | | Ihh2 | Ensembl | ENSTRUG00000013525 | Incomplete | All domains complete | N-Terminal |
| | | | Shh | NCBI | AJ507296.1 | Complete | All domains complete | na |
| Medaka | *Oryzias latipes* | Actinopterygii | Dhh | Ensembl | ENSORLG00000007595 | Complete | All domains complete | na |
| | | | Ihh | Ensembl | ENSORLG00000001666 | Complete | All domains complete | na |
| | | | Shh | Ensembl | ENSORLG00000010463 | Complete | All domains complete | na |
| Stickleback | *Gasterosteus aculeatus* | Actinopterygii | Dhh | Ensembl | ENSGACG00000009063 | Complete | All domains complete | na |
| | | | Ihh1 | Ensembl | ENSGACG00000015562 | Incomplete | All domains complete | N- and C-terminal |
| | | | Ihh2 | Ensembl | ENSGACG00000006349 | Incomplete | All domains complete | N-terminal |
| | | | Shh | Ensembl | ENSGACG00000003893 | Complete | All domains complete | na |
| Green spotted puffer | *Tetraodon nigroviridis* | Actinopterygii | Dhh | Ensembl | ENSTNIG00000015068 | Complete | All domains complete | na |
| | | | Ihh1 | Ensembl | ENSTNIG00000016449 | Incomplete | All domains complete | N- and C-terminal |
| | | | Ihh2 | Ensembl | ENSTNIG00000000900 | Incomplete | HhC complete | HhN |
| | | | Shh | Ensembl | ENSTNIG00000012780 | Incomplete | All domains complete | N- and C-terminal |

**Table S1**. List of Hh-coding sequences collected from currently available genomes. (*Continuation*)

| **Species** | | | **Gene** | **Database** | **Ref/ID** | **CDS State** | | |
|---|---|---|---|---|---|---|---|---|
| ***Common name*** | ***Scientific name*** | ***Class*** | | | | ***State*** | ***Present*** | ***Missing*** |
| Zebrafish | *Danio rerio* | Actinopterygii | Dhh | NCBI | NM_001030115.1 | Complete | All domains complete | na |
| | | | Ihha | NCBI | NM_001034993.2 | Complete | All domains complete | na |
| | | | Ihhb | NCBI | NM_131088.1 | Complete | All domains complete | na |
| | | | Shha | NCBI | NM_131063.1 | Complete | All domains complete | na |
| | | | Shhb | NCBI | NM_131199.2 | Complete | All domains complete | na |
| Cat Shark | *Scyliorhinus canicula* | Chondrichthyes | Shh | NCBI | HM991336.1 | Incomplete | All domains complete | N- and C-terminal |
| Sea lamprey | *Petromyzon marinus* | Agnatha | Hha (Ihh) | Ensembl | ENSPMAG00000004136 | Incomplete | HhC complete | HhN N-region |
| Lancelet | *Branchiostoma floridae* | Leptocardii | Hh | NCBI | XM_002592059.1 | Complete | All domains complete | na |
| Vase tunicate | *Ciona intestinalis* | Ascidiacea | Hh1 | NCBI | NM_001032462.1 | Complete | All domains complete | na |
| | | | Hh2 | NCBI | NM_001032463.1 | Complete | All domains complete | na |
| Fruitfly | *Drosophila melanogaster* | Insecta | Hh | Ensembl metazoa | GA18321-RA | Complete | All domains complete | na |
| Malaria mosquitoe | *Anopheles gambiae* | Insecta | Hh | Ensembl metazoa | AGAP001412 | Complete | All domains complete | na |
| Wasp | *Nasonia vitripennis* | Insecta | Hh | Nasonia genome project | | Complete | All domains complete | na |
| Beetle | *Tribolium castaneum* | Insecta | Hh | NCBI | NM_001114365.1 | Complete | All domains complete | na |
| Purple sea urchin | *Strongylocentrotus purpuratus* | Echinoidea | Hh | NCBI | NM_001012702.1 | Complete | All domains complete | na |
| Owl limpet | *Lottia gigantea* | Gastropoda | Hh | Lottia gigantea v1.0 | | Complete | All domains complete | na |

**Table S2**. Positively and negatively selected residues detected by SLAC and FEL. The percentage and the numbering of the negatively selected residues is in agreement with the Homo sapiens sequences and residues associated to a disease or to a functional process, as annotated on the GenBank and UniProt databases, are highlighted.

| *Gene* | *Model* | *p-value* | *Positive* | *Negative* | *%Negative residues* | *Negatively selected residues* |
|---|---|---|---|---|---|---|
| *Dhh* | SLAC | 0.05 | 0 | 112 | 28% | **G24**[¢], P25, G31, L40, Q47, P53, G58, A59, S60, G61, A63, R69, S71, I85, K88, D89, E91, A95, R97, R102, C103, K104, N108, L110, A111, I112, A113, D119, G120, R122, T126, E127, D130, E131, D132, G133, S139, L140, H141, Y142, E143, R145, A146, I149, T151, R154, R156, K158, Y159, L161, **L162°**, A163, A166, E168, A169, G170, E171, V174, Y176, E177, H183, V184, S185, V186, A188, S191, A193, G197, G198[¤], C199, F200, P201, G211, L219, G222, V225, G232, V239, L240, F253, R258, K264, T268, P269, H271, F274, F290, A291, R293, R295, G297, E320, G323, F325, A326, P327, T329, H331, G332, V336, A341, Y344, A345, S349, A353, H354, A356, F357, P359, R361, H381, L387 |
| | | 0.10 | 0 | 135 | 34% | Additionaly R32, R33, G94, R106, V107, R124, L147, S210, R217, L243, D246, A251, L272, L294, R318, E319, L335, L370, P371, S384, L386, L390, L395 |
| | FEL | 0.05 | 0 | 176 | 45% | Additionaly A17, L18, A20, Q39, P42, P50, E54, L57, V67, F74, L77, V78, N82, P83, E90, V121, L123, H134, Q137, D153, G160, L165, V167, D172, R179, V206, R207, K214, L226, R233, V234, L241, E259, A296, V316, V324, D338, S342, L347, W352, L366 |
| | | 0.10 | 0 | 203 | 51% | Additionaly K38, V41, Y45, S93, A109, G144, D148, N157, V182, D189, T205, R221, A227, T237, F242, P262, L265, R277, D298, L308, P310, L328, L340, L363, A365, Y383, R385 |
| *Ihh* | SLAC | 0.05 | 0 | 161 | 39% | **G29**[¢], G31, R32, R38, R39, P41, K43, L44, **P46**[†], L47, Y49, K50, F52, P54, N55, V56, T60, G62, R66, E68, G69, K70, S75, E80, T82, P83, N84, P87, I89, K92, T97, G98, A99, R101, C107, K108, D109, R110, S113, S117, P123, G124, V125,K126, R128, T130, **E131**[‡], D134, E135, D136, G137, H138, H139, E141, E142, S143, L144, H145, Y146, E147, G148, R149, A150, V151, D152, I153, T155, S156, R158, D159, R160, N161, K162, Y163, L165, A167, R168, L169, A170, E172, A173, G174, F175, D176, Y179, Y180, S182, K183, A184, H187, C188, S189, **V190**[†], K191, S192, E193, H194, S195, A196, A197, A198, G201, **G202**[¤], C203, F204, P205, A208, V210, G215, P225, G226, R228, V229, A231, G236, S241, F246, D248, I260, T262, P265, L271, T272, P273, A274, H275, L276, L277, N282, F294, A295, S296, V298, G301, Y303, A215, Y329, A330, P331, L332, T333, T337, V340, V344, S346, C347, F348, A349, L356, A357, Q358, F361, P363, L364, P365, L370, H382, Y384, L388,G392, L396, H402, P403 |
| | | 0.10 | 0 | 170 | 41% | Additionaly P40, I71, E181, L212, P266, H335, F367, P385, Y389 |
| | FEL | 0.05 | 0 | 223 | 54% | Additionaly G27, C28, R37, R42, Q51, S53, K59, L61, G65, A72, R73, E76, F78, L81, N86, D88, F91, E94, D100, L111, N112, A115, V118, L127, V129, G132, S140, T154, D157, V178, H185, V186, P211, V223, R224, P238, T239, D242, V243, L244, L247, R267, A280, Q299, P300, Q302, G336, A345, H355, L366, S369, R393, G405 |
| | | 0.10 | 0 | 237 | 58% | Additionaly D93, I116, L166, A222, G233, L269, H283, R284, L305, A318, E341, L359, W375, E377 |
| *Shh* | SLAC | 0.05 | 0 | 175 | 38% | A23, **C24**[¢], G25, **P26***, R28, R33, R34, K37, K38, **L39***, T40, P41, Y44, F47, I48, N50, V51, A52, L56, G57, G60, Y62, E63, G64, I66, S70, R72, E75, L76, P78, Y80, N81, P82, **D83***, **I84***, I85, F86, **D88***, E89, E90, T92, A94, R96, L97, **C102***, K103, K105, **L106***, **N107***, A108, **L109***, **A110***, S112, V113, **N115***, P118, G119, V120, K121, R123, **V124***, T125, D129, E130, D131, G132, H134, **E136***, E137, S138, L139, Y141, **G143***, **R144***, A145, **T150***, S151, D152, R153, D154, K157, Y158, G159, P163, L164, A165, E167, A168, G169, Y174, Y175, **E176***, **K178***, A179, H180, I181, **C183***, **S184***, **E188***, N189, S190, A192, A193, **G196***, **G197***[¤], **C198***, **F199***, **P200***, G201, S202, G210, K213, V215, **L218***, G221, K223, **V224***, **G231***, **K232***, L242, K249, F252, V254, **I255***, E256, T257, **R263***, **L266***, **T267***, **A268***, **A269***, **L271***, F273, V274, F305, P311, G312, R314, Y265, L327, S338, G343, A344, Y345, **A346***, **P347***, T349, A350, **I354***, L360, A361, **S362***, **C363***, **Y364***, A365, I367, E368, **A373***, **H374***, **A376***, **F377***, **A378***, **P379***, **R381***, L390, H433, **S436***, L439, T442, G443, D448, H453, P454, L455, **G456*** |
| | | 0.10 | 0 | 191 | 41% | Additionaly G21, S59, R68, H133, M160, L161, L207, K216, Q313, V315, A330, I356, H384, **A391***, L446, L447 |
| | FEL | 0.05 | 0 | 253 | 55% | Additionaly **G27***, H35, P36, L42, A43, K45, Q46, **E53***, K54, T55, A58, K65, K87, R101, **I111***, E126, G127, S135, **H140***, **D147***, I148, T149, R155, **S156***, **F170***, **D171***, V173, **S177***, H182, V185, V205, S219, **D222***, L225, **A226***, D228, **S236***, F238, L239, T240, **F241***, K250, Y253, P260, R261, H270, R308, V309, A331, E340, A341, L348, G352, L355, V366, W372, **L382***, A388, P392, A428, Y435, L438 |
| | | 0.10 | 0 | 271 | 59% | Additionaly C19, K32, P49, F73, N79, K186, L234, R244, D245, L272, P276, H277, **R310***, L328, P329, R358, A430, A451 |

[°] Gonadal dysgenesis 46,XY  [†] Brachidactyly type A-1  [‡] Acrocapitofemoral dysplasia  * Holoprosencephaly type 3  [¢] Palmitoylation  [¤] Cholesterol glycin ester

**Table S3**. Sites detected by SLAC and FEL with dN/dS values above 1 but not statistically positively selected.

| *Gene* | *Alignment position* | *dN/dS* | *LRT* | *p-value* |
|---|---|---|---|---|
| *Dhh* | 16 | 2.252 | 0.539 | 0.463 |
| | 17 | 1.095 | 0.005 | 0.943 |
| | 18 | 184702600 | 1.664 | 0.197 |
| | 30 | 1.001 | 0.000 | 1.000 |
| | 36 | 80280800 | 0.998 | 0.318 |
| | 47 | 34422800 | 0.610 | 0.435 |
| | 59 | 39755800 | 1.122 | 0.289 |
| | 96 | 35311600 | 0.284 | 0.594 |
| | 115 | 1.990 | 0.368 | 0.544 |
| | 159 | 133061000 | 0.612 | 0.434 |
| | 180 | 172775600 | 1.158 | 0.282 |
| | 186 | 1.248 | 0.055 | 0.814 |
| | 200 | 18325300 | 0.307 | 0.579 |
| | 212 | 1.140 | 0.019 | 0.890 |
| | 228 | 3.416 | 1.461 | 0.227 |
| | 236 | 2.418 | 0.431 | 0.512 |
| | 243 | 2.250 | 0.347 | 0.556 |
| | 246 | 2.303 | 0.164 | 0.685 |
| | 253 | 2.922 | 0.864 | 0.353 |
| | 268 | 1.717 | 0.289 | 0.591 |
| | 269 | 1.150 | 0.020 | 0.887 |
| | 281 | 1.108 | 0.005 | 0.942 |
| | 302 | 52066000 | 0.981 | 0.322 |
| | 310 | 1.174 | 0.022 | 0.882 |
| | 326 | 1.052 | 0.003 | 0.958 |
| | 328 | 1.473 | 0.089 | 0.766 |
| | 333 | 3.883 | 0.816 | 0.366 |
| | 343 | 2.494 | 0.596 | 0.440 |
| *Ihh* | 72 | 41162000 | 0.757 | 0.384 |
| | 140 | 11575420 | 0.059 | 0.808 |
| | 180 | 1.068 | 0.004 | 0.949 |
| | 182 | 1.430 | 0.081 | 0.776 |
| | 207 | 1.264 | 0.046 | 0.830 |
| | 217 | 1.391 | 0.137 | 0.711 |
| | 222 | 197888200 | 1.620 | 0.203 |
| | 224 | 1.106 | 0.005 | 0.946 |
| | 230 | 131731400 | 1.720 | 0.190 |
| | 264 | 53875600 | 0.431 | 0.511 |
| | 268 | 1.298 | 0.094 | 0.760 |
| | 270 | 1.700 | 0.470 | 0.493 |
| | 274 | 1.116 | 0.014 | 0.905 |
| | 298 | 1.519 | 0.015 | 0.901 |
| | 307 | 3.799 | 0.343 | 0.558 |
| | 333 | 9137740000 | 0.000 | 0.991 |
| | 339 | 1.046 | 0.002 | 0.963 |
| | 340 | 1.574 | 0.174 | 0.676 |
| | 347 | 4.197 | 0.747 | 0.388 |
| | 351 | 1.062 | 0.006 | 0.939 |
| | 352 | 1.404 | 0.132 | 0.716 |
| | 353 | 1.291 | 0.045 | 0.831 |
| *Shh* | 192 | 1.309.000 | 0.090 | 0.760 |
| | 232 | 3.290.000 | 1.160 | 0.280 |
| | 252 | 1.875.000 | 0.240 | 0.620 |

**Table S4**. Sites under strong positive selection (p < 0.001) on the three vertebrate Hh paralogs, according to TreeSAAP. Site numbering refers to the Homo sapiens sequences and an asterisk marks the sites which were detected as under negative selection with SLAC and FEL. A legend for properties symbols is presented on table S5. (*Continues*)

| **Codon Alignment Position** | ***Dhh*** | | ***Ihh*** | | ***Shh*** | |
|---|---|---|---|---|---|---|
| | Sites | Properties | Sites | Properties | Sites | Properties |
| 8 | | | 27* | $\mu$ | | |
| 36 | 55 | $pH_i$ | | | | |
| 43 | 62 | $K^0$ | 66* | $pH_i$ | | |
| 44 | | | 67 | $R_a$ $K^0$ | | |
| 56 | 76 | $pH_i$ | | | | |
| 58 | | | | | 77 | $R_a$ |
| 72 | 92 | $pH_i$ | | | | |
| 80 | 101 | $pH_i$ | | | | |
| 139 | | | 164 | $C_\alpha$ $M_w$ $V^0$ | | |
| 180 | | | 207 | $K^0$ $pH_i$ $C_\alpha$ $M_w$ $V^0$ | | |
| 182 | | | | | 204 | $pH_i$ $H_t$ |
| 184 | 207* | $pH_i$ | 211* | $pH_i$ $R_a$ | | |
| 186 | 209 | $pH_i$ $K^0$ $R_F$ $H_{nc}$ | 213 | $pH_i$ $K^0$ | | |
| 190 | | | 217 | $pH_i$ $K^0$ | | |
| 191 | | | 218 | $P$ | | |
| 192 | 215 | $B_l$ | | | | |
| 194 | 217* | $pH_i$ | | | | |
| 195 | | | 222* | $P_r$ | | |
| 197 | 220 | $pH_i$ $B_l$ $C_\alpha$ $M_v$ $M_w$ $V^0$ $\mu$ $H_t$ | | | 219* | $pH_i$ |
| 198 | | | | | 220 | $R_a$ $H_p$ |
| 206 | 230 | $K^0$ | 234 | $K^0$ | | |
| 207 | | | | | 230 | $pH_i$ $K^0$ |
| 209 | | | | | 232* | $pH_i$ |

**Table S4**. Sites under strong positive selection ($p < 0.001$) on the three vertebrate Hh paralogs, according to TreeSAAP. Site numbering refers to the Homo sapiens sequences and an asterisk marks the sites which were detected as under negative selection with SLAC and FEL. A legend for properties symbols is presented on table S5. (*Continues*)

| **Codon Alignment Position** | ***Dhh*** | | ***Ihh*** | | ***Shh*** | |
|---|---|---|---|---|---|---|
| | Sites | Properties | Sites | Properties | Sites | Properties |
| 210 | | | | | 233 | $K^0$ |
| 211 | | | 239* | $R_a$ | | |
| 217 | | | 245 | $R_a$ | | |
| 224 | 251 | $C_\alpha$ $M_v$ $M_w$ $V^0$ $\mu$ $H_t$ | 255 | $pH_i$ | | |
| 225 | 252 | $pH_i$ | | | | |
| 227 | | | | | 253* | $R_a$ |
| 228 | 255 | $R_a$ $H_p$ | | | | |
| 230 | | | 261 | $pH_i$ | | |
| 243 | 270 | $P$ | | | | |
| 249 | 276 | $P_r$ | 280* | $pH_i$ | | |
| 250 | 277* | $pH_i$ $K^0$ | | | 276* | $K^0$ |
| 251 | | | 282* | $\mu$ | | |
| 256 | 292 | $pH_i$ | | | | |
| 260 | | | 300* | $R_a$ $H_p$ | | |
| 263 | 299 | $R_F$ | | | | |
| 267 | | | | | 324 | $pH_i$ |
| 268 | 306 | $P_r$ $C_\alpha$ $M_w$ $V^0$ | | | | |
| 269 | 307 | $pH_i$ $H_{nc}$ | | | | |
| 274 | 312 | $pH_i$ | | | 331* | $R_a$ |
| 278 | 319* | $K^0$ | 323 | $pH_i$ | 339 | $pH_i$ $K^0$ |
| 280 | 321 | $pH_i$ | | | | |
| 289 | 330 | $K^0$ | 334 | $pH_i$ | | |
| 297 | 338* | $C_\alpha$ $M_w$ $V^0$ | | | | |
| 307 | | | 352 | $pH_i$ $K^0$ | | |

**Table S4**. Sites under strong positive selection ($p < 0.001$) on the three vertebrate Hh paralogs, according to TreeSAAP. Site numbering refers to the Homo sapiens sequences and an asterisk marks the sites which were detected as under negative selection with SLAC and FEL. A legend for properties symbols is presented on table S5. (*Continuation*)

| **Codon Alignment Position** | ***Dhh*** | | ***Ihh*** | | ***Shh*** | |
|---|---|---|---|---|---|---|
| | Sites | Properties | Sites | Properties | Sites | Properties |
| 308 | | | 353 | $pH_i$ | | |
| 310 | | | | | 371 | $pH_i$ |
| 317 | 358 | $B_l$ $C_\alpha$ $M_v$ $M_w$ $V^0$ $\mu$ $H_t$ | | | | |
| 323 | | | 368 | $K^0$ | | |
| 324 | | | | | 385 | $B_l$ $M_w$ $M_v$ $V^0$ $\mu$ $H_t$ $C_\alpha$ |
| 327 | | | 372 | $B_l$ $R_a$ $C_\alpha$ $M_w$ $V^0$ $M_v$ $\mu$ $H_t$ | | |
| 328 | | | 373 | $B_l$ $R_a$ $C_\alpha$ $M_w$ $V^0$ | | |
| 329 | | | | | | |
| 330 | | | | | 391* | $R_F$ |
| 331 | | | | | 392* | $pH_i$ |
| 332 | | | | | 393 | $P_r$ |
| 334 | 378 | $P_r$ | 374 | $P_r$ | 430* | $K^0$ |
| 339 | | | | | 437 | $pH_i$ |
| 342 | | | 389* | $H$ | | |
| 345 | 391 | $C_\alpha$ $M_w$ $V^0$ | | | | |
| 350 | 396 | $B_l$ $C_\alpha$ $M_v$ $M_w$ $V^0$ $\mu$ $H_t$ | 397 | $K^0$ | | |
| 351 | | | | | 449 | $pH_i$ |

**Table S5**. TreeSAAP properties and their categorization.

| ***Category*** | ***Property*** | ***Symbol*** |
|---|---|---|
| *Chemical* | Buriedness | $B_r$ |
| | Chromatographic index | $R_F$ |
| | Equilibrium constant | $pK'$ |
| | Hydropathy | $H$ |
| | Isoelectric point | $pH_i$ |
| | Long-range nonbonded energy | $E_l$ |
| | Normalized consensus hydropathy | $H_{nc}$ |
| | Polar requirement | $P_r$ |
| | Polarity | $P$ |
| | Refractive index | $\mu$ |
| | Short-range and medium range nonbonded energy | $E_{sm}$ |
| | Solvent accessible reduction ratio | $R_a$ |
| | Surrounding hydrophobicity | $H_p$ |
| | Thermodynamic transfer hydrophobocity | $H_t$ |
| | Total nonbonded energy | $E_t$ |
| *Other* | Composition | $C$ |
| | Molecular weight | $M_w$ |
| | Power to be at the c-terminal | $\alpha_c$ |
| | Power to be at the n-terminal | $\alpha_n$ |
| *Structural* | Alpha helical tendencies | $P_\alpha$ |
| | Average # surrounding residues | $N_s$ |
| | Beta structure tendencies | $P_\beta$ |
| | Bulkiness | $B_l$ |
| | Coil tendencies | $P_c$ |
| | Compressibility | $K^0$ |
| | Helical contact area | $C_\alpha$ |
| | Mean rms fluctuation displacement | $F$ |
| | Molecular volume | $M_v$ |
| | Partial specific volume | $V^0$ |
| | Power to be at the middle of the alpha helix | $\alpha_m$ |
| | Turn tendencies | $P_t$ |