MOLECULAR EVOLUTION OF SMALL PEPTIDE HORMONES AND THEIR RECEPTORS: THE CASE OF RELAXIN AND INSULIN-LIKE PEPTIDE SIGNALING SYSTEMS IN DEUTEROSTOMES

by

SERGEY YEGOROV

A thesis submitted to the Faculty of Graduate Studies

in partial fulfillment of the requirements for the

Master of Science degree

Department of Biology

Master of Science in Bioscience, Technology and Public Policy Program

University of Winnipeg

Winnipeg, Manitoba, Canada

October 2011

ABSTRACT

Relaxin family peptides are a diverse family of signalling molecules that play important roles in the regulation of reproductive and neuroendocrine processes in vertebrates. The signalling of relaxin peptides is mediated by G protein-coupled receptors of two distinct classes, small peptide receptors and leucine-rich repeat-containing receptors. The origins and evolutionary history of both relaxin family peptides and their receptors have been a matter of debate for several reasons, among which the small size of peptide molecules (~ 60 aa, often providing insufficient information for phylogenetic reconstructions) and low coverage of vertebrate taxa by functional studies have been most prominent. In this study, I combined traditional bioinformatic approaches with ancestral genome reconstructions to reassess some of the debated aspects of the evolution of relaxin peptides and their receptors. To cover a broad range of taxa, I performed thorough data mining of the focal genes in 29 publicly available genome databases of both vertebrate and invertebrate deuterostomes. Ancestral genome reconstruction-based analyses provided clear evidence for the strong influence of whole genome duplications (WGDs) on the diversification of the relaxin signaling system from a tripartite system, consisting of one hormone and two receptor-encoding genes in the vertebrate ancestor, to the present day system. The results presented here indicate that relaxin family peptide systems are more diverse than previously thought, in particular with respect to the number of genes present in different vertebrate lineages. Based on the duplication model presented here, I propose that the ancestral tripartite signaling system had a dual function which was partitioned after the first round of WGD such that two sets of ligand-receptor pairs subfunctionalized into predominantly neuroendocrine- or reproductive-focused functions. My further analyses indicated that the suite of four ligand-receptor pairs common to the majority of modern mammals and teleosts, and already present in their gnathostome ancestor, have mostly evolved under similar selection pressures, suggesting a similar function of the genes across vertebrates. However, there are some distinct patterns of selection and evidence of differential codon-specific selection in mammals versus teleosts. Lastly, the reconstruction of the ancestral states of relaxin family peptides demonstrates how the ancestral structure shared by all four peptides has changed over time and in different lineages to acquire the specific structural characteristics of the peptides that we are familiar with today. Overall, by creating an evolutionary framework for future analyses, this study should facilitate further investigation into the properties of relaxin family peptides and their receptors.

ACKNOWLEDGEMENTS

"The only place where success comes before work is in the dictionary"
Donald Kendall

This thesis is a result of a two-year long venture, the successful completion of which could not be possible without the generous contributions of a number of people who helped me in different ways to achieve my goals. I wish to thank these people here.

Sara Good: my "unofficial" supervisor, without whose involvement the world would not have seen this study. Sara's infinite knowledge of biology and unrivaled experience as a teacher and researcher, which she has kindly shared with me, played a key role in my successful career as a master student. Perhaps, more important than anything else for me was Sara's cheerfulness, generosity and patience with which she treated me at difficult times. Sara, thank you for all you have done to make things happen!

Jens Franck: my official supervisor, who kindly adopted me as a student, together with my project. Dr. Franck patiently led this project to its logical completion. Dr. Franck, thank you for your wisdom, expert guidance and real Canadian relaxed attitude when dealing with problems!

Kevin Campbell and Murray Wiegand: the members of my thesis evaluation committee, who had a great impact on the outcome of my work. Your ideas and critique finalized the shape of the project and directed my effort on the right path. Thank you for bearing with all the misplaced figures, references and appendices in the draft of this thesis!

Jan Bogerd: the external member on my thesis evaluation panel, who also initiated the ongoing collaboration between our labs in the Netherlands and Canada. Jan, it has been a great pleasure to work with you on the relaxin project, thank you for your expertise and constant support!

Sabine Hombach-Klonisch and the members of her lab: the people who taught me cell culture, Western blotting and laboratory discipline. Although I did not end up using any of these techniques in my current project (not even the discipline!), I gained valuable insight into the state of affairs in the relaxin community and the questions to be resolved. Sabine, a sincere thank you for all the knowledge you kindly conveyed to me!

Shirin Farag and the UofW Judo club: where I gained confidence in my own self and had a chance to take long-awaited breaks away from the computer. Sensei and my fellow judokas, thank you for your friendship and trust in me!

The members of Dr. Good and Dr. Franck's labs: it was a great pleasure to work with you. Thank you for all your assistance and friendship!

Department of Biology at the University of Winnipeg: where I performed all the work that appears in this thesis. Thank you everyone who contributed in some way to my experience as a master student.

Yumgiskor Holding: the company I have been working for since 7 July, 2011. My dear colleagues, thank you for your understanding and cooperation during the times when I was occupied with the writing of this thesis (which, of course, only happened after work hours) and could not pay enough attention to your needs.

I am most grateful for the financial support that I was provided with by the following funding agencies: the Province of Manitoba (Manitoba graduate student award), University of Winnipeg (Graduate studies award) and NSERC (Discovery grant to Sara Good).

Finally, I wish to very specially thank my mother for her constant support and eagerness to listen to my explanations on the importance of whole genome duplications in the diversification of vertebrate genes. At last I have an answer to her everlasting question (repeated on a weekly basis): "Son, are you done with your studies yet?"- "Undoubtedly, mom!"

TABLE OF CONTENTS

LIST OF FIGURES

BACKGROUND

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

CHAPTER 1

ACRONYMS

aa	amino acid
AGTR	Angiotensin receptor
AIC	Akaike information criterion
BLAST	Basic local alignment search tool
BLASTn/BLASTp	Nucleotide/Protein BLAST
BKRB	Bradykinin receptor B
CCR	Chemokine receptor
sCLG	Chordate linkage group
CNS	Central nervous system
dilp	Drosophila insulin-like peptide
dN	Non-synonymous amino acid changes
dS	Synonymous amino acid changes
DNA	Deoxyribonucleic Acid
ECL	Extracellular domain
EST (database)	Expressed sequence tags (database)
FSHR	Follicule stimulating hormone receptor
GAC	Gnathostome ancestor chromosome
GPCR (GPR)	G-protein coupled receptor
Hsap	Human chromosome
HPG (axis)	Hypothalamic-pituitary-gonadal (axis)
ICL	Intracellular domain
ilp	insulin-like peptide (invertebrate)
INS	Insulin
INSL	Insulin-like peptide (vertebrate)
Ins-l	Insulin-like peptide (Ciona spp.)
IGF	Insulin-like growth factor
JTT (model)	Jones, Taylor, Thornton (model)
LDLa	Low density lipoprotein a
LH	Luteinizing hormone

LGR	LRR-containing GPCR
LRR	Leucine-rich repeats
Mb	Mega base (pairs)
ML	Maximum likelihood
MYA	Million(s) years ago
NCBI	National center for biotechnology information
PHI-BLAST	Pattern-hit initiated BLAST
rfpl	Relaxin family peptide-like
RLN	Relaxin
RLN/INSL (peptides)	Relaxin and insulin-like (peptides)
RXFP	Relaxin family peptide receptor
(1/2/3) R	Round of whole genome duplication
SALPR	Somatostatin and angiotensin-like peptide receptor
SSD	Small-scale duplication
TM	Transmembrane
VAC	Vertebrate ancestor chromosome
WGD	Whole genome duplication

GENE/PROTEIN NOMENCLATURE

The following gene/protein naming guidelines were followed in this thesis:

- All gene- and/or mRNA-related names are italicized (e.g. *RLN*), proteins are not italicized (e.g. RLN);
- All mammalian gene/protein names are written in capitals (e.g. INSL4), other vertebrate genes/proteins are in lower case (e.g. rln3);
- When a group of genes/proteins is described, it is written in capitals if the discussed group contains mammalian sequences (e.g. INSL5 Tetrapods), otherwise it is written in lower case (e.g. insl5 Teleosts);
- The names for ancestral genes/proteins are in mixed case (e.g. Rln) and may have a prefix "Anc" (e.g. AncRln)

BACKGROUND

Relaxin family hormones: what are they?

Among the many types of mammalian signaling molecules, and particularly hormones, relaxin family peptides may well be one of the least familiar to the public. Although the first member of this hormone family was discovered shortly after insulin (Hisaw 1926), progress in the relaxin research field has until recently lagged behind the advances made in, for instance, our understanding of insulin (INS) and insulin-like growth factors (IGF), which are closely related to relaxin peptides both through their structural similarities and shared evolutionary origins. Notwithstanding, recent progress in molecular biology, the availability of whole genome sequence data and the emergence of novel research tools such as bioinformatics, have recently contributed enormously to our knowledge about the signaling systems regulated by the relaxin hormones and their receptors.

We now know that the repertoire of human relaxin family hormones consists of 7 molecules: three Relaxin-like (RLN) and four Insulin-like (INSL) peptides. Notably, the subdivision of relaxin hormones into 2 classes (RLN and INSL) is based primarily on early structural data and the order of their discovery (Sherwood 2004), and it does not take into account their evolutionary origins or physiological differences. Here, for simplicity, relaxin peptides will be introduced from the genomic perspective, i.e. according to the genes that encode them in the human genome.

The RLN/INSL-encoding genes in humans, as in many other tetrapods, exist in 4 distinct clusters. The largest cluster is made up of 4 loci: *RLN1, RLN2, INSL4* and *INSL6*, situated in tandem on human chromosome 9 (Hsap-9). This cluster of genes, “the *RLN*-cluster” arose from multiple local gene duplications that occurred in the ancestor of placental mammals (Wilkinson et al. 2005b). The products of these genes are primarily associated with reproductive functions, such as the relaxation of uterine musculature and of the pubic symphysis during labor (RLN1 & RLN2), the progression of spermatogenesis (INSL6) and possibly trophoblast development (INSL4) (Millar et al. 2005). At the same time human RLN2 (equivalent to RLN in other mammals), has more general functions than RLN1, and participates in collagen metabolism and angiogenesis in both reproductive and non-reproductive tissues. The other three *RLN/INSL* genes exist as single loci in two linkage groups: *RLN3* (Hsap-19), *INSL3* (Hsap-19, 3.8 Mb apart from *RLN3*) and *INSL5* (Hsap-1).

The physiological action of RLN and INSL3 has been quite well studied in human and mouse, but the functions of INSL5 and RLN3 are relatively unexplored, especially outside of placental mammals. Both RLN3 and INSL5 are thought to play important roles in neuroendocrine regulation. In the case of INSL5 this hypothesis is based on its expression (and also co-expression with its receptor) in the central nervous system (CNS), intestine and lymph nodes. At the same time, RLN3 is predominantly localized in the brain and locally affects selected regions of the CNS, such as those responsible for the sense of appetite and stress regulation. Moreover, it has been shown that RLN3 stimulates the hypothalamic-pituitary-gonadal (HPG) axis and hence affects the levels of luteinizing hormone (LH) in the blood (McGowan et al. 2008). LH is essential for normal functioning of both male and female reproductive systems, where it triggers ovulation (in females) or stimulates production of testosterone (in males).

Finally, *INSL3* is highly expressed in gonads, where it enhances the survival of germ cells (Kawamura et al. 2004). In males, the levels of INSL3, secreted by testicular Leydig cells, are significantly higher than those in females, in which the hormone is probably produced by ovarian theca cells. Thus, in both males and females, INSL3 is implicated in germ cell survival, but may play a more central role in cell development in males (Kawamura et al. 2004). Interestingly, the INSL3 secreted by fetal Leydig cells is a crucial factor regulating the descent of testicles in some placental mammals (Feng et al. 2009), such as mice, but this role of INSL3 in humans has not been fully established.

Structure of RLN/INSL peptides

Most members of the human relaxin peptide family in their mature form consist of two ~30 amino acid long peptide chains, named “A” and “B”-chains, interlinked by disulfide bridges. The exception is INSL4, which contains an insertion and is longer than the other peptides. The double-chain structure, characteristic of members of the insulin superfamily, is a result of post-translational modification of the *RLN/INSL*-gene products (Figure B1).

The receptors of relaxin family hormones

The endogenous signals transmitted by Relaxin family hormones reach their target tissues by different means. While some of these peptides (e.g. INSL3 and RLN3) are believed to be mostly paracrine (Kawamura et al. 2004) and exert their effect primarily on the cells surrounding the hormone producing site, others (e.g. RLN2) are of endocrine nature and may affect multiple tissues by being transported in the blood stream (Sherwood 2004).

Regardless of how a hormone reaches its target cell, it will eventually have to bind a specific receptor to initiate communication with the cell. The receptors for most of the human relaxin hormones have been established and only INSL4 and INSL6 remain “orphan” (Kong et al. 2010).

The receptors for the RLN/INSL peptides are collectively called “Relaxin family peptide receptors (RXFPs)”. RXFPs were discovered relatively recently, and, somewhat surprisingly, have been found to regulate signaling pathways that differ significantly from those employed by the insulin and IGF receptors. There are two distinct families of RXFPs, all of which are cell membrane-associated and coupled to G-proteins (hence they are known as G protein-coupled receptors [GPCRs]). All GPCR-type receptors are embedded in the plasma membrane with the aid of seven transmembrane (7TM) spanning helices and they interact with G proteins via their intracellular parts (Fredriksson et al. 2003). Although there is evidence that some relaxin hormones may also be able to interact with glucocorticoid-type nuclear receptors, which are found floating freely between the cytoplasm and nucleoplasm (Dschietzig et al. 2006), in this study these receptors will not be covered due to limited information available about them.

The most thoroughly studied RXFP family in humans consists of two members, RXFP1 and RXFP2, which are closely related to the receptors of glycoprotein hormones (Figure B2). This class of receptors has a distinctly large domain consisting of a number of leucine-rich repeats (LRR) on the N-terminus, which *in vivo* is situated extracellularly and plays a key role in ligand recognition (Halls et al. 2007).

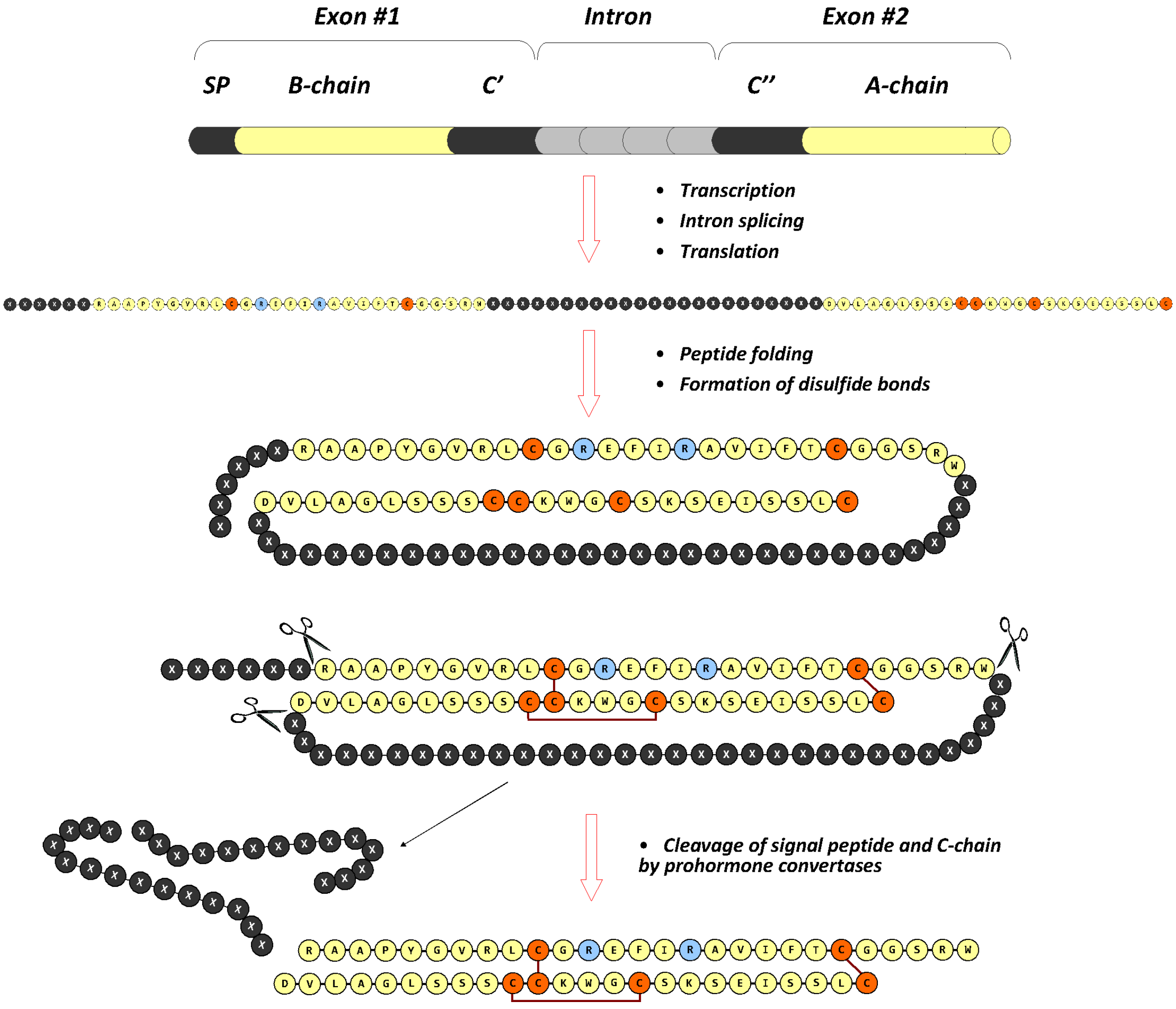

Exon #1
Intron
Exon #2
SP
B-chain
C'
C''
A-chain
Transcription
Intron splicing
Translation
Peptide folding
Formation of disulfide bonds
Cleavage of signal peptide and C-chain by prohormone convertases
R A A P Y G V R L C G R E F I R A V I F T C G G S R W
D V L A G L S S S C C K W G C S K S E I S S L C
X

Figure B1. Structural characteristics of relaxin family peptides.
RLN/INSL genes contain 2 exons and a single intron. The mature *RLN/INSL* mRNA is translated into a preprohormone with a signal peptide (SP) that is co-translationally removed yielding a prohormone (consisting of domains named B-, C- and A-chains), which is further processed by prohormone convertases to produce a double-chained mature peptide, which is ready to be released from the cell.

The endogenous ligands of RXFP1 and RXFP2 are RLN1-2 and INSL3 peptides respectively. The other class of human RXFP receptors also has only two members, RXFP3 and RXFP4. These receptors are both structurally and functionally distant from RXFP1/2s, and are more related to the receptors of small peptides, such as somatostatin, angiotensin and bradykinin (Figure B3). The endogenous ligands of RXFP3 and RXFP4 are RLN3 and INSL5 respectively. Notably, although RXFP1-RLN2, RXFP2-INSL3, RXFP3-RLN3 and RXFP4-INSL5 are well established endogenous receptor-ligand pairs in human, the ligands overlap in their abilities to bind the receptors *in vitro*. For instance, RLN2 can bind RXFP1 and RXFP2, and RLN3 can bind RXFP1, RXFP3 and RXFP4 (Figure B2). The promiscuous interaction of RLN/INSL with the two diverse classes of RXFP receptors is unique among GPCR ligands and its evolutionary significance has yet to be clarified (Gloriam et al. 2009).

Interestingly, before the discovery of RXFP-type receptors (which occurred only several years ago), it was thought that the receptors of RLN/INSL peptides should be similar to those of their closest relative, INS and IGF, based on the general rule that signaling molecules of similar structure pair with similar receptors (Halls et al. 2007). As it turned out, the receptors of RLN/INSL share little with the tyrosine kinase-type INS/IGF receptors. Moreover, the latter do not express any binding affinity toward relaxin-like peptides and appear evolutionarily very distant to both subgroups of RXFPs (Halls et al. 2007).

To this extent I have introduced the members of the RLN/INSL-RXFP signaling system as they exist in humans. Relatively little is currently known about the biology of this system in non-human vertebrates.

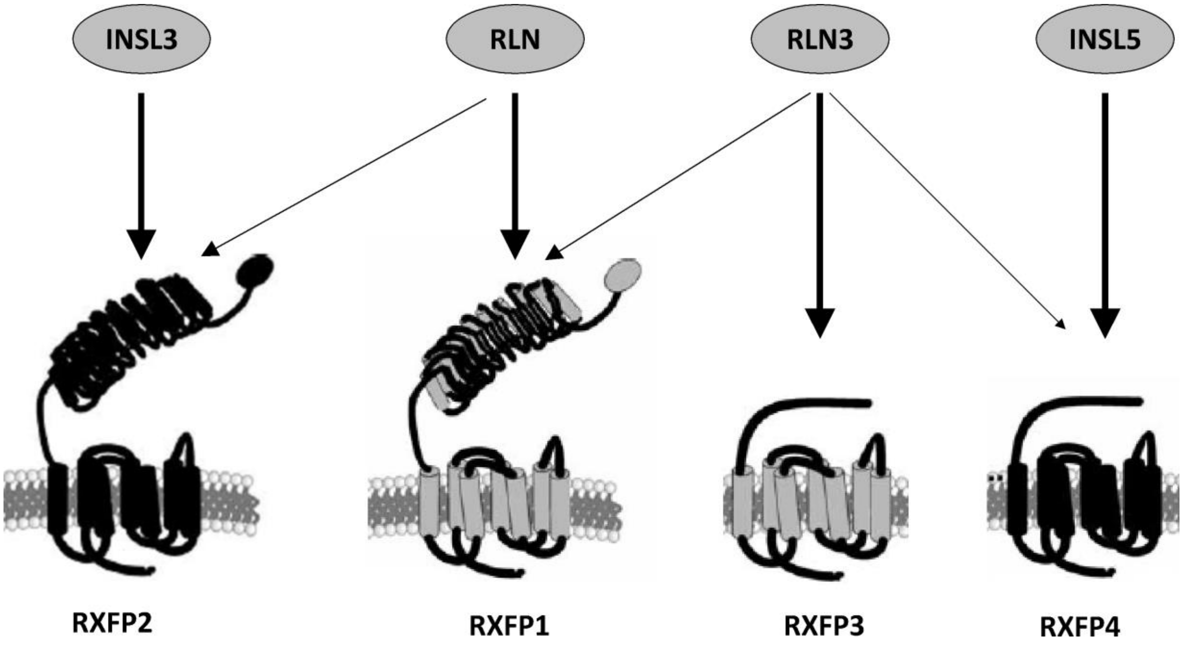


Figure B2. Receptor-ligand interactions among relaxin family peptides and their receptors in human.
RLN/INSL peptides are promiscuous in their interactions with their receptors. Thus, although each peptide is believed to endogenously interact with a single specific receptor (shown with thick arrows), RLN (both RLN1 and RLN2 in human) and RLN3 show affinity toward additional receptors (shown with narrow arrows). Note that structurally RXFP1/2 are different from RXFP3/4 receptors owing primarily to the longer and more complex N-terminus, which consists of the LRR and LDL domains (see text). The 7TM domains are colored in either black or grey to distinguish between receptors of same class. Based on the functional affinities of relaxin peptides and their receptors as described in Halls et al. (2007). *The receptor images used with permission from John Wiley & Sons, Inc.*

Figure B3. The receptors of relaxin family peptides belong to two evolutionarily distant groups of Rhodopsin class GPCR receptors.
The receptors of relaxin family peptides belong to two evolutionarily distant groups of Rhodopsin class GPCR receptors. RXFP1/2 (also known as LGR7 and LGR8, green boxed) receptors are classified as subclass delta (δ) receptors and are related to glycoprotein receptors (e.g. FSHR) and other LRR-containing GPCRs (LGR4-6). RXFP3/4 (for which the alternative name is SALPR and GPR100, green boxed) receptors are in the gamma (γ) subclass of Rhodopsin GPCRs, which also includes bradykinin (BDKRB), angiotensin (AGTR) and other receptors with small peptide ligands (e.g. chemokine receptors, CCR). SALPR: Somatostatin and Angiotensin-Like Peptide Receptor; GPR: G protein-coupled receptor. Adopted from Fredriksson et al. (2003), *with permission from ASPET Journals Department.*

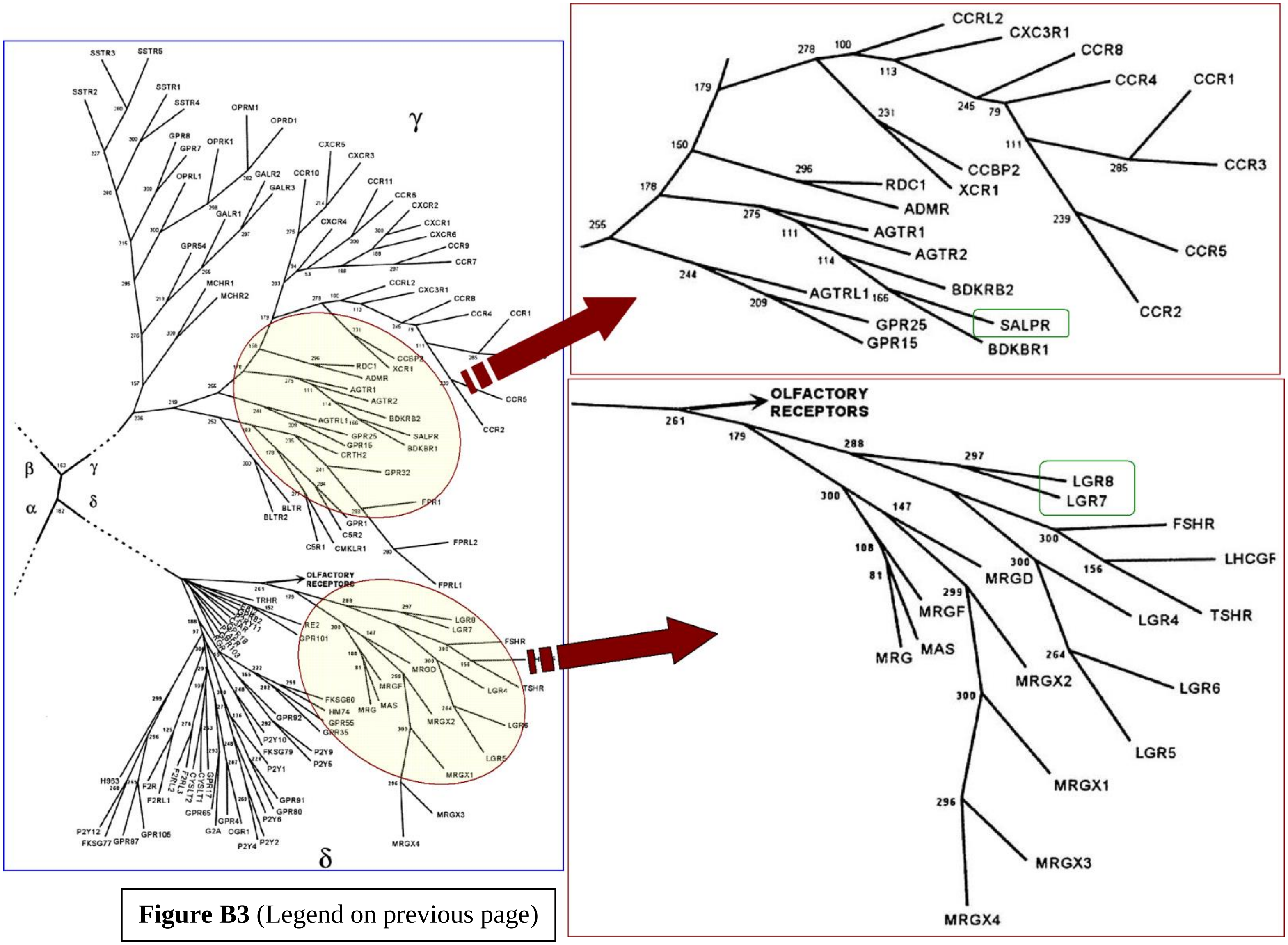


Figure B3 (Legend on previous page)

The molecular evolution of the RLN/INSL-RXFP ligand-receptor systems: current opinions

The emergence of new signaling pathways that coordinate novel functions in biological systems is an important part of the evolutionary process. At the molecular level, the appearance of new functional networks is often associated with the expansion of gene families through gene duplications and ensuing neo- or subfunctionalization (Lynch and Conery 2000) and the relaxin family peptide and their receptor systems do not seem to be exceptional in this regard. However, the roles of gene duplication, gene loss, gene retention, and selection in the diversification of these signaling systems have only recently attracted the attention of researchers. Even then, studies on the evolutionary history of the family have predominantly focused on the evolution of ligands, in most cases avoiding the receptors, and have not addressed the question of how the RLN/INSL-RXFP system evolved as one biological unit throughout evolutionary history.

Previous work identified that four distinct *RLN/INSL* loci, *RLN*, *RLN3*, *INSL3* and *INSL5,* were present prior to the divergence of teleosts and tetrapods (Good-Avila et al. 2009, Park et al. 2008). The three additional relaxin family genes in humans and apes (*RLN1*, *INSL4* and *INSL6*) are the result of recent local duplications of the more ancient *RLN2* locus. It has been proposed that the peptide family and its receptors originated from the ancestral RLN3-RXFP3 system and that the INSL5-RXFP4 pair arose from this ancestral ligand-receptor pair as a result of a duplication of both the ligand and receptor genes (Wilkinson and Bathgate 2007). The other two members of the relaxin family, RLN and INSL3, were hypothesized to also have arisen from RLN3 (Park et al. 2008, Wilkinson and Bathgate 2007) and to subsequently have recruited a new pair of receptors, RXFP1 and RXFP2, that are associated with reproductive processes specific to placental mammals (Wilkinson and Bathgate 2007).

More recently, it was hypothesized that the diversification of relaxin family genes in vertebrates occurred as early as 550 million years ago (MYA) through the two rounds of whole genome duplication (2R, WGD) that took place in early chordate evolution (Hoffmann and Opazo 2011). In addition, it was proposed that the family arose via duplication of the ancestral insulin/IGF locus in the common ancestor of urochordates and modern vertebrates and originally functioned using an insulin receptor-related receptor, prior to switching to a RXFP-type GPCR (Olinski et al. 2006b). Genomic analyses of protochordates (Holland et al. 2008, Olinski et al. 2006a) further indicated that the chromosomal regions hosting multiple *insulin-like peptide (ilp)* genes in a tunicate, *Ciona intestinalis,* and amphioxus, *Branchiostoma floridae*, share equal amounts of synteny with both vertebrate relaxin and insulin families of genes. Despite the interest in the evolutionary history of the family, no clear method has been available to resolve the origins of both RLN/INSL peptides and their receptors and the knowledge of the evolutionary history of this ligand-receptor pair has remained equivocal.

Evolution of gene families through gene duplication

The fate of duplicated genes: gene loss and gene retention

Gene duplication is considered to be one of the major forces of molecular evolution (Lynch and Conery 2000). Genes may duplicate as a consequence of polyploidization, i.e. doubling of the complete DNA set of already diploid cells (otherwise known as whole genome duplication, WGD), or as a result of an amplification of a short stretch of a chromosomal segment containing one or several genes (small-scale duplication, SSD). Following a duplication event and depending on the mechanism involved (WGD or SSD), the resulting duplicates may experience different evolutionary fates. For instance, while the majority of duplicates are essentially lost from the genome due to accumulation of missense mutations in their coding sequence (“pseudogenization”), others may be retained and even acquire novel functions. The process of gene loss and gene retention plays a key role in shaping metabolic networks across lineages. Thus, studying the retention patterns of gene families in a broad range of taxa may be helpful for understanding the co-evolutionary processes among different gene families. Interestingly, because polyploidization has the power to amplify entire metabolic networks of genes while conserving the individual gene niches, there is a general tendency for WGD-duplicates to be retained more frequently than for their SSD-counterparts, which find themselves “looking for new jobs” in the unchanged genomic environment (Hakes et al. 2007). Moreover, WGDs are thought to have had a strong influence on the evolution of ligand-receptor systems because of their ability to duplicate both ligands and their receptor genes (assuming that both ligand and receptor are gene-encoded proteins) avoiding the necessity of genetic linkage (Huminiecki and Heldin 2010).

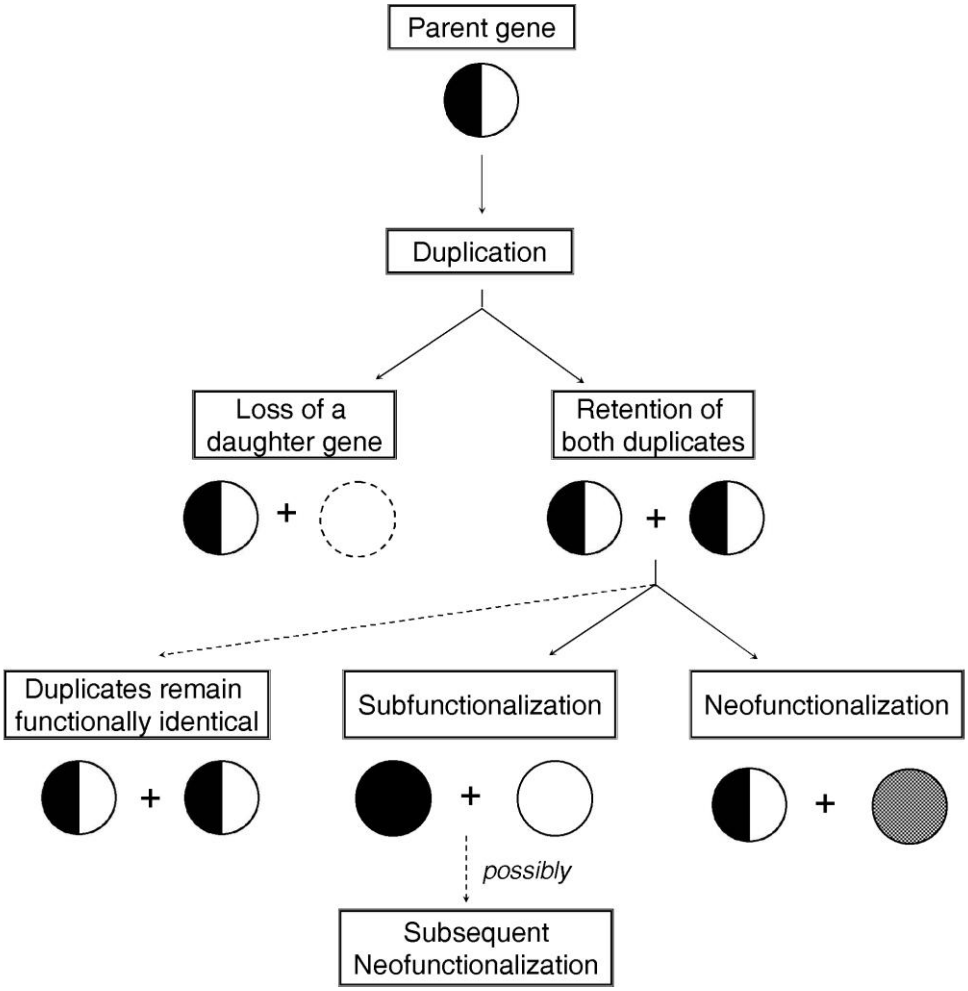

Parent gene
Duplication
Loss of a daughter gene
Retention of both duplicates
+
Duplicates remain functionally identical
Subfunctionalization
Neofunctionalization
possibly
Subsequent Neofunctionalization

Figure B4. A hypothetical perspective on the fate of genes arising via duplication. Following a gene duplication event, the parent gene gives rise to two daughter genes, which are functionally identical to their progenitor and to each other. Depending on different factors (see text) both daughter genes may either be retained in the genome or one of them can be lost. In the event of loss of one of the descendant genes, the retained duplicate will keep performing the function of the original parent gene. Alternatively, in the event of retention of both daughter genes, at least three possible scenarios of gene evolution could be expected to take place: 1) both duplicates remain functionally identical, which leads to an increase in the amount of produced protein product; 2) subfunctionalization, where the ancestral gene function (black-and-white) is split between the two duplicates (one becomes white, the other black), thus reducing their functional overlap; 3) neofunctionalization, where one of the duplicates acquires a function which is not directly related to that of the ancestral gene (neither white, nor black), while the other duplicate retains the function of the ancestral gene (black-and-white). Over long evolutionary time scales it may be difficult to discern the difference between sub- and neofunctionalization processes. Dashed arrows depict less frequently occurring, while bold arrows represent more frequently occurring processes.

The fate of duplicated genes: sub- and neofunctionalization

The retention of novel duplicate genes is driven by the needs of the genome and organism at large: “novice” genes may be recruited to perform one of the functions of the parent gene (subfunctionalization) or to fill a niche previously unoccupied by any other genes (neofunctionalization) (Figure B4). Whereas subfunctionalization may lead to an increasing complexity of gene regulation, neofunctionalization expands the functional boundaries of the organism’s molecular machinery (Force et al. 1999). It is also thought that subfunctionalization is a more frequently occurring process, because splitting a parent gene’s function into two components is considered a more likely event than the adoption of a completely novel function by one of the daughter genes. However, even though subfunctionalization may occur first, as time passes and the two genes evolve independently, one of the subfunctionalized genes may acquire novel functions, i.e. become neofunctionalized (Force et al. 1999).

In this regard, an illustrative example of duplication followed by subfunctionalization is a two gene system composed of a ligand and its receptor, which duplicates and gives rise to a four gene system (2 ligands and 2 receptors, Figure B5).

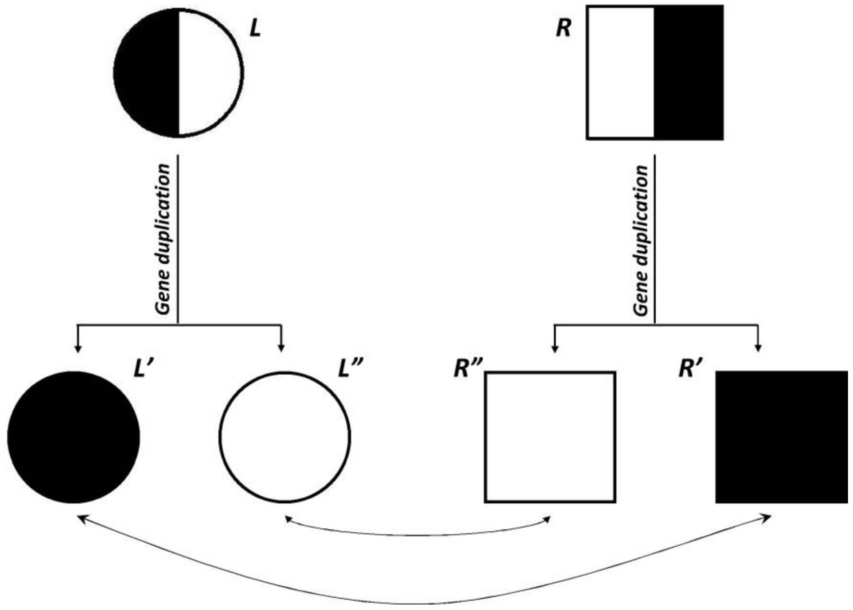


Figure B5. Subfunctionalization in a two component receptor-ligand system. The duplication of a ligand-receptor system encoded by 2 genes, the ligand gene (*L*) and the receptor gene (*R*)**.** After gene duplication, both daughter ligand (*L'* and *L"*) and receptor genes (*R'* and *R"*) subfunctionalize. The ancestral "black-and-white" function is split among the daughter genes, such that L'-R' (black) and L"-R" (white) become novel ligand-receptor pairs (shown with arrows).

At first, the ligand-receptor interactions in this system will be promiscuous, i.e. due to their structural identity both receptors will be capable of interacting with both ligands. Then, unless the redundancy of the system has a selective advantage, the system will either split its original function between its two subcomponent receptor-ligand pairs or it will rid itself of redundancy by means of gene loss.

Exploring gene evolution: bioinformatics tools

With the availability of whole genome sequence data from a variety of vertebrates, the task of characterizing the relaxin family peptide system in non-human species has become more achievable using traditional bioinformatics approaches. Multiple authors have used methods in bioinformatics to address unclear aspects in the evolution of relaxin family peptides, although not all of them successfully resolved their questions (see, for example, the discussion of Hoffman et al.'s (2011) work in Chapter 2). Therefore, a major goal of this thesis was to further demonstrate the utility of synteny, phylogeny and selection analyses in deciphering the mechanisms involved in the diversification of the relaxin peptides. A second aim was to focus on the evolution of their receptors, a subject that has received much less attention using other methodologies such as analyses of selection and ancestral state reconstruction.

Data mining

Most, if not all, bionformatic studies begin with data mining, which involves searching various databases with the goal of obtaining the protein or DNA sequences of interest for further analyses. In this study, the data mining for the relaxin family peptides and their receptor genes was carried out using publicly available databases of both raw and processed genomic data. The novelty of this study is largely attributable to the novel and unannotated sequences that were for the first time analyzed here thanks to the multiple publicly available sequenced and assembled genomes of vertebrates and invertebrates. These databases are Ensembl and Pre-Ensembl (maintained by the European Bioinformatics and Welcome Trust Sanger Institutes), GenBank in NCBI (National Center for Biotechnology Information) and DOE JGI (Joint Genome Institute, USA) just to name a few (for a complete list of databases refer to supplementary materials). Albeit the degree of gene/protein annotation in some databases (for instance in the human or mouse genome assemblies in Ensembl or NCBI) allows for gene identification with the easiness of modern internet search engines, this is not the case with most other genome assemblies, where the useful coding sequences of genes may be “hidden” in the midst of “junky” non-coding material of chromosomes. To aid with gene identification and comparison, this study employed various software algorithms, such as the BLAST package available through NCBI, along with syntenic and phylogenetic analyses, which are briefly described below.

Phylogeny

The most intuitive way to determine the evolutionary relatedness of DNA or protein molecules is to compare their nucleotide (in the case of DNA) or amino acid (in the case of protein) sequences side by side. The more similar the sequences are the more confident one can be regarding the relatedness of the sequences. The results inferred from such comparisons of gene sequences are often depicted using clustering methods resulting in diagrams called “phylogenetic trees” (see Figure B3 for an example of a tree depicting the evolutionary relationships among rhodopsin class GPCR receptors). However, phylogenetic comparisons have some caveats, like any other analyses, and a major caveat relevant to this study is their inability to correctly identify the evolutionary relatedness of short genes under selection pressures. It is always a good idea, hence, to supplement phylogenetic analyses with other methodologies.

Synteny

In classical genetics, synteny describes the physical co-localization of genetic loci on the same chromosome within an individual or species. In modern bioinformatic terms synteny refers to the similarity in the genetic background among two or more genes of

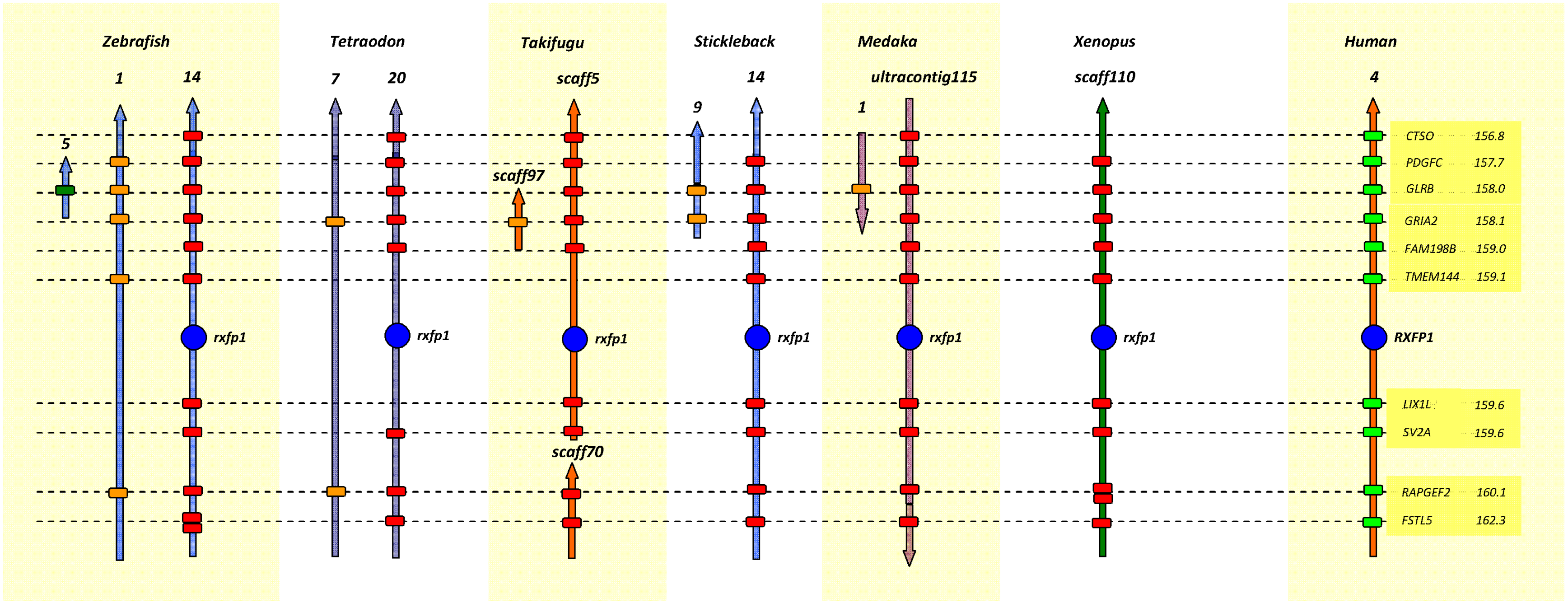


Figure B6. Diagram showing the syntenic relationships among the *RXFP1* genes of the five species of teleost fish, frog (*Xenopus*) and human.
The *RXFP1* genes across taxa have similar genetic backgrounds, i.e. they share a number of flanking genes (e.g. *TMEM144* or *SV2A*), which indicates that these *RXFP1* genes are orthologs, i.e. they are derived from one ancestral gene. Arrows depict chromosomes, boxes and circles depict genes. Gene names are shown on the right with numbers indicating their chromosomal locations in Mb. Scaff=scaffold.

interest (or focal genes). The degree of similarity (or synteny) is measured by looking at the loci flanking the focal genes and by counting the number of flanking loci shared among the genes of interest. Generally, the more syntenic similarity there is between two focal genes, the more evolutionarily related they are. Figure B4, for example, shows the syntenic relationships among the *RXFP1* orthologs of different vertebrates.

Selection analyses

Depending on the level of functional constraint experienced by proteins, they may evolve slowly and be subject primarily to purifying selection in which mutations are purged from populations. Alternatively, if the protein is subject to less functional constraint, then some mutations, especially synonymous mutations (i.e. those resulting in non-radical changes in the amino acids of the protein) may become fixed within lineages, and peptides from different lineages will exhibit greater sequence divergence caused by neutral evolution of the proteins. Lastly, some proteins, or a few codon positions within otherwise more conserved proteins, may be subject to positive selection in which an amino acid replacement becomes favoured within a lineage. Traditionally, the extent of functional constraint on a protein is measured by calculating the average number of mutations resulting in non-synonymous (dN, amino acid residue changing) to synonymous (dS, silent mutations) changes. If dN>dS, proteins are said to be subject to positive selection. If dN=dS, proteins are said to be evolving neutrally, whereas if dN<dS, they are subject to purifying selection (Hughes L. A. and Nei 1988). Additionally, more recently, tests of codon-specific positive selection have frequently been employed because positive selection frequently operates at a local scale, on select amino acids or lineages (Zhang et al. 2005). Thus, selection analyses provide yet another

perspective on the diversification of genes by giving an estimate of the rates at which orthologous and paralogous genes evolve in different lineages or within the same lineage.

Ancestral gene reconstruction

Ancestral gene reconstruction is a method that allows one, with some caveats, to study the properties of long lost genes and their products from ancestral organisms. This method makes use of the phylogenetic reconstructions of the evolutionary relationships among related genes to infer the structure of the gene(s) that gave rise to the gene family(ies) of interest. There are a number of methodological approaches to performing ancestral state reconstruction. In this work, for instance, I employ maximum likelihood (ML) methods to infer the ancestral peptide sequences of relaxin family genes at distinct points in their evolutionary history.

Brief overview of the contents of this thesis

Chapter 1: Written as an independent manuscript and consists of four sections: introduction, methods, results and discussion. Here I describe a novel method to study the evolutionary origins of both relaxin family peptides and their receptors using ancestral deuterostome genome reconstruction models. Principally, this method takes advantage of the large-scale synteny analyses performed on both vertebrate and invertebrate deuterostome genomes. By combining the information about chordate ancestral linkage groups with small-scale synteny and phylogenetic analyses, I resolved some controversial issues regarding the evolution of *RLN/INSL* genes in early vertebrates. In addition, I reconstructed the evolutionary relationships among *RXFP* genes and show, for the first time, that there are multiple duplicates of *RXFP* genes that arose independently during vertebrate evolution. Finally, I looked for evidence of the presence of *RLN/INSL-RXFP* system genes in invertebrate deuterostomes, such as protochordates and echinoderms.

Chapter 2: Here I discuss the results of a thorough bioinformatic survey of vertebrate *RXFP* genes (performed in Chapter 1) and expand upon previous analyses of *RLN/INSL* genes to show how both whole genome and small-scale duplications coupled with differential gene loss resulted in the diverse array of relaxin system genes among vertebrate lineages. In addition, I use the model of duplication of *RLN/INSL* (ligand) and *RXFP* (receptor) genes as a theoretical basis to explain the functional diversification of relaxin peptide systems through subfunctionalization in vertebrates. This chapter consists of three sections: introduction, methods and results & discussion.

Chapter 3: Written as an independent manuscript and consists of four sections: introduction, methods, results and discussion. In this chapter, I examine the influence of purifying, neutral and positive selection on the relaxin family peptides and their receptors. Co-evolutionary theory predicts that ligands and receptors with functions that are conserved across lineages, should exhibit similar levels and types and selection within and between lineages. I thus estimated the proportion of amino acids subject to different forms of selection for all ligand and receptor genes in teleost and mammalian lineages and examined whether selection has played a similar role in the distinct ligand-receptor pairs in the two groups. Then, I assessed the role of codon-specific selection on both ligand and receptor genes to assess 1) the main regions of the receptor genes that have been targets of positive selection in vertebrates and 2) if there is evidence of lineage-specific codon-selection, particularly in mammalian versus teleostean lineages. The presence of lineage-specific codon selection would suggest that positive selection might have lead to sub- or neo-functionalization of genes between lineages.

Chapter 4: Here I focus on the early state of the RLN/INSL peptides and hypothesize about their early structura l and functional evolution. Thus I reconstructed the ancestral states of all four *RLN/INSL* ohnologs. I also looked for evidence of selection on codons to assess the role of selection on *RLN/INSL* genes over evolutionary time and in distinct vertebrate lineages. This chapter consists of three sections: introduction, methods and results & discussion.

Supplementary materials:

- The detailed methods explaining the use of ancestral genome reconstructions to trace the evolutionary history of individual focal genes (Appendix A);
- All accession numbers and map locations for genes used in the study (Appendix B);
- The supplementary figures, Figures C1-C2, cited in Chapter 1 (Appendix C);
- The supplementary figures, Figures D1a and D1b, cited in Chapter 2 (Appendix D)
- The supplementary table, Table E1, with additional data for Chapter 3 (Appendix E)

CHAPTER 1: *Uncovering the origin, linkage relationships and duplication history of the relaxin family hormones and their receptors*

INTRODUCTION

Analyses of whole genome sequence data suggest that three rounds of WGD occurred and contributed immensely to the diversification of vertebrates (Abi-Rached et al. 2002, Dehal and Boore 2005, Jaillon et al. 2004); two rounds of WGD (2R) occurred in early chordate evolution, probably before the divergence of agnathans and gnathostomes (Kuraku et al. 2009), while a third round (3R) of WGD occurred only at the base of the teleostean lineage. Even though gene duplication has long been recognized as a major factor in the evolution of biological diversity (Ohno 1970, Taylor and Raes 2004), determining the evolutionary relationships among members of gene families that arose via duplication is not always easy because individual genes originated via both small-scale and whole genome duplication events, could have been modified by selection or concerted evolution, and may have experienced differential loss across lineages (Nei et al. 1997, Ohno 1970, Taylor and Raes 2004). Although the ready availability of small-scale synteny data has facilitated the determination of the orthologous and paralogous relationships among genes, and thus the factors influencing gene diversification, some aspects of gene family evolution, such as their ancient origins and the timing and kind of duplication events they underwent, continue to elude investigation and are difficult to resolve using traditional bioinformatic approaches.

Recently, large-scale synteny analyses comparing entire genomes of evolutionarily distant taxa have been employed to reconstruct the karyotypes of extinct ancestors and to look back at the events that shaped the appearance of modern genomes (Muffato and Roest Crollius 2008). Ancestral genome reconstruction models depict metazoan chromosomes as composed of segments, originating from one or more linkage groups of a distant ancestor, which became united following repeated chromosomal fission and fusion events to form the karyotypes of modern taxa. By tracing the syntenic relationships among such chromosomal segments from two or more extant taxa, it is possible to reconstruct the linkage groups of their common ancestor at the time of taxon divergence. For example, comparison of the genomes of tetrapods and teleosts allows one to infer the chromosomes of the hypothetical ~450 MY old gnathostome ancestor and to also outline the linkage groups of the ~500 MY old ancestor of all extant vertebrates (Nakatani et al. 2007).

Reconstructions of ancestral genomes in the chordate lineage are particularly interesting, because they shed light on important WGD events and the intensive karyotype rearrangements that played key roles in the evolution of the vertebrate genetic portfolio. Although it has been suggested that genome reconstructions provide principally a heuristic tool for understanding genome evolution (Muffato and Roest Crollius 2008), in this chapter I demonstrate how such models can be used to trace the evolutionary history and linkage relationships of genes, thereby giving further power to elucidate both the origin and duplication history of gene families. Although it has been shown that orthologous copies of four *RLN/INSL* genes (*RLN*, *INSL3*, *INSL5* and *RLN3*) are present in teleosts and mammals, the exact mechanisms giving rise to their diversification in non-placental vertebrates have remained elusive (see Background)

In this chapter, I employ ancestral genome reconstruction models to examine the origin and linkage relationships of RLN/INSL peptide and RXFP receptor genes, and to determine the role of WGDs in their diversification. This chapter provides evidence that WGDs played a central role, larger than previously appreciated, in the evolution of the family and suggests that the system originated in the chordate ancestor from a trio of 2 receptors with a single ligand, in which the ligand and one receptor were initially linked. In addition, this chapter also sheds light on the origin of the gene ancestral to all members of relaxin family genes (*AncRln-like*) in protochordates and echinoderms. Since *AncRln-like* is believed to have arisen from the duplication of the Insulin locus, I discuss the evolution of the insulin-relaxin superfamily as a whole in deuterostomes. I find support for the hypotheses generated from the ancestral genome reconstruction models by using traditional small-scale synteny analyses and phylogenetic reconstructions performed on a broad repertoire of focal genes, and ultimately show the broad utility, with some caveats, of incorporating ancestral genome reconstruction data for understanding the evolution of gene families.

METHODS

Tracing of the duplication history of RLN/INSL, INS/IGF and RXFP genes

Detailed methods used to trace the evolutionary history of genes are provided in Appendix A. A brief overview of the procedure is given here: first, using their exact map positions, I mapped the *RLN/INSL*, *RXFP* and *INS/IGF* genes found in human, medaka and chicken to their corresponding chromosomal segments. These chromosomal segments were then matched to the linkage groups in ancestral genomes primarily according to Nakatani et al.'s (2007) model, but I also invoked other vertebrate genome reconstructions (Kasahara et al. 2007, Kemkemer et al. 2009) as needed. Finally, I compared the results obtained for each of the three taxa to resolve the positions of the focal genes at consecutive stages of the vertebrate genome evolution. Where discrepancies arose and the genes reported as "orthologous" were traced to different ancestral linkage groups, I performed small-scale synteny analyses (details below) to clarify the relationship of individual genes among taxa.

Identification of RLN/INSL and RXFP(-like) sequences across vertebrates

All annotated *RLN/INSL* and *RXFP* coding sequences with their genomic positions were retrieved from the Ensembl v.60 database (http://ensembl.org) for 13 mammals (11 placentals, opossum and platypus), three reptiles (anole lizard, chicken and zebrafinch), two amphibians (clawed frog and edible frog) and five teleosts (Tables B1-B5 in Appendix B). The annotated sequences for rhesus monkey were obtained from NCBI (http://ncbi.nlm.nih.gov/gene). When multiple splice variants were available, the longer variant was chosen, unless shorter variants had been confirmed to be functional. For five placental species, the *RLN* locus was found to be duplicated 1-5 times (Table B2), of those only one gene was retained for phylogenetic analyses.

Using the more or less complete sets of human, mouse, zebrafish and medaka sequences as reference, I performed searches of the databases at both Ensembl and NCBI to look for unannotated and/or yet unidentified genes in other tetrapods and teleosts using the NCBI BLAST package (Altschul et al. 1997). Additionally, to either confirm the identity of sequences obtained using the above procedure or to search for other difficult to identify genes, I searched the genomic regions syntenic to previously determined human/teleost *RLN/INSL* (Good-Avila et al. 2009) or *RXFP* genes in Ensembl by using the Genscan tool or the MIT Genscan server (http://genes.mit.edu/GENSCAN.html) in combination with the conserved-domain search tool (http://ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), or by blasting the entire syntenic regions via BLASTn in NCBI with a *RLN/INSL* query. The synteny analysis for the *RXFP* genes was done using either the Genomicus v.60.01 server (http://dyogen.ens.fr/genomicus-60.01/cgi-bin/search.pl), the appropriate Ensembl tools and/or manual identification of orthologous regions through subjecting genes to BLASTp at NCBI.

RXFP-type and Insulin-Relaxin superfamily genes in pre-2R taxa

Three *ilp* sequences were retrieved from GenBank for the tunicate *C. intestinalis* and used to perform additional searches on the *C. intestinalis* and *C. savignyi* proteomes using PHI-BLAST (Altschul et al. 1997) at NCBI (details below). For *C. intestinalis,* eight *rxfp1/2*-type genes were retrieved from Ensembl and two candidate *rxfp3/4*-type genes were obtained from ANISEED (http://crfb.univ-mrs.fr/aniseed). Two *C. productum ilp* genes were obtained from McRory and Sherwood (McRory and Sherwood 1997).

PHI-BLAST searches in the Ciona genome

Successful PHI-BLAST searches were conducted in the Ciona genome; analogous searches were performed in other invertebrate deuterostomes, but either did not yield any result at all (as in sea urchin) or only confirmed the known gene sets (as in amphioxus). For the PHI-BLAST searches, vertebrate RLN/INSL, INS/IGF and starfish rln-like

(“GSS”-gonad stimulating substance) peptide sequences were used as queries along with the amino acid patterns constructed using both the A and B-chains of RLN/INSL and INS/IGF peptide sequences based on their alignment and the GSS sequence from starfish. This led to the identification of one novel *C. intestinalis* candidate rln-like sequence that was then used to identify a similar rln-like gene in the *C. savignyi* database in Ensembl. Analysis of the Ciona proteome with the traditional vertebrate RLN/INSL B-chain pattern “CGR-x(3)-R-x(5)-CG” did not yield any results. At the same time using a more simplified B-chain pattern “C-X(11)-C” yielded numerous sequences rich in cysteines, most of which however did not possess the two-chain peptide structure characteristic to the Insulin-Relaxin superfamily. Using the A-chain pattern “CC-x(2)-GC-x(8)-C” derived from the combined alignment of the starfish and vertebrate RLN/INSL sequences with various (vertebrate and starfish) Rln/Insl as queries, yielded a RLN/INSL-like protein predicted as a functional gene in Ensembl. According to the ANISEED database (http://crfb.univ-mrs.fr/aniseed) this gene is expressed in the nervous system of C. intestinalis. Interestingly, the hits in the search performed with this pattern also contained ins-l1 and ins-l3. Using the INS/IGF A-chain pattern “CC-x(3)-C-x(8)-C” and INS/IGF sequences as queries yielded ins-l1 as the first hit and also identified the RLN/INSL-like gene as the second hit.

The six *ilp* genes from the amphioxus database were previously analyzed and shown to have syntenically shared genes with the vertebrate Insulin-Relaxin loci (Holland et al. 2008). VISTA Point (http://pipeline.lbl.gov/cgi-bin/gateway2?bg=Brafl1&selector=vistapoint) was used to further look at the synteny between the amphioxus scaffolds hosting *ilp* genes and the human genome (Appendix A). Five amphioxus *rxfp1/2*-type genes were retrieved from GenBank.

My searches in the sea urchin database (http://www.spbase.org/SpBase/) yielded 27 *rxfp1/2*-like sequences, but no *ilp* sequences. Two *ilp* sequences were retrieved for two lancelet species (*Branchiostoma belcheri* and *B. californiensis*) and one *ilp*, GSS, obtained for starfish (*Asterina pectinifera*) from GenBank; seven *ilp* and two *rxfp1/2*-like genes (*lgr3* and *lgr4*) were obtained from Ensembl Metazoa (http://metazoa.ensembl.org) for fruit fly (*Drosophila melanogaster*). All accession numbers for the pre-2R taxa are available in Tables B6-B8 in Appendix B. The vertebrate *INS/IGF* sequences used to root the *RLN/INSL* phylogeny (described below) were retrieved from GenBank (Tables B9-B10).

Phylogenetic reconstruction of the relation among RLN/INSL and RXFP genes

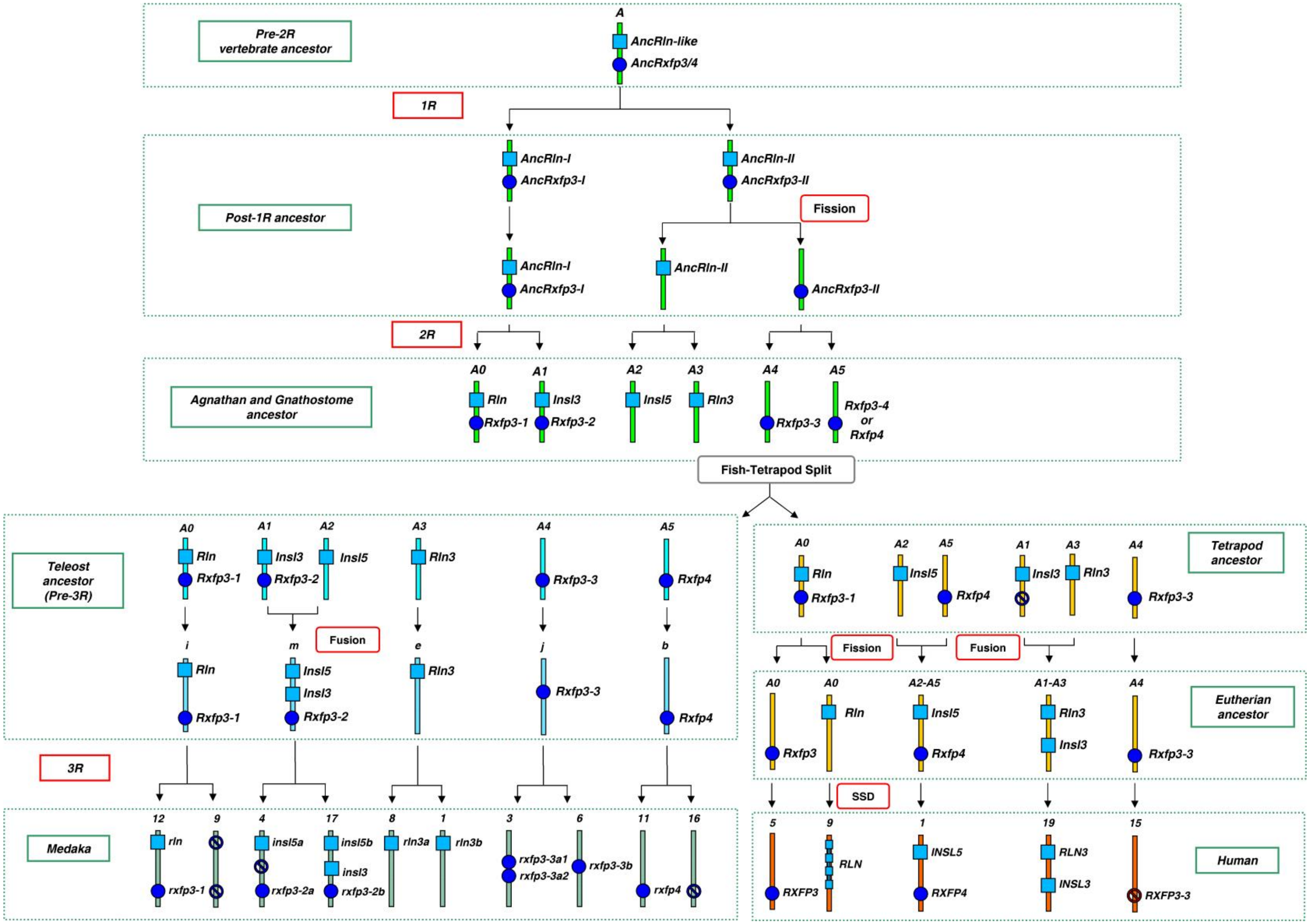
The amino acid alignment of RLN/INSL and INS/IGF was performed as outlined in Good-Avila et al. (2009). The alignment of RXFP proteins was accomplished using MUSCLE (Edgar 2004) as implemented in MEGA v. 5.01 (Tamura et al. 2011) and through manual adjustments. Phylogenetic reconstruction of protein sequences was carried out in Phyml (Guindon et al. 2010) using: for *RXFP* genes, the LG model of sequence evolution and with estimated or fixed values for *G*, the shape parameter for the gamma distribution, and *I*, the proportion of invariant sites, depending on what was determined to be the best model of amino acid sequence evolution using AIC as implemented in ProtTest (Abascal et al. 2005); for *RLN/INSL* genes, the JTT model with *G=1.063* and *I=0.045*. Confidence in the phylogenetic reconstruction was assessed using

1000 replicate bootstrap samples. The phylogenetic relationship among invertebrate *rxfp1/2*-type genes was reconstructed separately following the methods described above for the vertebrate *RXFP* sequences.

RESULTS

In the first part of this study, I inferred the origins of the *RLN/INSL* and *RXFP* gene sets by comparing the ancestry of large chromosomal fragments in a teleost fish (Japanese medaka), a bird (chicken) and human using a model of vertebrate genome evolution (Nakatani et al. 2007), the "N-model" (for a full explanation of the method, see Appendix A). Since, with some exceptions, *RLN/INSL-RXFP* genes in non-mammals have been primarily characterized by automated gene scan tools and are poorly annotated, I searched a number of available vertebrate genomes (25 species) for the focal genes (235 total genes) to ensure that all potential ligand and receptor ohnologs were considered (see Tables B1-B5, Appendix B). Thus for human, chicken and medaka, I mapped the genomic positions of 4, 3 and 6 *RLN/INSL* (ligand) and 6, 4 and 9 *RXFP* (receptor) genes (or pseudogenes), respectively, onto the linkage groups composing each of the 3 vertebrate genomes (according to the N-model) and "traced" their origins to the gnathostome ancestor chromosomes (*GAC*), i.e. linkage groups of the hypothetical post-2R ancestor of jawed (and possibly jawless, see Kuraku et al. (2009) vertebrates. According to the N-model, each of the 40 post-2R reconstructed *GAC*s (*A0-J1*) originate from 10 vertebrate ancestral chromosomes (*VAC, A-J*), i.e. linkage groups that existed in the hypothetical pre-2R genome. For 3 of the *VACs* (*A*, *B* and *F*), Nakatani et al. (Nakatani et al. 2007) were able to reconstruct the major chromosomal fission events that multiplied the chromosome numbers in the pre-1R, post-1R and post-2R vertebrate ancestor genomes. The occurrences of several of my genes-of-interest on these *GAC*s allowed me to not only trace their pre-2R origins, but also to assess the number and linkage relationships of ligand and receptor genes in the intermediate post-1R vertebrate ancestor. In their work, Nakatani *et al.* (2007) proposed two alternative scenarios for the duplication and rearrangement history of *VAC* “*A*” (found to host the predecessors of both *RLN/INSL* and *RXFP3/4* genes, see below). I considered both scenarios and adopted the more parsimonious one, which minimizes the overall gene loss and duplication concerning our focal genes. As described in detail in Appendix A, the primary difference of the main (“fusion”) and the alternative (“fission”) models concerns the time of duplication of *AncRln-I/AncRln-II* and *AncRxfp3-I/AncRxfp3-II* genes. In the alternative scenario these duplication events occurred in the proto-pre-2R genome, while in the main model, they occurred commensurate with 1R. Not only is the model adopted here more parsimonious, but it is also supported by the phylogenetic data (see below and Figure 1.5a), which indicates a short evolutionary period (measured in branch lengths) separating the divergence of *AncRxfp3-I/AncRxfp3-II* (the 1R event in the main model) from the divergence of *Rxfp3-1/Rxfp3-2* and *Rxfp3-3/Rxfp3-4* (the 2R event).

Figure 1.1. Reconstruction of the genetic events that led to the diversification of RXFP3-type receptors and RLN/INSL hormones in vertebrates.
The genomic origins of the hypothetical ancestral relaxin (*AncRln-like*) and *Rxfp3/4* receptor (*AncRxfp3/4*) genes can be traced to a single chromosome in the vertebrate ancestor that had not yet been through the two rounds of WGD, 2R (Pre-2R vertebrate ancestor). The ancestral linkage group harbouring *AncRln-like* and *AncRxfp3/4* genes sequentially underwent duplication, fission and another duplication yielding 5 distinct linkage groups (agnathan and gnathostome ancestor) harbouring the ligand and receptor genes. Subsequently, tetrapods completely lost *RXFP3-2* and often *RXFP3-3* genes, but retained all of the post-2R *RLN/INSL* gene duplicates. Teleosts, on the other hand, retained all of the ligand and receptor post-2R gene duplicates, suggesting that *RXFP3-2* and *RXFP3-3* acquired important functions in the pre-3R teleost ancestor. The duplicates of *rxfp3-2* and *rxfp3-3* were again retained in the post-3R teleost ancestor along with those of *rln3* and *insl5* (indicating their possible ligand-receptor relationships). Lastly, in placentals the *RLN* locus underwent multiple local duplications (depicted as multiple boxes in the human *RLN* locus), resulting in the emergence of *INSL4* in all eutherians, and *INSL6* and *RLN1* only in apes, whose *RLN2* is orthologous to *RLN* of other eutherians. For simplicity, tetrapod and eutherian ancestor linkage groups are only shown to contain the fragments (e.g. *A0, A2-A5*) harbouring the genes of interest; thus they should not be confused with actual chromosomes. Blue circles and squares represent receptor and their ligand genes respectively. Crossed circles represent pseudogenes (red, if they are verified in databases, blue if they are hypothetical). SSD: small-scale duplication. The first letter of ancestral gene names is capitalized.


Pre-2R
vertebrate ancestor
A
AncRln-like
AncRxfp3/4
1R
Post-1R ancestor
AncRln-I
AncRxfp3-I
AncRln-II
AncRxfp3-II
Fission
AncRln-I
AncRxfp3-I
AncRln-II
AncRxfp3-II
2R
Agnathan and Gnathostome ancestor
A0
Rln
Rxfp3-1
A1
Insl3
Rxfp3-2
A2
Insl5
A3
Rln3
A4
Rxfp3-3
A5
Rxfp3-4
or
Rxfp4
Fish-Tetrapod Split
Teleost ancestor (Pre-3R)
A0
Rln
Rxfp3-1
A1
Insl3
Rxfp3-2
A2
Insl5
A3
Rln3
A4
Rxfp3-3
A5
Rxfp4
Fusion
i
Rln
Rxfp3-1
m
Insl5
Insl3
Rxfp3-2
e
Rln3
j
Rxfp3-3
b
Rxfp4
3R
Medaka
12
rln
rxfp3-1
9
4
insl5a
rxfp3-2a
17
insl5b
insl3
rxfp3-2b
8
rln3a
1
rln3b
3
rxfp3-3a1
rxfp3-3a2
6
rxfp3-3b
11
rxfp4
16
Tetrapod ancestor
A0
Rln
Rxfp3-1
A2
Insl5
A5
Rxfp4
A1
Insl3
A3
Rln3
A4
Rxfp3-3
Fission
Fusion
Eutherian ancestor
A0
Rxfp3
A0
Rln
A2-A5
Insl5
Rxfp4
A1-A3
Rln3
Insl3
A4
Rxfp3-3
SSD
Human
5
RXFP3
9
RLN
1
INSL5
RXFP4
19
RLN3
INSL3
15
RXFP3-3

RLN/INSL and RXFP3/4 originate from one ancestral linkage group, while RXFP1/2 originates from another

Tracing of human, medaka and chicken genes to ancestral chromosomes revealed that *RLN*, *RLN3*, *INSL3*, *INSL5* and their orthologs in teleosts originated from one location in *VAC "A"* in the pre-2R vertebrate ancestor (see Table A1, Appendix A). Since each of the four *RLN/INSL* genes can be mapped to 4 distinct 2R-derived *GAC*s (*A0*, *A1*, *A2* and *A3*), I infer that modern vertebrate relaxin family genes arose from a single ancestral gene, *AncRln-like*, as a result of 2R (Figure 1.1).

The origins of the receptor *RXFP3* and *RXFP4* genes from tetrapods and teleosts were traced to four *GAC*s (*A0*, *A1*, *A4* and *A5;* two of which, *A0* and *A1*, are the same as those hosting *RLN* and *INSL3)*, which suggests that vertebrate receptors *RXFP3* and *RXFP4* originated from one gene, *AncRxfp3/4-like,* located on *VAC* "*A"* (see Table A1 in Appendix A). This indicates that the ancestral genes for *RLN/INSL* and *RXFP3/4* were physically linked before 2R took place (Figure 1.1).

The high number of receptor *rxfp3*-type genes in teleosts is explained by the post-2R retention of all four *rxfp3/4* ohnologs in the teleost ancestor. Additionally, the fish-specific 3R coupled with a few local duplications increased the number of *rxfp3*-like genes in teleosts to 7 (Figure 1.1 and Tables B4-B5 in Appendix B). Interestingly, my data mining uncovered that a few tetrapods retained *RXFP3-3*, but *RXFP3-2* appears to have been completely lost in the early tetrapod ancestor (Figure 1.1 and Tables B1-B3, Appendix B). Using the available *RXFP3-3* sequences from opossum, cow and pig, the *RXFP3-3* pseudogene was located in human and its common origin (*GAC "A4"*) with its medaka orthologs (Table A1, in Appendix A) was confirmed.

The tracing of the ancestral origins of *RXFP1* and *RXFP2* receptors in human and medaka showed that both of these genes originated from *VAC "C"* (Table A1, Appendix A). Thus I concluded that 2R led to the duplication of an ancestral gene, *AncRxfp1/2* with the retention of only 2 orthologs (*RXFP1* and *RXFP2*) in human and medaka (Figure 1.2). Interestingly, duplicates of *rxfp1* and *rxfp2* were also lost after 3R in stickleback (*Gasterosteus aculeatus*), tetraodon (*Tetraodon nigroviridis*) and fugu (*Takifugu rubripes*), but were partly retained in zebrafish, in which I found two *rxfp2* orthologs (Figure C1 in Appendix C; Tables B4-5, Appendix B).

The two genes reported as *RXFP1* and *RXFP2* in chicken, turned out to have an evolutionary history that was slightly different from that of their counterparts in other vertebrates. Chicken *RXFP1* was traced to *GAC* "*C1"* (implying its orthology to the *RXFP1* of human and medaka), but the chicken *RXFP2* gene was traced to a different ancestral linkage group (*GAC "B0"* or "*F4"*) than the expected *GAC* "*C2"* (Table A1 in Appendix A). Further analyses confirmed that this gene does not share synteny with either *RXFP1* or *RXFP2* and I therefore rename it *RXFP2-like*. Subsequently, I identified an ortholog of this *RXFP2-like* gene in some other vertebrates, such as zebrafish and opossum, and found a pseudogene of the *RXFP2-like* gene on the human X chromosome next to *STARD8,* its neighbouring gene in chicken (Table B3 in Appendix B). Convincingly, BLASTn searches also revealed a pseudogene of *RXFP2* in the region of the chicken genome orthologous to that hosting *RXFP2* in other vertebrates. The tracing of *RXFP2-like* to a separate *VAC* ("*B"* or "*F"*) from that of *AncRxfp1/2 (VAC "C")* indicates that either it originated from a pre-2R locus independent from that of *AncRxfp1/2* or that it is the ohnolog of *RXFP1/2* and was translocated shortly after

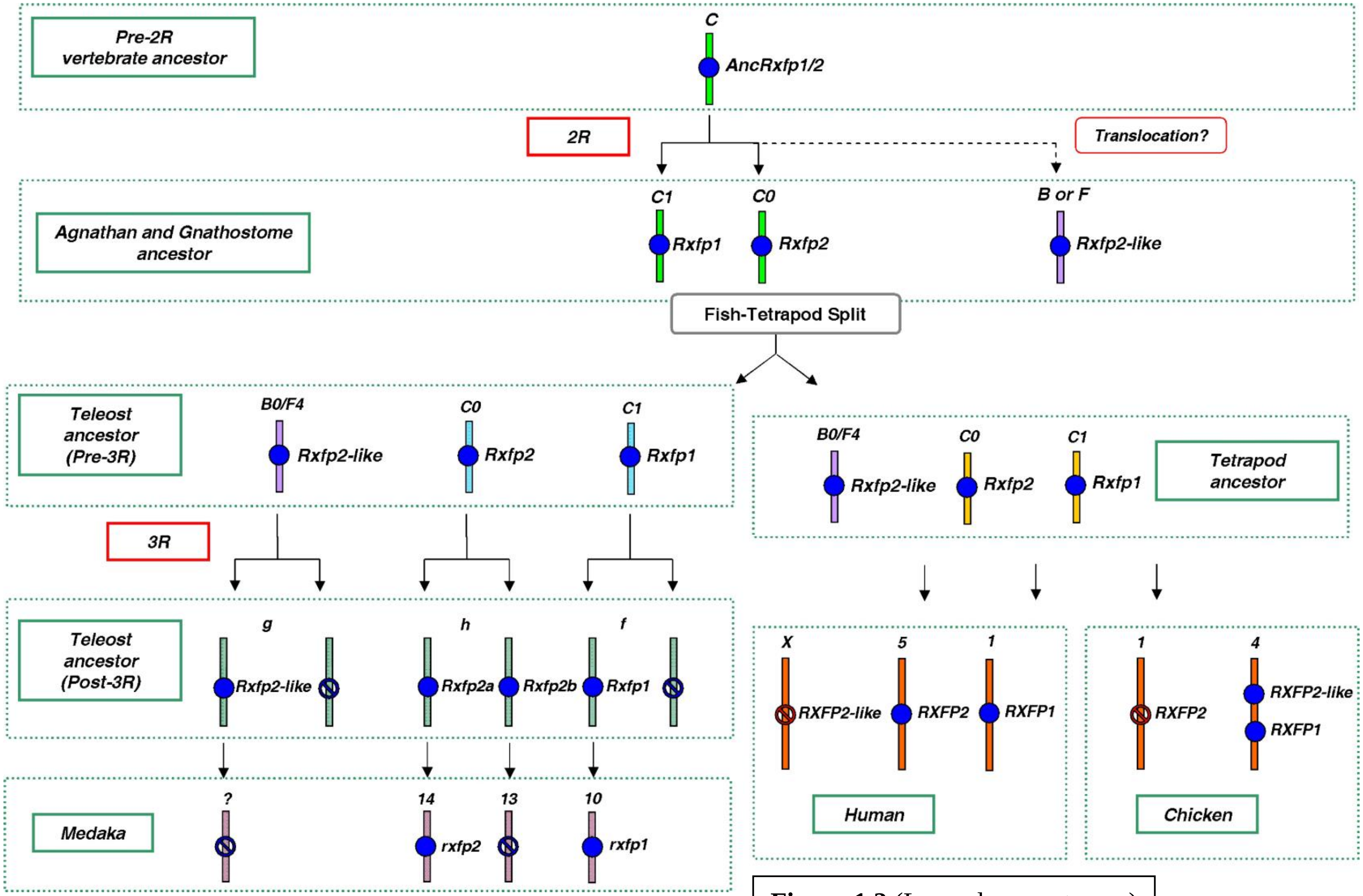


Figure 1.2 (Legend on next page)

Figure 1.2. Reconstruction of the genetic events that led to the diversification of *RXFP1/2*-type receptor genes in vertebrates.
Symbols and linkage group numbering same as in *Figure 1.1*

duplication from one of the *RXFP1/2* loci. Here I adopt an origin-based nomenclature for the novel genes identified in this study, in which I aim to reflect their relationship to their hypothetical ancestors while retain, as much as possible, the traditional naming scheme for the RLN/INSL peptide and their RXFP receptor genes (Table 1.1).

Linkage relationships among RLN/INSL and RXFP genes have changed over evolutionary time

In the pre-2R vertebrate ancestor, *AncRxfp3/4* (receptor) was in the same linkage group as *AncRln-like* (ligand). My reconstruction shows that two of the *RXFP3* 2R-ohnologs (*RXFP3-1* and *RXFP3-2*) were linked to *RLN* and *INSL3* (Figure 1.1), while the remaining ohnologs became unlinked. These ancestral genetic linkage relationships have mostly persisted in teleosts (Figure 1.3 and Figure C1 in Appendix C), but they have dynamically changed in tetrapods resulting in different combinations of linkage pairs such as *INSL5-RXFP4*, *RLN3-INSL3* and *RXFP1-RXFP2-like* to name a few (Figure 1.4).

RXFP phylogenetic reconstruction supports strong role of WGDs in gene duplication events

The second goal of this study was to use other types of analysis, such as phylogeny and small-scale synteny, to corroborate the above model of evolution of the vertebrate RLN/INSL-RXFP systems in a broader range of vertebrates. I created a protein database and subsequently phylogenetic trees of RLN/INSL and RXFP-type genes for vertebrates and a few pre-2R diverging taxa, based on publicly annotated genes and included a few that I identified *de novo* (Tables B1-B5 for post-2R vertebrates, Tables B6-B8 for pre-2R deuterostomes, in Appendix B). Overall, I find that the phylogenetic relationship of the receptor RXFP3/4 sequences clearly recapitulates their proposed WGD-driven diversification: the 1R descendants cluster into two groups, *AncRxfp3-I* versus *AncRxfp3-II*, while the 2R descendants are sister clades, i.e. RXFP3-1/RXFP3-2 and RXFP3-3/RXFP3-4 as expected (Figure 1.5a). Because most tetrapods lost half of their post-2R *RXFP3* ohnologs, the RXFP3-2 and RXFP3-3 clades mostly contain teleostean sequences.

The RXFP1/2 phylogenetic tree (Figure 1.5b) also generally supports the reconstruction model: there are 3 distinct clades for RXFP1, RXFP2 and RXFP2-like, and the RXFP2-like clade is sister to RXFP2, a clustering that supports the ohnologous nature (i.e. their orthology and origination by means of WGD) of the relationship between *RXFP2-like* and *RXFP1/2* genes, rather than the pre-2R origins of *RXFP2-like*. To examine this more closely, I analyzed several vertebrate RXFP1/2 and RXFP2-like sequences together with invertebrate rxfp1/2-type proteins, and found that all vertebrate sequences clustered together (Figure C3, Appendix C), indicating that all 3 genes (i.e. *RXFP1*, *RXFP2* and *RXFP2-like*) originated after the divergence of protochordates.

Lastly, as found in previous studies (e.g. Good-Avila et al. (2009) , the RLN/INSL phylogeny does not clearly reflect their WGD ancestry: the small size and differential selection pressures on peptides in tetrapod versus teleostean lineages renders it impossible to resolve orthologous relationships among *RLN/INSL* genes across vertebrates in the absence of synteny data. Here I combined phylogeny with small-scale synteny and also used a reconstruction of the ancestral chordate karyotype

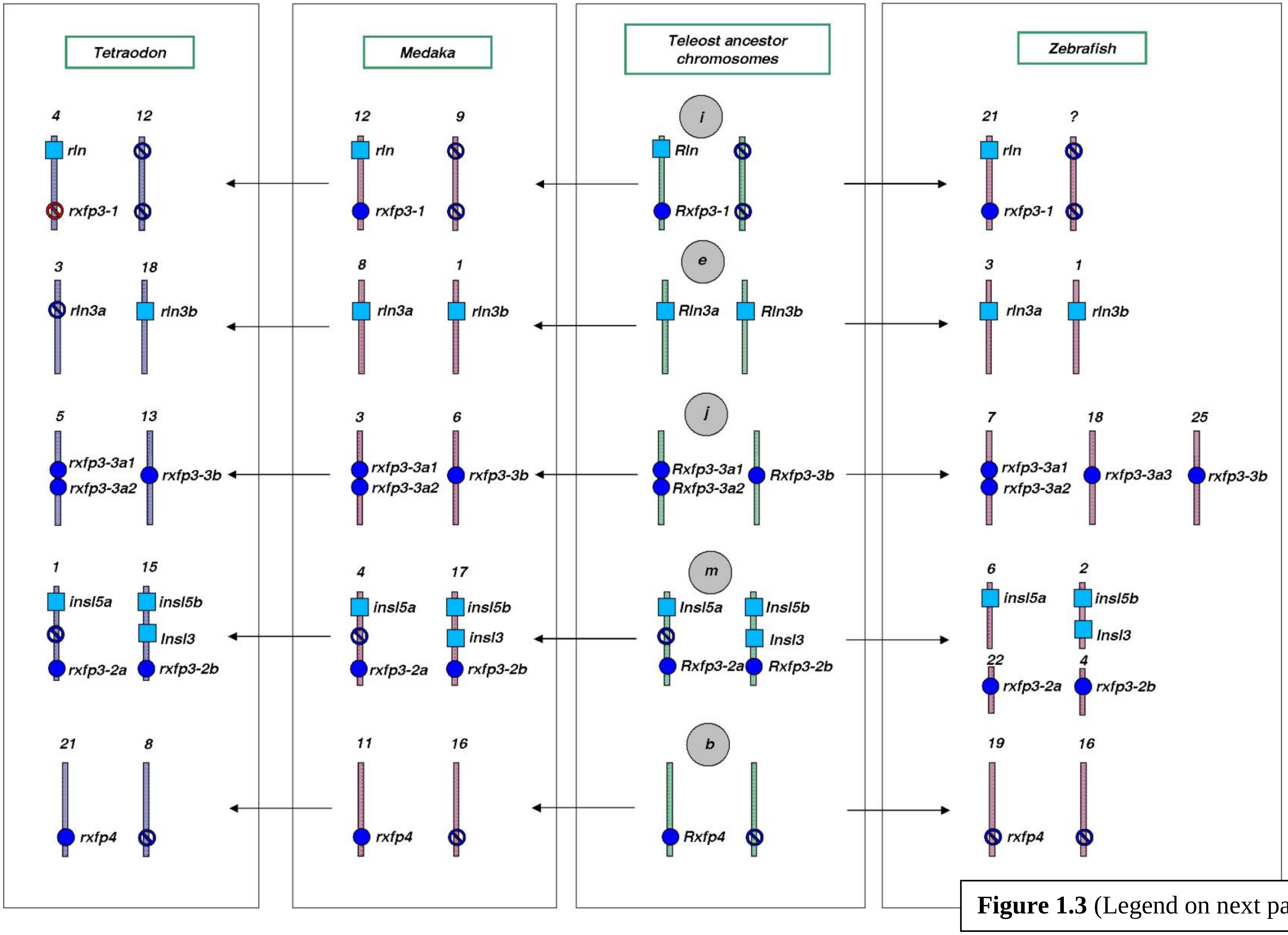


Figure 1.3 (Legend on next page)

Figure 1.3. The evolution and genetic linkage of *RLN/INSL* (ligand) and *RXFP3/4* (receptor) loci in the pre-3R teleost ancestor and three species of teleost fish. Notice that among the three analyzed fish species, medaka's genome and *rln/insl-rxfp* gene sets are the most preserved and resemble those of the teleost ancestor. Tetraodon experienced lineage-specific loss of two genes, *rln3a and rxfp3-1*, which may indicate their co-evolution as a ligand-receptor pair. The *rxfp4* gene in zebrafish seems to have been replaced with an extra (zebrafish-specific) copy of an *rxfp3-3* gene. Alternatively, rxfp3-3 could be a product of gene conversion that occurred between rxfp4 and one of the *rxfp3-3* paralogs. Overall this scheme demonstrates that the rln/insl-rxfp system in teleosts has taken a slightly different, and seemingly more complicated, evolutionary pathway compared to other vertebrates. Chromosome numbers in extant species are shown as numbers and in the teleost ancestor as letters (Kasahara et al. (2007) nomenclature).

(Putnam et al. 2008) to explore the evolutionary pathway of the entire insulin superfamily in deuterostomes. This was done to determine whether the observed trends, such as the number of potential *RLN/INSL*-like sequences in pre-2R organisms, fit into the model of evolution proposed for relaxin family peptides in post-2R vertebrates.

Presence of both RLN/INSL-like and INS/IGF-like genes in amphioxus, but only INS/IGF-like in tunicates: loss of RLN/INSL in the common ancestor of ascidians?

There are 2 classes of *ins-like* genes in protochordates and echinoderms. Both lancelets and tunicates possess proteins clustering inside the vertebrate INS/IGF clade (Figure C2a, Appendix C); whereas starfish, 2 *Ciona* species (*C. intestinalis* and *C. savignyi*) and amphioxus have proteins that group sister to the vertebrate RLN/INSL clade (1 starfish *gss* and 3 amphioxus insulin-like peptides *(ilps*) or just outside it (amphioxus *ilp6* and ciona *ilp3*-*ilp4*) (Figure 1.7; Figure C2a-b, Appendix C). Somewhat surprisingly, amphioxus *ilp5* clusters in the vertebrate INSL5 clade (Figure 1.7).

Mapping the amphioxus scaffolds hosting *ilp* genes to ancestral chordate linkage groups (*CLG*) as outlined by Putnam et al. (2008) revealed that three of these genes (*ilp1*-*3*) originate from the same ancestral linkage group as the human *INS/IGF* loci, while the other 3 *ilp* genes either do not share *CLG*s with any genes coding for members of the human relaxin-insulin superfamily (*ilp6*) or no *CLG* could be assigned to them due to insufficient information available in the short scaffolds of amphioxus (*ilp4-5*, Figure A4, Appendix A). These findings imply that the *AncIns/Igf* gene may have duplicated before the emergence of cephalochordates to give rise to one *ins*-type and five *rln/insl*-type genes (through multiple duplications) in lancelets. While the ancestral chordate *Ins* gene survived until the emergence of vertebrates (later giving rise to the ancestor of *Igf* as a result of local duplication); it is unclear whether one of the amphioxus *rln/insl*-type genes is orthologous to the predecessor gene of the vertebrate relaxin family or whether a later duplication of the *ins* locus (*i.e.* after the divergence of cephalochordates) gave rise to an *AncRln-like* gene.

Figure 1.4. Dynamic changes in the chromosomal linkage relationships of *RLN/INSL* and *RXFP* genes in tetrapods.
Each bar represents a chromosome (IDs not shown for simplicity). Symbols and linkage numbering are as in *Figure 1.1*.

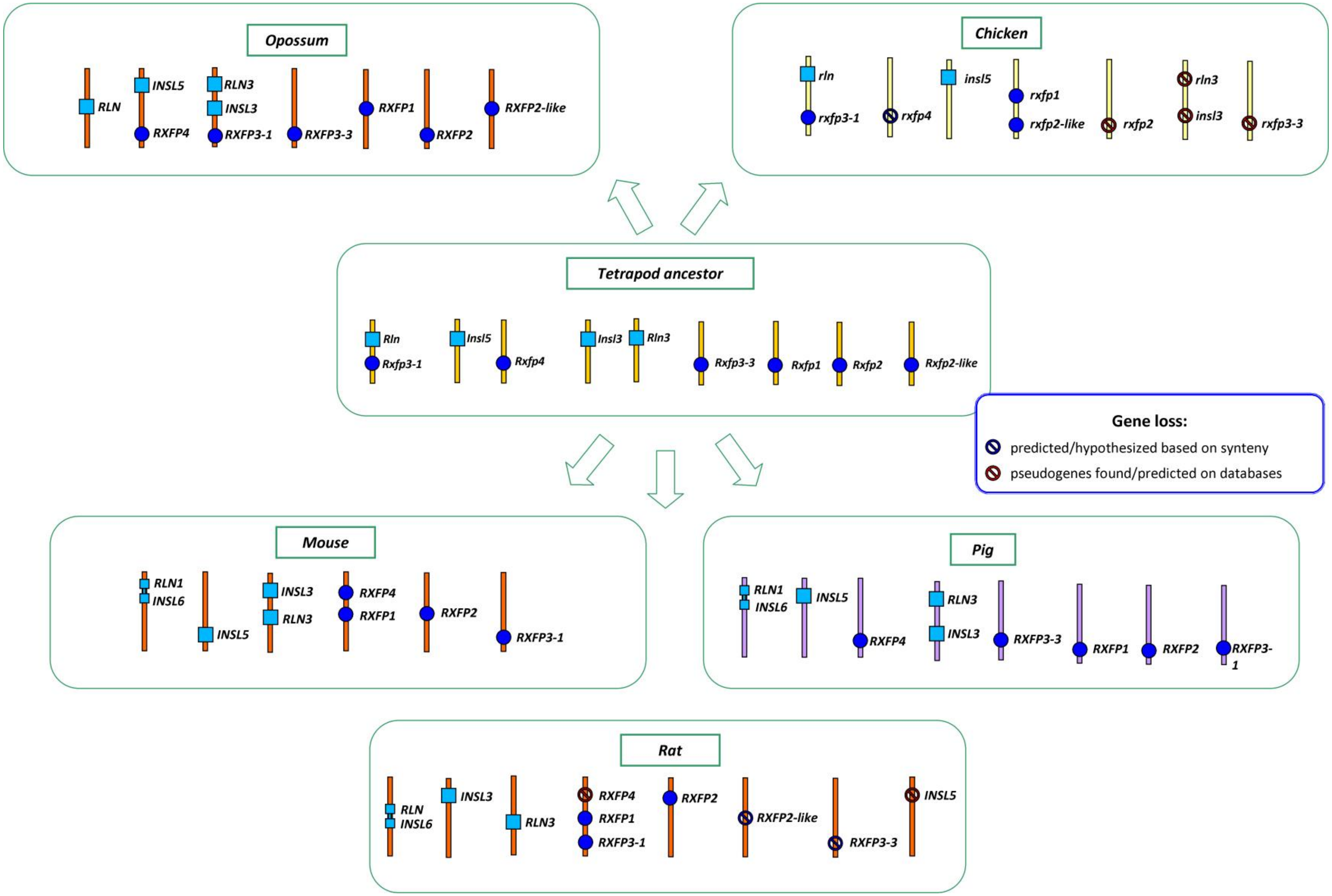
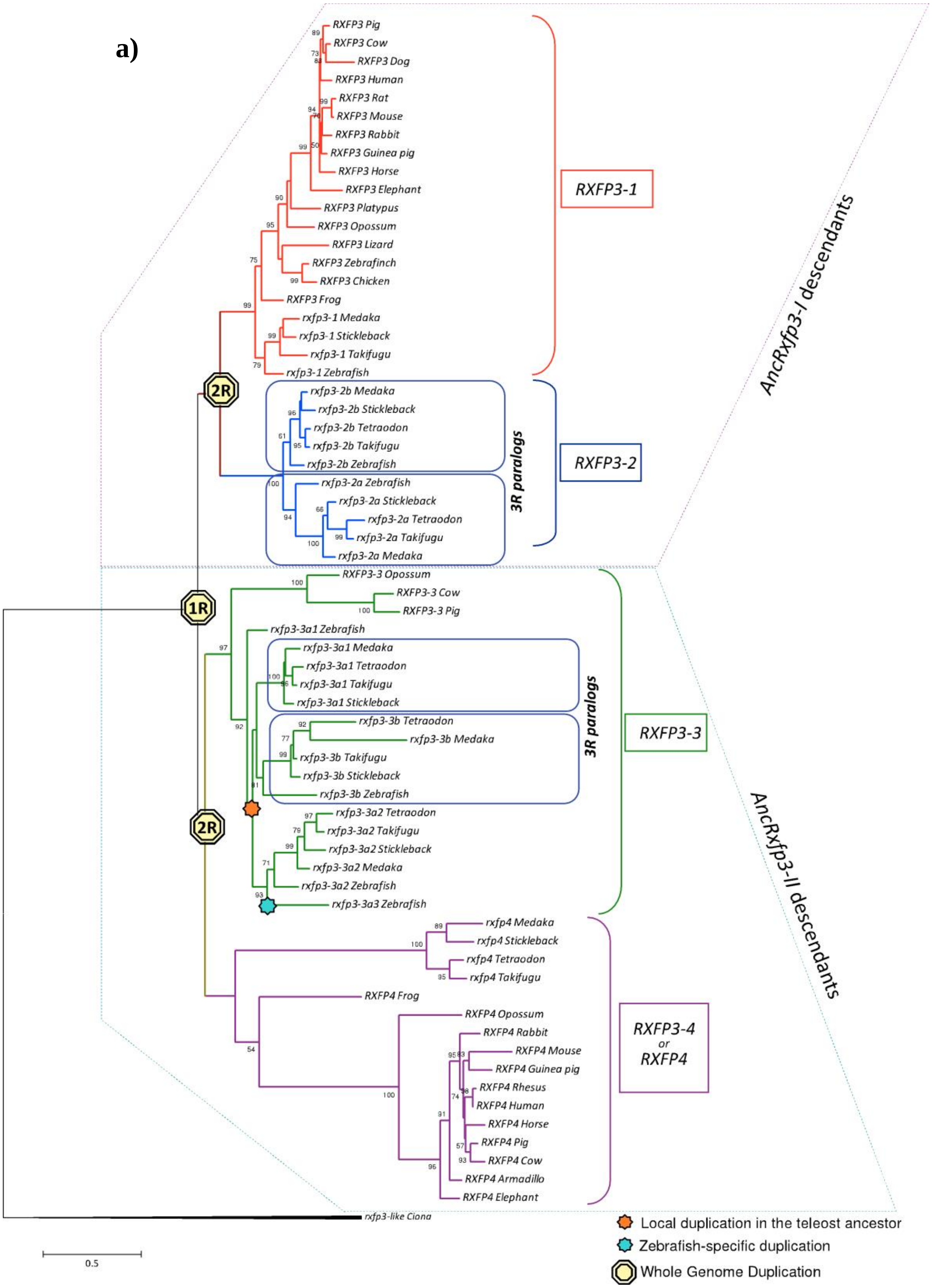


Figure 1.4 (Legend on previous page)


a)
RXFP3 Pig
RXFP3 Cow
RXFP3 Dog
RXFP3 Human
RXFP3 Rat
RXFP3 Mouse
RXFP3 Rabbit
RXFP3 Guinea pig
RXFP3 Horse
RXFP3 Elephant
RXFP3 Platypus
RXFP3 Opossum
RXFP3 Lizard
RXFP3 Zebrafinch
RXFP3 Chicken
RXFP3 Frog
rxfp3-1 Medaka
rxfp3-1 Stickleback
rxfp3-1 Takifugu
rxfp3-1 Zebrafish
RXFP3-1
rxfp3-2b Medaka
rxfp3-2b Stickleback
rxfp3-2b Tetraodon
rxfp3-2b Takifugu
rxfp3-2b Zebrafish
rxfp3-2a Zebrafish
rxfp3-2a Stickleback
rxfp3-2a Tetraodon
rxfp3-2a Takifugu
rxfp3-2a Medaka
3R paralogs
RXFP3-2
AncRxfp3-I descendants
2R
1R
RXFP3-3 Opossum
RXFP3-3 Cow
RXFP3-3 Pig
rxfp3-3a1 Zebrafish
rxfp3-3a1 Medaka
rxfp3-3a1 Tetraodon
rxfp3-3a1 Takifugu
rxfp3-3a1 Stickleback
rxfp3-3b Tetraodon
rxfp3-3b Medaka
rxfp3-3b Takifugu
rxfp3-3b Stickleback
rxfp3-3b Zebrafish
3R paralogs
RXFP3-3
rxfp3-3a2 Tetraodon
rxfp3-3a2 Takifugu
rxfp3-3a2 Stickleback
rxfp3-3a2 Medaka
rxfp3-3a2 Zebrafish
rxfp3-3a3 Zebrafish
2R
rxfp4 Medaka
rxfp4 Stickleback
rxfp4 Tetraodon
rxfp4 Takifugu
RXFP4 Frog
RXFP4 Opossum
RXFP4 Rabbit
RXFP4 Mouse
RXFP4 Guinea pig
RXFP4 Rhesus
RXFP4 Human
RXFP4 Horse
RXFP4 Pig
RXFP4 Cow
RXFP4 Armadillo
RXFP4 Elephant
RXFP3-4
or
RXFP4
AncRxfp3-II descendants
rxfp3-like Ciona
0.5
Local duplication in the teleost ancestor
Zebrafish-specific duplication
Whole Genome Duplication

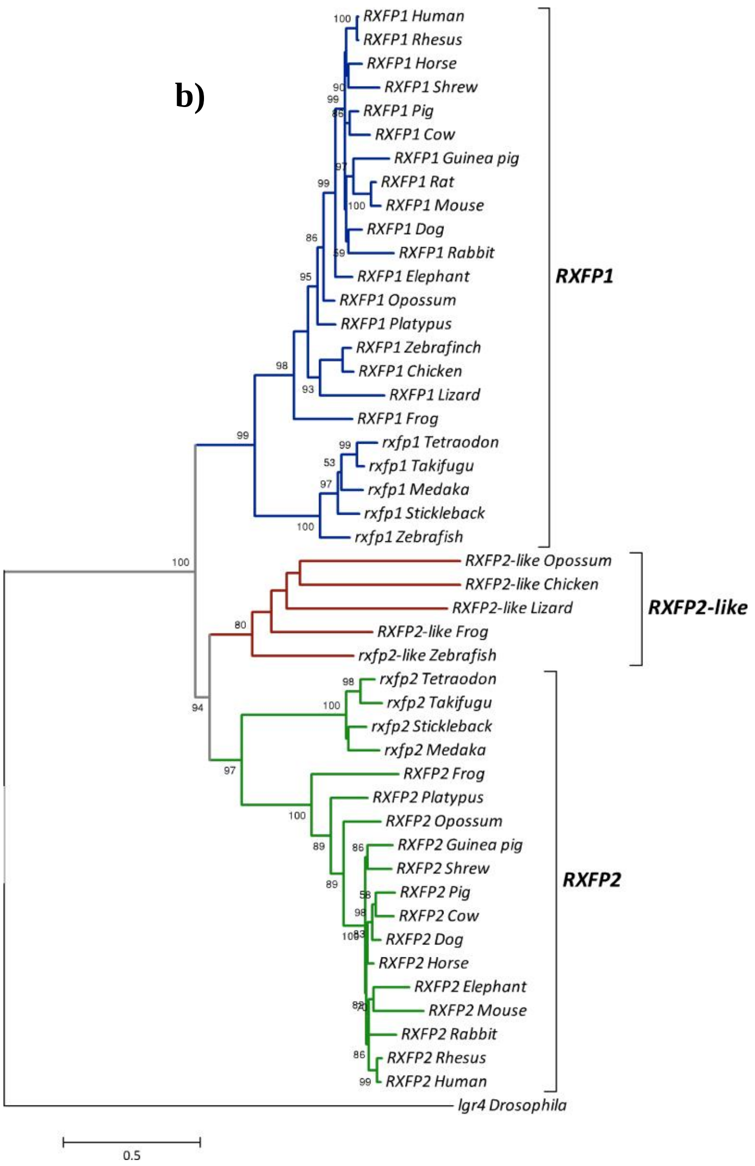


Figure 1.5. Phylogenetic reconstruction of the evolutionary relationship among vertebrate RXFP protein sequences.
a) RXFP3/4. Reconstruction performed as outlined in methods with *G=0.91* and *I=n/a*. Numbers at each node indicate the bootstrap values (only values exceeding 50% shown). Teleost *rxfp3-2* underwent duplication yielding two 3R-paralogs, *rxfp3-2a* and *rxfp3-2b*, while teleostean ancestral *rxfp3-3* was duplicated typically giving rise to three *rxfp3-3* loci in modern teleosts: 3R generated *rxfp3-3a* and *rxfp3-3b*, while a local duplication generated *rxfp3-3a1* and *rxfp3-3a2* (Table 1.1, Tables B4-5 in Appendix B). Solely in zebrafish, *rxfp3-3a2* duplicated again giving rise to *rxfp3-3a3*, an event which appears to have occurred coincidently with the exclusive loss (or gene conversion, see Chapter 2) of *rxfp4* in zebrafish. **b)** RXFP1/2. Phylogenetic tree reconstructed as outlined in methods with *G=0.958* and *I=0.034*. Numbers at each node indicate the bootstrap values (only values exceeding 50% shown). Due to their incomplete nature, not all sequences were included in this tree (e.g. zebrafish rxfp2a and rxfp2b and medaka rxfp2).

Origins of the vertebrate INS/IGF locus

To determine whether *AncRln-like* and the *AncIns/Igf* loci possibly occupied the same linkage group in the pre-2R vertebrate ancestor, the "N-model" tracing procedure (see Appendix A) was applied to the *INS/IGF* genes from tetrapods and teleosts. This revealed some surprising findings about vertebrate *INS/IGF* genes. Tetrapods possess three *INS/IGF* genes, two of which (*INS* and *IGF2*) are tightly linked on the chromosomal segment traceable to *GAC "D0"* and the third one, *IGF1*, is located in a different linkage group, traceable to *GAC "D1"*(Figure 1.8), indicating that *IGF1* and *IGF2* are ohnologs. On the other hand, in teleosts, the *ins* locus is not linked to *igf*, and the number of insulin genes ranges from 2-3, raising the question of the orthology of teleost and tetrapod *INS* genes. To investigate this, I used the genome reconstruction performed on *Tetraodon nigroviridis* (Kasahara et al. 2007) due to insufficient genomic data for medaka. Tetraodon has two *igf* genes (*igf1* and *igf2*) and two *ins* genes (*ins1* and *ins2*). Tracing of the chromosomal segments hosting *igf* loci in tetraodon indicated that the tetraodon *igf* genes share origins with their tetrapod counterparts.

Figure 1.6. Phylogenetic reconstruction of RXFP1/2 sequences from vertebrates, protochordates and an echinoderm.
All of the amphioxus *rxfp1/2*-like genes cluster closely to vertebrate *RXFP1/2*'s and *RXFP2-like* genes, while two of the 27 sea urchin *rxfp1/2*-like genes are found in a clade with fruit fly *lgr3* and *lgr4* in another sister clade to *RXFP1/2*-like genes. The Ciona *rxfp1/2*-like genes appear distantly related to the entire protostome-deuterostome *RXFP1/2* cluster.

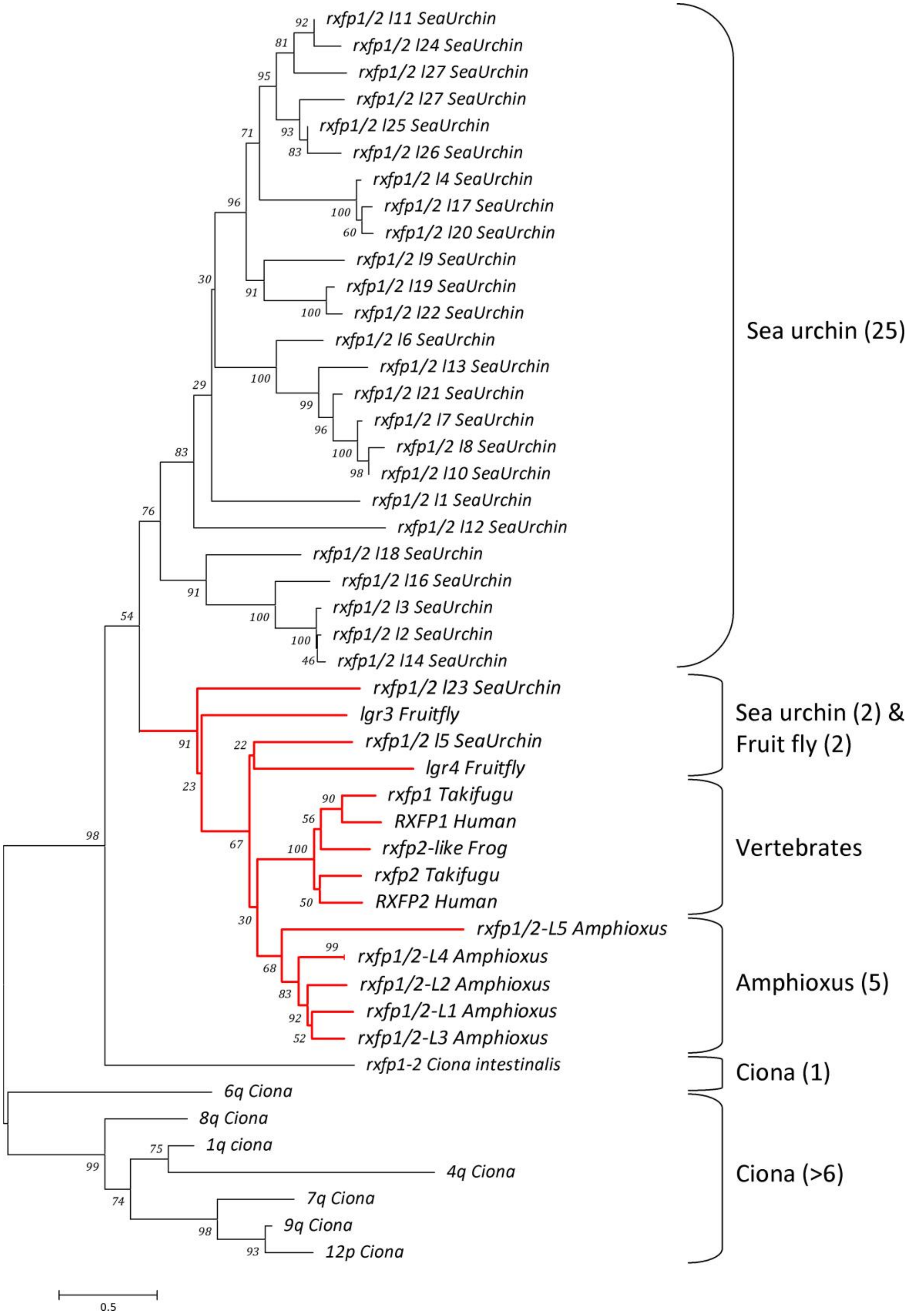


Figure 1.6 (Legend on previous page)

To the contrary, the *GAC*s for *ins*-hosting chromosomes were ambiguously mapped to three *GAC*s (*B0, C2 or F4)*, which do not host any *ins-igf* loci in other vertebrates, a situation reminiscent of that for *RXFP2-like* in chicken (see above). However, further syntenic (not shown) and phylogenetic (Figure 1.7) analyses of teleost *ins* genes could not clearly resolve their evolutionary pathway. Although, I find that some teleost *ins* loci share synteny with human chromosome 11 (Hsap11) that hosts *INS/IGF2,* others share synteny with certain regions of Hsap10 and Hsap5, which have no detectable relationship to the insulin superfamily loci. Notably, zebrafish *insb* (and *ins-l,* which appears to be a pseudogene) along with tetraodon *ins1* appear to be phylogenetically most distant from the rest of the *ins* genes (Figure 1.7), which may indicate that certain teleost-specific evolutionary pressures on the *ins* locus could have fostered the duplication and subsequent translocation of *ins*-duplicates. Overall my analyses make it clear that *AncIns/Igf* and *AncRln-like* loci were already in separate linkage groups before the onset of 2R, but leave unresolved the origin of several teleost *ins*-duplicates.

DISCUSSION

Although it is now widely accepted that the two rounds of WGD that took place early in vertebrate evolution played a crucial role in the diversification of many vertebrate gene families (Kasahara 2007), the processes by which WGD-driven gene family evolution occurred are not easy to determine. This has been shown to be true for the three vertebrate gene families encoding relaxin hormones and their receptors (*RLN/INSL, RXFP1/2* and *RXFP3/4*), whose duplication history and invertebrate origins I analyzed here.

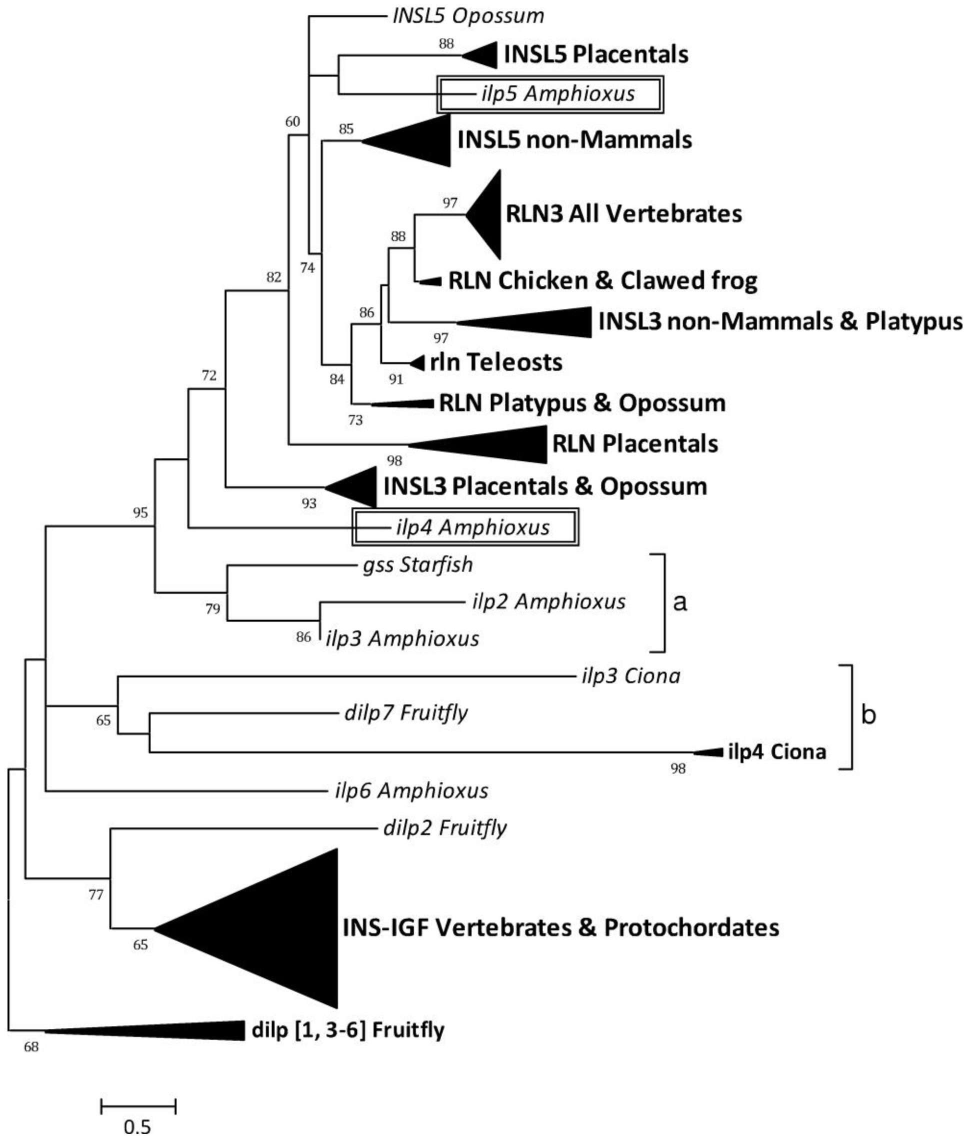


Figure 1.7. The relationship among vertebrate and invertebrate Insulin superfamily protein sequences depicted using an extremely collapsed phylogenetic tree.
Only for the most conserved peptide, RLN3, do all vertebrate orthologs cluster together, while for the remaining peptides (INSL3, RLN, and INSL5) differential selection pressures and the small size of the ligand peptides prevent the orthologs from grouping together across lineages (Good-Avila et al. 2009). Within the strongly supported clade (at 95%) that contains all vertebrate RLN/INSL sequences, there are additionally 4 ilps from amphioxus and the relaxin-like peptide (gss) from starfish (ilps4-5 are boxed, ilp2-3 and gss shown with bracket “a”), providing support that all of these sequences are of the RLN/INSL-type rather than INS/IGF-type. The tree was rooted with the clade containing five of the seven Fruitfly dilp peptides. Probably due to the known phenomena that divergent sequences tend to group together (so-called “long-branch attraction”), the two *Ciona* ilp genes (ilp3 and ilp4) group with dilp7, although the latter has been hypothesized to have “relaxin-like” reproductive roles in fruitflies , shown with “b”. For expanded versions of this phylogeny see Figure C2 in Appendix C. ilp: insulin-like peptide; dilp: drosophila insulin-like peptide

By combining information from ancestral genome reconstructions with phylogenetic and syntenic data, I was able to elucidate the origin of the *RLN/INSL-RXFP* genes. My study also revealed the intriguing linkage of the ancestral *RLN/INSL* (ligand) and *RXFP3/4* (receptor) loci in the pre-2R vertebrate ancestor genome, and the strong role of both 2R and 3R in the diversification of the focal genes.

The reconstruction of the *RLN/INSL-RXFP* gene history was principally based on Nakatani et al.'s (2007) model of the pre-2R vertebrate genome. It has been proposed that the major vertebrate novelties, such as their structurally complex nervous, immune and reproductive systems, arose as a result of the massive amplification of genes that occurred during 2R (Huminiecki and Heldin 2010, Kasahara 2007). By making the necessary assumption that my focal genes remained in the given linkage groups since the pre-2R vertebrate ancestor, I deduced that the diversification of *RLN/INSL* and *RXFP* genes was coincidental with 2R events, suggesting that the roles played by the *RLN/INSL* hormones in neuroendocrine and reproductive regulation were important in early vertebrate evolution. Interestingly, two of the post-2R *RLN/INSL* ohnologs, *RLN* and *INSL3*, that derived from one of the 1R duplicates (*AncRln-I*) are both involved in reproductive functions, while the other 2R ohnologs, *INSL5* and *RLN3*, play roles in the neuroendocrine system (Halls et al. 2007). This suggests that the pre-2R duplicates (*AncRln-I* and *AncRln-II*) probably differed from the original *AncRln-like* locus in the tissues they targeted: i.e. they were subfunctionalized for expression in either reproductive or neuroendocrine tissues following 1R. This provides an explanation for the retention of the 1R duplicates and may reflect the evolutionary origin of the systems as will be discussed below.

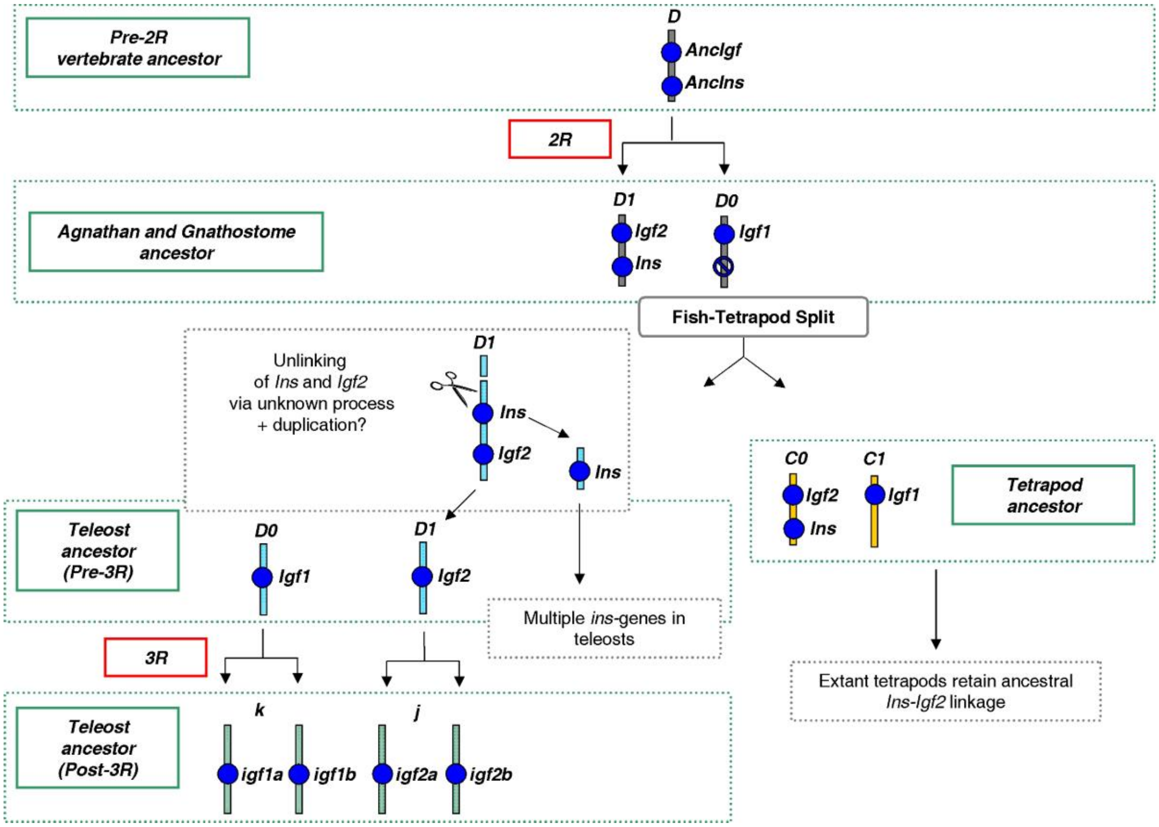

Pre-2R vertebrate ancestor
D
AncIgf
AncIns
2R
Agnathan and Gnathostome ancestor
D1
D0
Igf2
Ins
Igf1
Fish-Tetrapod Split
Unlinking of Ins and Igf2 via unknown process + duplication?
D1
Ins
Igf2
Ins
C0
C1
Igf2
Ins
Igf1
Tetrapod ancestor
Teleost ancestor (Pre-3R)
D0
D1
Igf1
Igf2
Multiple ins-genes in teleosts
3R
Extant tetrapods retain ancestral Ins-Igf2 linkage
k
j
Teleost ancestor (Post-3R)
igf1a
igf1b
igf2a
igf2b

Figure 1.8. Reconstruction of the possible genetic events that led to the diversification of *INS* and *IGF* genes in vertebrates.

Symbols and linkage numbering are as in *Figure 1.1*. The origin and evolutionary relationships of *ins*-genes in extant teleost fish are unclear.

I also observe that the teleost-specific 3R, which strongly contributed to the genetic richness of teleosts and their biological success, further increased the number of *rln/insl* and *rxfp* genes in teleosts. However, in contrast to the 1R and 2R events, only those genes potentially involved in neuroendocrine regulation (*rln3*, *insl5* and half of the *rxfp3/4*-type receptors), but not reproduction (*rln*, *insl3* and *rxfp1/2*-type receptors) were retained after 3R in teleosts. The post-3R retention of *rln3* and *insl5* was paralleled by the retention of duplicates of *rxfp3-2* and *rxfp3-3* suggesting both co-functioning but also subfunctionalization of their neuroendocrine functions. Overall, I demonstrate that the large number of teleost receptor *rxfp3* genes is only partly attributable to teleost-specific duplications (which was proposed as the sole factor driving their diversification (Wilkinson et al. 2005a), but rather also resulted from the loss of *RXFP3* ohnologs in tetrapods.

By thus elucidating the origin of genes, my model underscores the somewhat artificial nature of both ligand and receptor nomenclature, which is primarily based on the early physiological data. For ligands, I show that all *INSL* (insulin-like) genes independently a from *RLN* (relaxin-like) genes, and not from an ancestral *INSL* gene, as previously hypothesized (Hoffmann and Opazo 2011). For receptors, currently only 4 *RXFP* genes (*RXFP1-4*) are recognized, those present in humans and some other placentals, while, in fact, there are at least seven *RXFP* genes of independent origin in vertebrates (three *RXFP1/2* and four *RXFP3/4*-type), at least six of which are ohnologs, and six of which are present in at least one copy in tetrapods. I also show that *RLN3* and *INSL3* are ohnologs, and not closely related genes that arose from a tandem duplication event as previously hypothesized (Park et al. 2008). Furthermore, all four *RLN/INSL* ohnologs were retained after 2R, which contradicts a less parsimonious scenario discussed by Hoffman and Opazo (2011), in which one of *RLN/INSL* genes is lost in all vertebrates. Overall, my model for *RLN/INSL* evolution in vertebrates is consistent with the hypothesis postulating that *INSL3* and *RLN* evolved as a subfamily distinct from that formed by *RLN3* and *INSL5* (Wilkinson and Bathgate 2007). My model, however, dates the diversification of the two subfamilies back to the agnathan and gnathostome ancestors, while Wilkinson and Bathgate (2007) refer it to the more recent appearance of mammals.

Pre-vertebrate relaxin peptides

My study clearly indicates that the *RLN/INSL* and *INS/IGF* loci were already separate before the onset of 2R, but my search for *AncRln-like* genes in protochordates and echinoderms produced equivocal results. Both invertebrate deuterostomes and protostomes possess genes coding for insulin-like peptides (ilp's) recognized by the presence of disulfide bond-forming cysteines. Nevertheless, invertebrate proteomes are devoid of peptides with traditional RXFP-binding "R-XXX-R-I/V" motifs (Wilkinson et al. 2005b). Historically vertebrate *RLN/INSL* genes and their products have been characterized and classified according to the presence of their B-chain motif. Authors have therefore assumed that molecules without the motif could not bind the relaxin receptors, and therefore should not be considered to be related to vertebrate *RLN/INSL* (Wilkinson et al. 2005b). However, this assumption seems to contradict the co-evolutionary process of hormones with their cognate receptors, and I propose a different evolutionary model for early relaxin family peptides and their receptors.

I explored the origin of the *RLN/INSL* genes in early deuterostomes and found evidence of multiple relaxin-like genes, contrary to previous claims of an absence of *rln/insl*-like genes in invertebrates (Wilkinson and Bathgate 2007). Three of the amphioxus ilp peptides are phylogenetically close to a relaxin-like peptide, termed gonad stimulating substance (GSS), which is produced by the radial nerves in starfish and induces oocyte maturation and ovulation (Mita et al. 2009). Two of the amphioxus peptides, ilp4 and ilp5 (their genes are linked in one scaffold), cluster in close proximity or inside the vertebrate RLN/INSL clade, which may hint at their orthology with vertebrate *RLN/INSL* genes. Unfortunately synteny data was unavailable to confirm this. However, although I find evidence for relaxin-like genes in starfish and amphioxus, I failed to identify any *rln/insl*-type genes in the later-diverging tunicates. I confirm the presence of *ins/igf*-type genes in *Chelyosoma productum* (a tunicate), whose two ilp peptides cluster with lancelet ins/igf-like peptides (Figure C2a, in Appendix C), but the four *ilp* genes identified in *Ciona intestinalis* appear to be lineage-specific duplicates of the *ins* locus: two of them cluster basal to the INS/IGF-like peptides of other deuterostomes while the remaining two are intermediate between INS/IGF and RLN/INSL-type peptides. Olinski et al. (2006b) hypothesized that one of the *Ciona ilp* genes could be ancestral to vertebrate *RLN/INSL*, but I find no evidence for (or against) this although the lack of synteny and sequence identity I observe may be due to the accelerated evolution of the *Ciona* genome (Hughes A.L. and Friedman 2005). I advocate further studies in other tunicates such as *C. productum* to clarify this missing stage in the evolutionary pathway of the insulin superfamily between cephalochordates and pre-2R vertebrates.

Pre-vertebrate relaxin receptors

With respect to the origin of the *RXFP* receptors, I find that while both amphioxus and sea urchin genomes seem to be devoid of *rxfp3*-type genes (Nordstrom et al. 2008, Sodergren et al. 2006), and the two *rxfp3*-type genes in *C. intestinalis* are very divergent from their vertebrate analogs (Figure 1.5a), early deuterostome lineages witnessed many lineage-specific expansions of the *Rxfp1/2* locus (Figure 1.6). Intriguingly, three of the five amphioxus *rxfp1/2*-type genes appear orthologous to human *RXFP1* and *RXFP2* based on synteny. Thus collectively, given the observation of multiple *rln/insl* and *rxfp1/2*-type genes, which are unmistakably evolutionarily related to their vertebrate counterparts, combined with the virtual absence of *rxfp3*-type genes in echinoderms and cephalochordates, I propose that the signaling of the ancestral RLN/INSL peptide in the chordate ancestor occurred via RXFP1/2-type receptors. Only at the onset of 2R, was the RXFP3/4-type receptor recruited to produce a signaling system encoded in total by three genes and composed of two receptors and a single ligand. It is tempting to hypothesize that this ancestral two-receptor system had a dual function and played a role both in the regulation of reproduction (using RXFP1/2-type receptor), and in neuroendocrine processes (via RXFP3/4-type receptor). This hypothesis is supported by several lines of evidence: 1) the dual functionality of human RLN3, i.e. its integration of neuropeptide signaling with the ability to trigger reproductive response (McGowan et al. 2008), 2) the discovery of relaxin-like nature of a starfish gonadotropin which is produced by the echinoderm's nervous system and directly influences the maturation of eggs in the ovary of starfish (Mita et al. 2011) and 3) my prediction that the 1R duplicates of the *AncRln-like* gene were subfunctionalized into reproductive and neuroendocrine functions (see above).

Genetic linkage of receptors and their ligands

To the best of my knowledge, this is the first study to reveal the dynamic nature of the changing linkage relationships among the genes encoding relaxin family peptides and their receptors. The association of *RXFP3/4* genes with the *RLN/INSL* paralogon was first documented by Olinski et al. (2006a) and the linkage of human *INSL5* and *RXFP4* on one chromosome also mirrors their ligand-receptor interaction (Liu et al. 2005) . Nevertheless, I show that *INSL5* and *RXFP4* occupied different linkage groups in the gnathostome and tetrapod ancestors and only became linked in the eutherian ancestor (Figure 1.1). Chromosomal linkage of receptor and ligand genes has been known for a number of unrelated gene families and is more common in the human genome than expected by chance (Hurst and Lercher 2005). To explain this phenomenon, it has been proposed that receptor-ligand linkage could be advantageous for the creation of new receptor-ligand pairs when they result from block duplications (Hurst and Lercher 2005). However, this beneficial effect of linkage would not pertain to genes duplicated via WGDs, as is the case of *RLN/INSL-RXFP* loci. Instead, I propose that the *RXFP3/4-RLN/INSL* linkage may reflect the original need for connection during the recruitment of a new receptor by the *AncRln-like* peptide product. This linkage may have caused their co-expression and, consequently, increased the frequency of their interaction. Although the original linkage was disrupted in one of the post-1R ohnologs, linkage of certain *RLN/INSL-RXFP3* pairs has been conserved in some organisms, e.g. in medaka, while not in others, such as in rat (Figures 1.3 and 1.4). In this regard, it is interesting that the chromosomal sections harboring the *INSL/RLN* paralogons contain many other conserved gene families, such as the major histocompatibility complex genes, whose origins are traceable to singular pre-2R ancestor genes (Kasahara 2007). This suggests that conservation of the linkage relationship among the *RLN/INSL-RXFP* genes may result from conservation of synteny at a larger-scale. At the same time, vertebrates have also acquired novel and lineage-specific gene linkages, such as that of *RLN3-INSL3* in opossum, human and pig and *RXFP1-RXFP2-like* in chicken (Figure C2), which could be explained by other factors such as recurrent evolutionary chromosomal breaks in the fragile parts of genomes containing these genes (Bailey et al. 2004).

Finally, the difficulty in resolving the origin of two genes examined in this study, *RXFP2-like* and teleost *ins*, highlights an important weakness of the approach. If a gene maps to a different ancestral linkage group than expected, it is difficult to determine if 1) this gene has independent origins from its expected ohnologs or whether 2) it underwent a single-gene translocation that caused it to move from its authentic chromosomal fragment. For *RXFP2-like* it seems possible that there were two genes present in the ancestral pre-2R genome, because there are many *rxfp1/2*-type genes in primitive chordates (Figure 1.6). However, the few *RXFP2-like* genes from vertebrates cluster closely within the vertebrate *RXFP1/2* sequences (Figure 1.6), suggesting that *RXFP2-like* is an ohnolog of *RXFP1/2* that was translocated from its authentic position in the gnathostome ancestor. In the case of teleost insulin, I encountered an unexpected result: although some of the teleost *ins* genes seem to share their origin with the tetrapod insulin gene, some clearly do not. Again, based on the overall similarity of the teleost and tetrapod insulin genes, I favor the translocation hypothesis. However, it is also possible that teleost *ins* genes arose as local duplicates and subsequently moved to locations syntenically unrelated to the original chromosomal location of *INS/IGF* in the ancestor teleost genome. There is a documented example of a similar translocation of a duplicated insulin gene in rodents (Shiao et al. 2008). When traced using the N-model the murid-specific *INS*-paralog maps to an ancestral linkage group different from the expected *VAC "D"*(not shown here) owing to a single gene-translocation that took place early in the evolution of mice and rats (Shiao et al. 2008). Two resolve the origins of the vertebrate *INS/IGF* loci, examination of a wide range of taxa using independent methods should be performed.

Table 1.1. Explanation of the nomenclature used for the hypothetical ancestral genes that gave rise to the three gene families discussed in this study

<table>
<tr><th>Origin</th><th colspan="8">Gene Family: <i>RXFP3/4</i></th></tr>
<tr><td><i>Pre-2R</i></td><td colspan="8"><i>AncRxfp3/4</i></td></tr>
<tr><td><i>Post-1R paralogs</i></td><td colspan="4"><i>AncRxfp3-I</i></td><td colspan="4"><i>AncRxfp3-II</i></td></tr>
<tr><td><i>Post-2R paralogs</i></td><td colspan="2"><i>RXFP3-1</i></td><td colspan="2"><i>RXFP3-2</i></td><td colspan="2"><i>RXFP3-3</i></td><td colspan="2"><i>RXFP3-4*</i></td></tr>
<tr><td><i>Post-3R paralogs</i></td><td><i>rxfp3-1</i></td><td>†</td><td><i>rxfp3-2a</i></td><td><i>rxfp3-2b</i></td><td><i>rxfp3-3a</i>‡</td><td><i>rxfp3-3b</i></td><td><i>rxfp4</i></td><td>†</td></tr>
<tr><th>Origin</th><th colspan="8">Gene Family: <i>Rxfp1/2</i></th></tr>
<tr><td><i>Pre-2R</i></td><td colspan="8"><i>AncRxfp1/2</i></td></tr>
<tr><td><i>Post-1R paralogs</i></td><td colspan="4"><i>AncRxfp1</i></td><td colspan="4"><i>AncRxfp2</i></td></tr>
<tr><td><i>Post-2R paralogs</i></td><td colspan="2"><i>RXFP1</i></td><td colspan="2" rowspan="2">†</td><td colspan="2"><i>RXFP2</i></td><td colspan="2" rowspan="2">[<i>RXFP2-like</i>]**</td></tr>
<tr><td><i>Post-3R paralogs</i></td><td><i>rxfp1</i></td><td>†</td><td><i>rxfp2a</i></td><td><i>rxfp2b</i></td></tr>
<tr><th>Origin</th><th colspan="8">Gene Family: <i>RLN/INSL</i> <i>(Relaxin peptides)</i></th></tr>
<tr><td><i>Pre-2R</i></td><td colspan="8"><i>AncRln-like</i></td></tr>
<tr><td><i>Post-1R paralogs</i></td><td colspan="4"><i>AncRln-I</i></td><td colspan="4"><i>AncRln-II</i></td></tr>
<tr><td><i>Post-2R paralogs</i></td><td colspan="2"><i>RLN</i></td><td colspan="2"><i>INSL3</i></td><td colspan="2"><i>RLN3</i></td><td colspan="2"><i>INSL5</i></td></tr>
<tr><td><i>Post-3R paralogs</i></td><td><i>rln</i></td><td>†</td><td><i>insl3</i></td><td>†</td><td><i>rln3a</i></td><td><i>rln3b</i></td><td><i>insl5a</i></td><td><i>insl5b</i></td></tr>
</table>

* I show that the gene known as "RXFP4" is one of the three ohnologs of RXFP3-1, hence based on its origin it should be termed "RXFP3-4";
** The origins of *RXFP2-like* (present in zebrafish, amphibians, reptiles and marsupials) remain controversial, it is possible that *RXFP2-like* is a post-2R descendant of *AncRxfp2*, in which case it should be called "*RXFP2-2*", while the ortholog of human *RXFP2* should be called "*RXFP2-1*";
‡ *rxfp3-3a* was locally duplicated in the Post-3R ancestor of zebrafish, medaka, stickleback and pufferfishes (hence *rxfp3-3a* and *rxfp3-3b*); in zebrafish there are three paralogous *rxfp3-3a* genes: *3-3a1*, *3-3a2* and *3-3a3*;
† Gene loss

CHAPTER 2: *Gain and loss of RLN/INSL and RXFP genes across vertebrate lineages and roles of duplication and subfunctionalization in the diversification of the signaling systems*

INTRODUCTION

In the previous chapter, I introduced the duplication history of the relaxin family peptides and their receptors throughout vertebrate evolution. Briefly, I postulated that a receptor-ligand system encoded by three genes (*AncRln/Insl, AncRxfp1/2 and AncRxfp3/4*) already existed in the vertebrate ancestor before the onset of the two rounds (2R) of whole genome duplication (WGD) ~550 MY ago (see Figure 2.1 for a summary). In addition, using the model of vertebrate karyotype evolution, I demonstrated that the genetic origination of the *RLN/INSL-RXFP* system throughout vertebrate evolution has been strongly influenced by WGD events, which amplified the numbers of both ligand and receptor genes. The modern sets of genes present in vertebrate genomes were ultimately defined by differential gene loss and retention across lineages. For instance, in teleosts post-3R gene loss primarily affected the ligand-receptor pairs potentially involved in reproduction (rln-rxfp1, insl3-rxfp2), whereas *rln3/insl5* ligand and *rxfp3/4* receptor genes (potentially involved in neuroendocrine regulation) experienced comparably high retention rates. The other group of vertebrates, tetrapods, experienced a significantly higher loss of RXFP3/4 receptors in comparison to teleosts, and in some tetrapods, gene loss reduced the size of relaxin peptide and their receptor families to, probably, a biological minimum (as in reptiles, which will be discussed further in this chapter).

To date, the functional diversification of relaxin family peptides and their receptors has remained enigmatic, primarily due to insufficient information available on the molecular biology of the signaling system in non-placental vertebrates. One notable property of relaxin peptides, which has remained unexplained in evolutionary terms, is their ability to promiscuously interact with various receptors. Thus, although RLN/INSL3 and RLN3/INSL5 appear to belong to two functionally distinct signaling niches, one member of each peptide pair expresses a strong overlap in its binding ability of RXFP receptors from one or both classes (Figure B2). Hence, while the endogenous receptor of RLN3 is RXFP3, this peptide is also capable of activating RXFP4 and RXFP1 receptors in mammals (Halls et al. 2007) and, moreover, rxfp1 in zebrafish (Park et al. 2008). Another peptide, RLN, apart from its cognate receptor RXFP1, can also induce RXFP2-mediated signaling. Conversely, neither human INSL3 nor INSL5 show promiscuity in their interactions with RXFP receptors (Figure B1, in Background). The question arising from this is whether the observed promiscuity in ligand-receptor interactions of relaxin peptides and RXFPs is an artifact of shared ancestral origins of the four RLN/INSL ohnologs or whether this property was acquired *de novo* to accommodate certain needs of the vertebrate organism.

In this chapter I demonstrate how knowledge about the origins of the individual components of the system has a potential to predict the functional relationships among yet poorly understood genes in vertebrates. In particular, here I will:

a) look in detail at the loss and retention of both *RLN/INSL* (ligand) and *RXFP* (receptor) genes in different taxa, such as reptiles (including birds), non-placental mammals and teleosts, which have been, for the most part, neglected in the literature, to gain further insight into the ligand-receptor co-evolutionary process. In addition, I will discuss the novel *RLN/INSL* and *RXFP* genes found in elephant shark and lamprey.

b) In addition I will use the *RLN/INSL-RXFP* duplication model established in Chapter 1 as a theoretical basis to explain the functional diversification of relaxin peptide systems in vertebrates.

Figure 2.1. The number and identity of *RLN/INSL* (ligand) and *RXFP* (receptor) genes found and/or predicted in major extant and ancestral vertebrate taxa.
It has been postulated that the pre-2R vertebrate ancestor probably had two *RXFP* (*AncRxfp3/4* and *AncRxfp1/2*) genes and a single *RLN/INSL* (*AncRln-like*) gene. The first round of WGD (1R) duplicated the original set of genes to produce a total of 6 post-1R genes. The second round of WGD raised the number of *RLN/INSL* and *RXFP* genes to 12. The third round of WGD in teleosts led to another increase in numbers of both ligand and receptor genes. Crossed circles indicate gene loss.

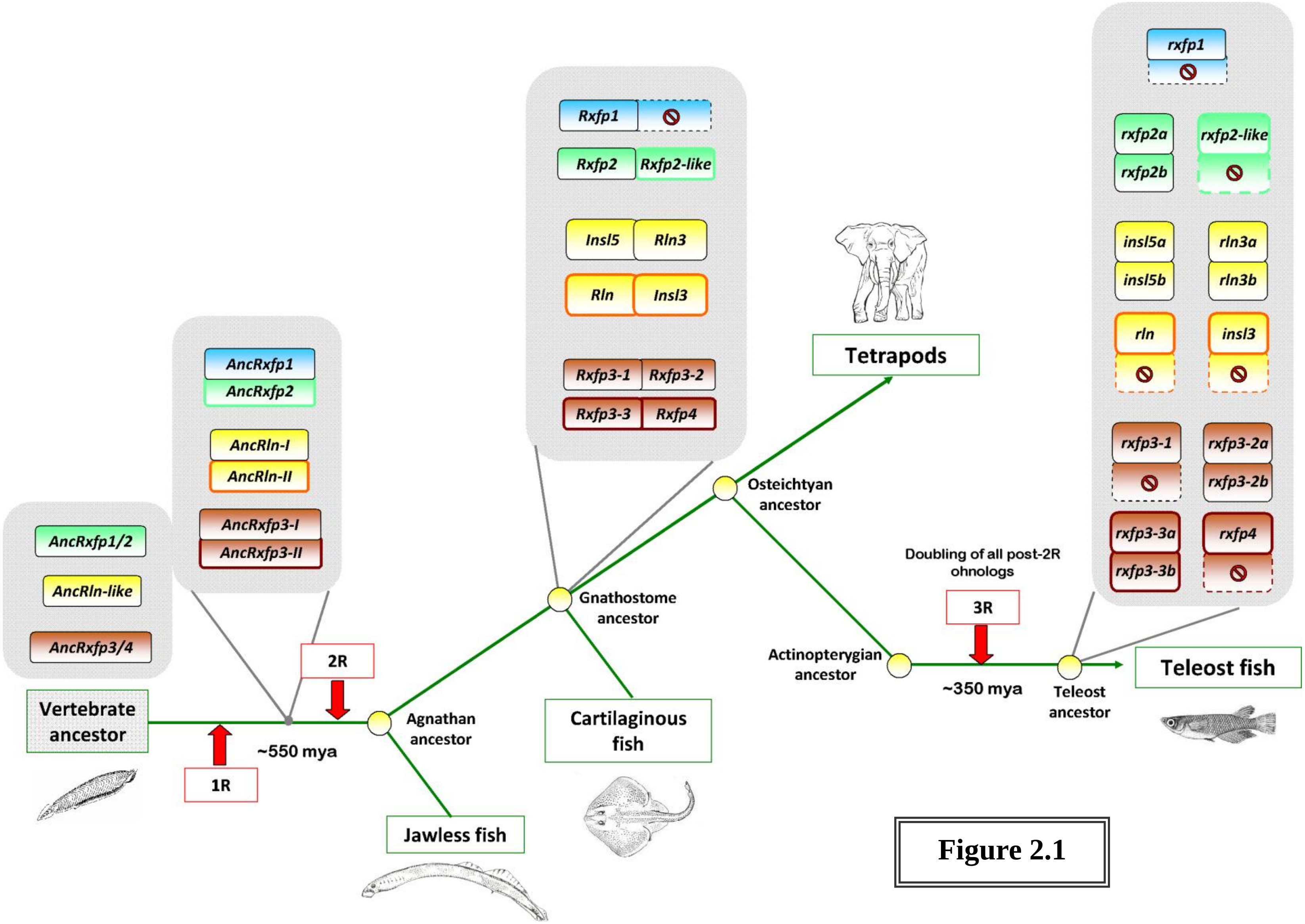

AncRxfp1/2
AncRln-like
AncRxfp3/4
Vertebrate ancestor
1R
~550 mya
AncRxfp1
AncRxfp2
AncRln-I
AncRln-II
AncRxfp3-I
AncRxfp3-II
2R
Agnathan ancestor
Jawless fish
Rxfp1
Rxfp2
Rxfp2-like
Insl5
Rln3
Rln
Insl3
Rxfp3-1
Rxfp3-2
Rxfp3-3
Rxfp4
Gnathostome ancestor
Cartilaginous fish
Osteichtyan ancestor
Tetrapods
Actinopterygian ancestor
Doubling of all post-2R ohnologs
3R
~350 mya
Teleost ancestor
Teleost fish
rxfp1
rxfp2a
rxfp2b
rxfp2-like
insl5a
insl5b
rln3a
rln3b
rln
insl3
rxfp3-1
rxfp3-2a
rxfp3-2b
rxfp3-3a
rxfp3-3b
rxfp4

Figure 2.1

METHODS

Retrieval of sequences for lamprey and cartilaginous fish

Tetrapod and teleost gene and protein sequences analyzed in this chapter are part of the dataset created for Chapter 1. Lamprey and elephant shark whole genome sequences (WGS) were searched in the NCBI Trace Archives (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using discontiguous megaBLASTn (Altschul et al. 1997) with teleost/tetrapod *RLN/INSL* and *RXFP* sequences as queries. The resulting trace hits were then checked for the presence of RLN/INSL and RXFP amino acid signatures by manual inspection and by employing the NCBI conserved-domain search tool or BLASTp (Altschul et al. 1997). Where necessary, multiple traces were combined based on overlapping regions. The lamprey candidate sequences were mapped to the pre-assembled lamprey genome (http://pre.ensembl.org) and were also used to blast the lamprey EST database. Due to the low quality of the current shark database, my findings were limited to shark *rln/insl*-like B-chain sequences. Additionally, *rln/insl-like* sequences were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/GenBank) for four cartilaginous fishes (dogfish, sand tiger shark, little skate and Atlantic stingray, one peptide sequence per species). Phylogenetic reconstruction was performed using the methodology presented in Chapter 1.

RESULTS and DISCUSSION

Part 1- Gene Loss and Gene Gain in Vertebrates

A close inspection of gene gain and loss patterns among different vertebrate taxa revealed interesting patterns of loss/retention and sometimes duplication of *RLN/INSL* (ligand) and *RXFP* (receptor) genes. The following briefly describes some of the highlights of the data mining (performed as part of Chapter 1) in teleost fish and tetrapods. In addition, gene gain and gene loss across placental mammals are discussed by looking at 12 placentals of different evolutionary origins.

TELEOSTS

Due to high retention rates and 3R duplication of both *rln/insl* (ligand) and *rxfp* (receptor) genes, the gene set in teleosts is the largest among vertebrates. Data mining and other analyses performed in this study in five teleost genomes confirmed that after 3R, teleosts retained 50% of the *rxfp3/4* and all of the *rln3* and *insl5* duplicates. At the same time, teleosts only retained single copies of *rln* and *insl3* and their receptors *rxfp1* and *rxfp2*. Zebrafish differs from the other four analyzed percomorphan species- stickleback, takifugu, tetraodon and medaka (Kinoshita et al. 2009) - in that it has retained both *rxfp2b* (3R duplicate of *rxfp2*) and *rxfp2-like* (2R-ohnolog, lost in most other vertebrates), but lost *rxfp4*. Additionally, a local duplication event in zebrafish led to the origination of *rxfp3-3a3* from duplication of *rxfp3-3a2* (Figure 2.3). The loss of *rxfp4* with the co-incidental gain of *rxfp3-3a3* in zebrafish, coupled with no changes in the number of their ligands, suggests that zebrafish rxfp3-3a3 may work with one or both copies of insl5. Differences between zebrafish and other teleosts are not limited to the existence of

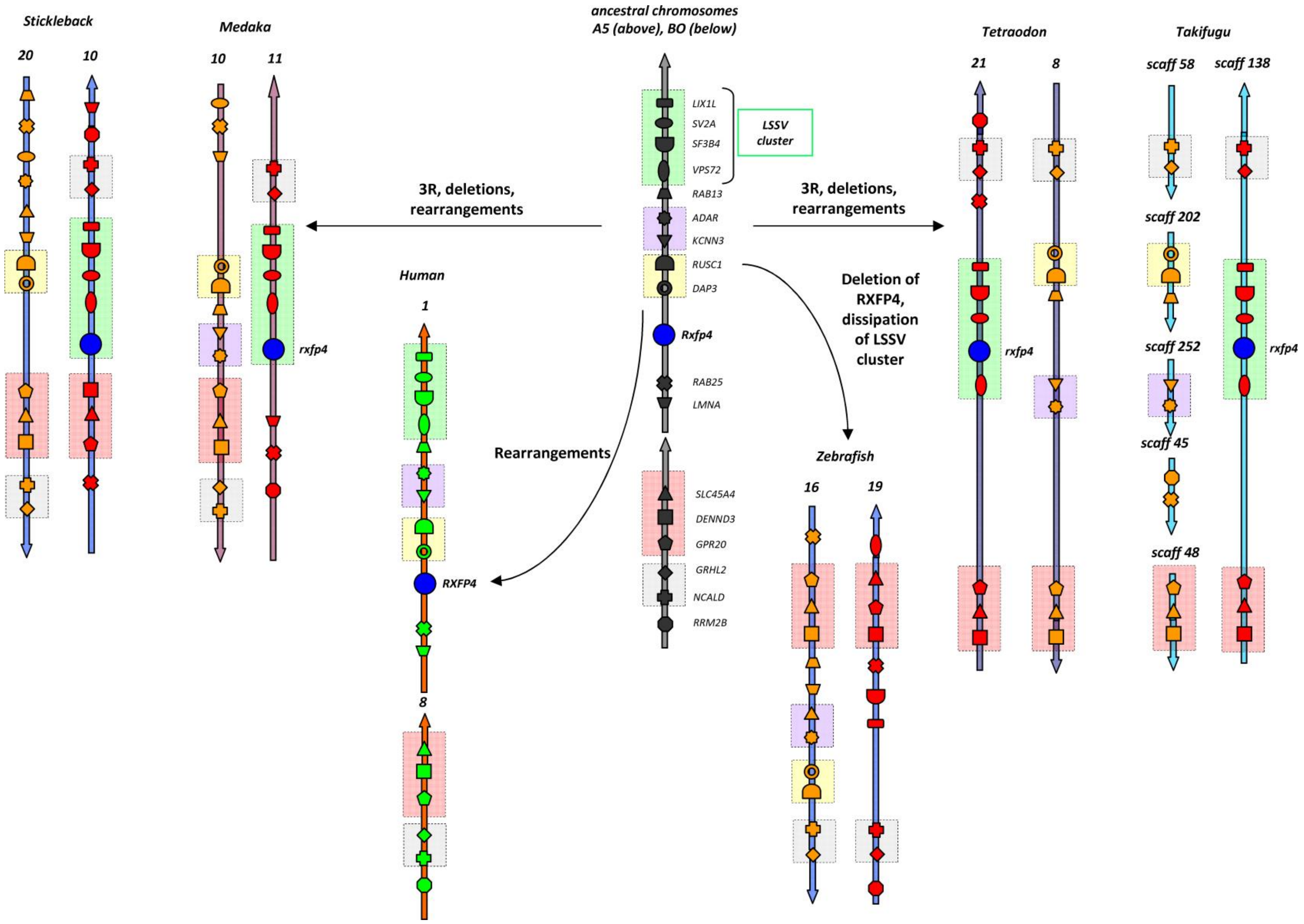

Stickleback
20
10
Medaka
10
11
rxfp4
ancestral chromosomes
A5 (above), BO (below)
LIX1L
SV2A
SF3B4
VPS72
LSSV cluster
RAB13
ADAR
KCNN3
RUSC1
DAP3
Rxfp4
RAB25
LMNA
SLC45A4
DENND3
GPR20
GRHL2
NCALD
RRM2B
3R, deletions, rearrangements
3R, deletions, rearrangements
Human
1
RXFP4
8
Rearrangements
Deletion of RXFP4, dissipation of LSSV cluster
Zebrafish
16
19
Tetraodon
21
8
rxfp4
Takifugu
scaff 58
scaff 138
scaff 202
scaff 252
scaff 45
scaff 48
rxfp4

Figure 2.2. Synteny map depicting rearrangements in the structure of the *RXFP4* paralogon in teleosts and human.
The hypothetical ancestral chromosome is shown to highlight the various recombination/duplication/deletion events that occurred in fish and human. 3R-paralogs in teleosts are denoted in orange and red. Colored boxes outline clusters of genes shared between organisms. The "LSSV" cluster of 4 genes (*LIX1L*, *SV2A*, *SP3B4* and *VPS72*) is linked to the *RXFP4* gene in all organisms but zebrafish, which seems to have lost its *rxfp4* gene together with the LSSV cluster. Alternatively the *rxfp4* gene in zebrafish may have undergone gene conversion (see *Figure 2.10*). Chromosomes depicted as arrows pointing upstream. Not to scale.

lineage-specific duplicates, but are also mirrored in phylogenetic trees, where zebrafish proteins are always found basal to other teleosts' proteins (see Figure 1.5). Such differences are attributable to the early divergence of zebrafish (~320 MYA) from the rest of the teleost lineage (Percomorphans diversified ~190 MYA) (Kasahara et al. 2007). Data mining of the smallest known teleost genome, the genome of of *Tetraodon nigroviridis*, was unable to identify *rln3b* or a pseudogene of it, but detected the presence of an *rxfp3-1* pseudogene. This suggests that, at least in tetraodon, rln3b could be a cognate ligand of rxfp3-1 and that this ligand-receptor system became non-functional in this species. This observation is interesting in light of the results of the experimental work on the expression of *rln3* genes in teleosts, which point at the subfunctionalization of teleost *rln3* paralogs and hence, most likely, also of their rxfp3 receptors (see section "Subfunctionalization in teleosts" below).

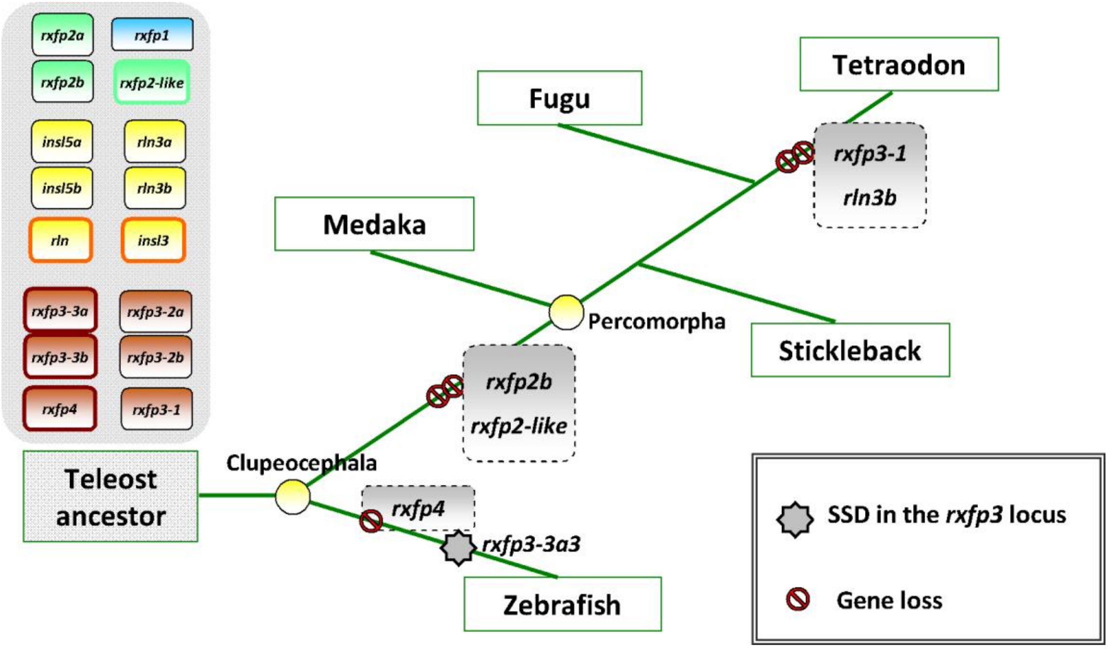


Figure 2.3. Post-3R gene loss and gain in five teleost fish species.
Teleosts start with the gene set composed of ten receptors and six ligands. Zebrafish, which diverged early from the ancestral teleost lineage, retain most of the focal genes, except *rxfp4*. At the same time, possibly as a replacement for *rxfp4*, zebrafish gain an additional copy of the *rxfp3-3* gene. Other teleosts lose *rxfp2-like* and also the 3R-duplicate *rxfp2b*. SSD: small-scale (local) duplication. Phylogeny and classification of fish adapted from Kinoshita, et al. (2009).

TETRAPODS

Many tetrapods have experienced loss of both *RLN/INSL* (ligand) and *RXFP* (receptor) genes. Notably, apart from the expansion of the RLN locus in placental mammals, differential retention of *RLN3/INSL5* genes and *RXFP3/4* ohnologs is the main cause of differences in the genes present in the relaxin peptide systems of tetrapods and teleosts. For instance, while teleosts fully retained (and duplicated) the post-2R repertoire of *rxfp3/4* ohnologs and *rln3/insl5* genes, tetrapod *RXFP3-2* completely disappeared while

RXFP3-3 was retained only in a few tetrapod groups, such as cow, pig and opossum (Figures 2.4 and 2.5). The presence of 4 ligands and 4 receptors in many tetrapods therefore implies a 1-to-1 ligand-receptor interaction, while in teleosts, in which the number of rxfp3/4 receptors exceeds that of their ligands, these interactions appear more promiscuous.

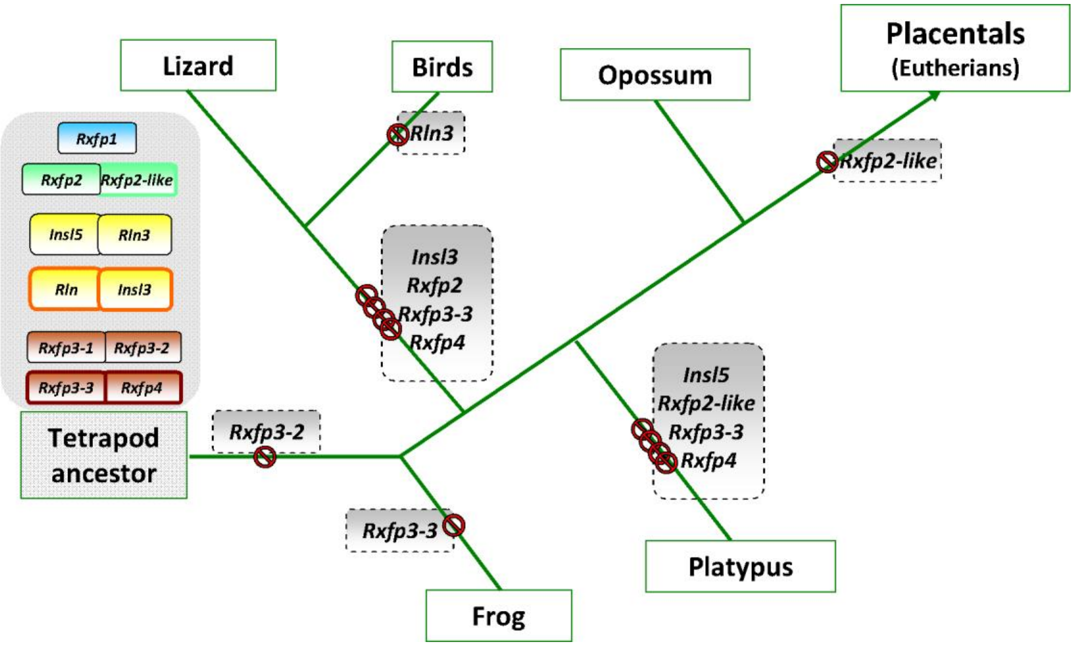


Figure 2.4. Post-2R gene loss and gain in tetrapods.
Tetrapods start with the gene set composed of seven receptors and four ligands. *RXFP3-2* is lost before the divergence of all extant tetrapod taxa. Interestingly, opossum is the only tetrapod (of the ones analyzed here), which has preserved the number of *RLN/INSL-RXFP* genes hypothesized to exist in the tetrapod ancestor. Massive gene loss occurred in reptiles (including the loss of the rxfp2-insl3 receptor-ligand pair in all reptiles and of rln3 in birds) and in platypus. See *Figure 2.5* for an expanded view of the situation in placental mammals.

While the phylogenetic relationships I observe among most tetrapod sequences are in agreement with other studies, there are also some unexpected findings, which include the relaxin from anole lizard (discussed below) and armadillo RLN branching off the base of the entire RLN/INSL clade (Figure 2.7). It is interesting that armadillo is well-known for

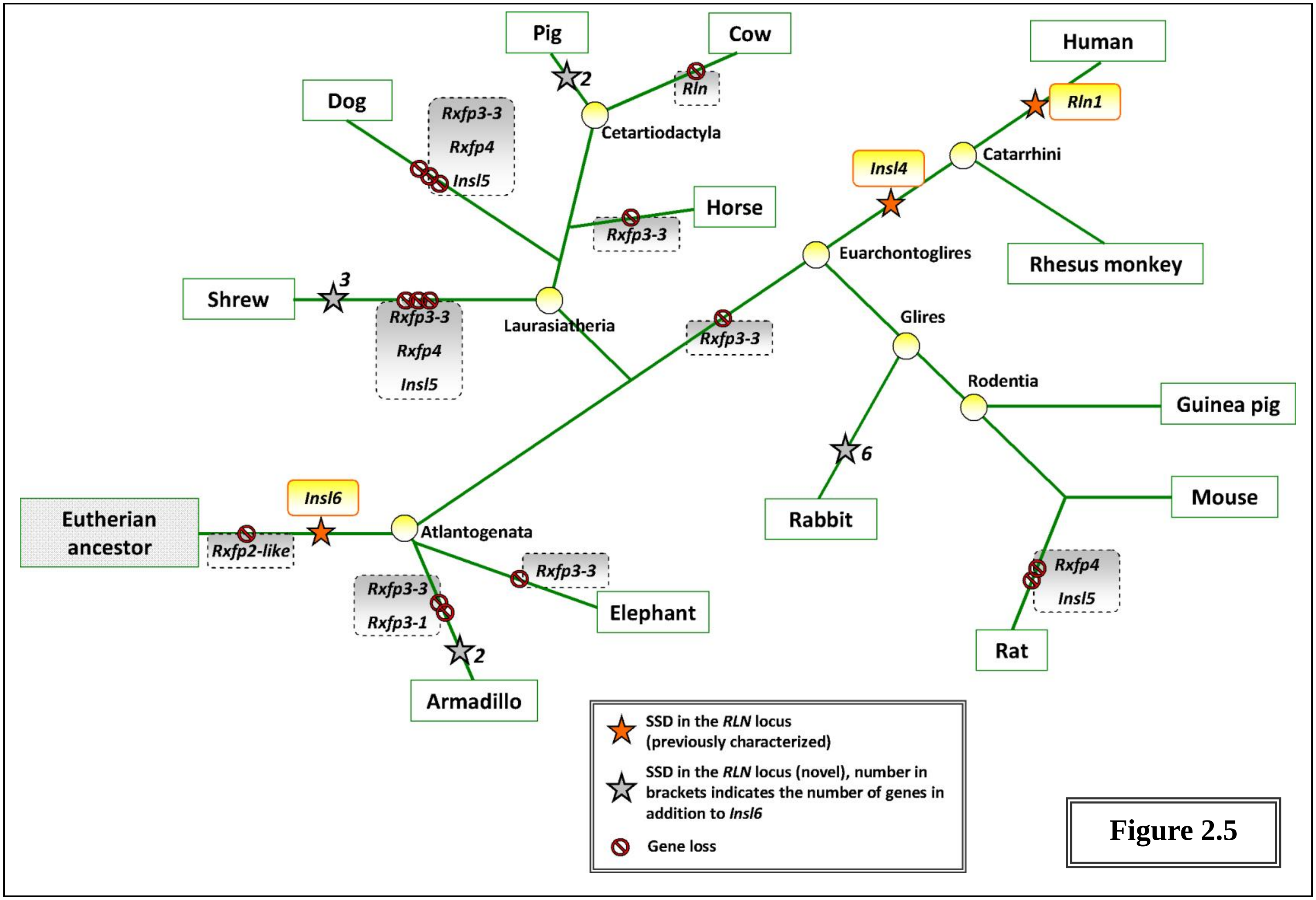

Pig
Cow
Human
Dog
2
Rln
Rxfp3-3
Rxfp4
Insl5
Cetartiodactyla
Rln1
Catarrhini
Insl4
Horse
Rxfp3-3
Euarchontoglires
Rhesus monkey
3
Shrew
Rxfp3-3
Rxfp4
Insl5
Laurasiatheria
Rxfp3-3
Glires
Rodentia
Guinea pig
6
Rabbit
Mouse
Eutherian ancestor
Insl6
Rxfp2-like
Atlantogenata
Rxfp3-3
Rxfp3-3
Rxfp3-1
Elephant
Rxfp4
Insl5
2
Rat
Armadillo
SSD in the RLN locus (previously characterized)
SSD in the RLN locus (novel), number in brackets indicates the number of genes in addition to Insl6
Gene loss

Figure 2.5

Figure 2.5. Gene loss and gain in placentals.
The evolution of placental RLN/INSL-RXFP systems is marked by a complete loss of *RXFP2-like*. Many placental lineages also lost the RXFP4-INSL5 receptor-ligand pair and most (except cow and pig) did not retain RXFP3-3, suggesting that their role has diminished in these lineages. At the same time, the *RLN*-locus tremendously expanded in multiple placental lineages through independent local duplications. Note: the branching of this tree reflects the branching order (adapted from Prasad et al. (2008), but not the timing of the divergence of taxa. SSD: small-scale duplication.

its ability to give birth to monozygotic quadruplets in every litter (Cyranoski 2009). This strategy is thought to be advantageous for overcoming the physical constraints imposed by the shape of the mammal's reproductive tract (Cyranoski 2009) and may potentially involve the functioning of RLN.

Reptiles (both reptiles *per se* and birds) present a particularly interesting model in which to study the diversification of the relaxin system. In addition to the gene loss encountered in other tetrapods, all reptiles appear to have lost the insl3-rxfp2 system and birds additionally lost rln3. The insl3-rxfp2 system has been shown to be important in the regulation of testicular descent in placental mammals (Feng et al. 2009) and is involved in the survival of both male and female germ cells in mammals (Kawamura et al. 2004). Furthermore, in teleosts, the expression and localization of insl3 hormone and rxfp2 receptor are similar to those in mammals (Good-Avila et al. (2009); Dr. J. Bogerd, personal communication). Thus it appears that the insl3-rxfp2 system plays an ancient role in germ cell survival and therefore the observed loss of both the *insl3* (ligand) and *rxfp2* (receptor) genes in reptiles implies that they use mechanisms different from other vertebrates for the regulation of germ cell survival.

Interestingly, while the loss of INSL5 and RXFP4 in rat was used to infer ligand-receptor specificity of the molecules in mammals (Wilkinson et al. 2005b), reptiles have lost rxfp4 but seem to have a functional insl5. Moreover, in birds insl5 is the only ligand that could potentially function via the single rxfp3/4-type receptor, rxfp3-1, retained in these organisms. Although RXFP3-1 is known to be the primary receptor for RLN3 in mammals (Halls et al. 2007), the finding that rxfp3-1 probably functions with insl5 in birds, suggests that either a switch in the ligand-receptor pairings occurred in reptiles, or that the rxfp3/4-type receptors can act promiscuously with either rln3 or insl5 ligands. Another interesting finding pertaining to the reptilian relaxin signaling peptides is the high sequence similarity of lizard rln and other vertebrate INSL5 peptides (Figure 2.7), suggesting that either gene conversion or selection has strongly influenced the evolution of the lizard rln. Overall these findings show that the reptile rln/insl-rxfp systems have evolved in a lineage-specific way, apparently different from other vertebrates, and that there may have been rearrangements in traditional receptor-ligand interactions.

PLACENTALS

Placental mammals are the largest source of information about the relaxin peptide signaling systems. In fact, to date the majority of the functional and bioinformatic studies on the RLN/INSL ligand and RXFP receptors has been done in a narrow range of placental taxa, such as murids, apes and human, which has ultimately limited our current understanding of the signaling system’s diversification to a few organisms. In this study, screening of multiple recently sequenced placental genomes has revealed that the placental *RLN/INSL* and *RXFP* gene sets are less uniform than previously thought, which implies that this signaling system has assumed various roles in different placental lineages to allow for lineage-specific adaptations. For instance, many placental mammals lost *RXFP4* and *INSL5* genes, and only a few (cetartiodactylans: cow and pig) have retained *RXFP3-3* in addition to *RXFP3-1* (Figure 2.5). Another example is that several placental taxa appear to possess lineage-specific duplicates of the RLN locus (Figure 2.5). While a few such duplicate genes (INSL6 in all placentals, INSL4 in catarrhines and apes and RLN1 in apes) were documented well by other authors and are believed to represent neofunctionalizations of the RLN locus (Wilkinson et al. 2005b), there are additionally multiple lineage specific duplications of the RLN locus that appear to have occurred independently. Specifically, there are three copies of a RLN-like gene in shrew, two in pig, two in armadillo and six in rabbit (Figure 2.5). Since some of the duplicate RLN genes could be attributed to errors in the current genome assemblies, further investigation into their properties is deemed important to establish their identities and possible roles.

Origins of INSL4 in placental mammals

Previous bioinformatic analyses revealed that the appearance of INSL4, at least in catarrhines, has been associated with the viral retroposition of the RLN locus that resulted in a tandem duplication (Bieche et al. 2003). Despite searching for the presence of INSL4-type genes, Bièche et al (2003) failed to find sequence evidence of INSL4 in earlier diverging mammals. Recently, Hoffman and Opazo (2011) proposed that the duplication of the RLN locus that gave rise to INSL4 occurred long before the emergence of catarrhines. Based on phylogenetic evidence in which catarrhine INSL4 always groups basal to RLN1/RLN2, the authors suggested that INSL4 appeared for the first time in Euarchontoglires (~103 MYA), but then was lost in all lineages with the exception of catarrhines and apes (which diverged ~ 30 MYA). One problem with the Hoffman and Opazo (2011) hypothesis is that it is based purely on phylogenetic evidence. When a gene has an accelerated rate of evolution compared to its ancestor, as is the case for INSL4 (see Background), which underwent neofunctionalization relative to its progenitor RLN (Wilkinson et al. 2005b), the branch often comes out basal to a clade (because purifying selection is relaxed, and the gene diverges), and indeed this is one way that pseudogenes are detected. Consequently, the phylogenetic position of catarrhine/ape INSL4 as basal to the RLN clade is not unexpected.

Second, both the bioinformatic analyses by Bièche et al. (2003) and the database screening performed here failed to find any evidence of INSL4 in earlier diverging placental mammals, such as murids. Third, a problem in the analyses of the INSL4 locus is that its peptide product is longer than the other RLN/INSL peptides, and the C-peptide can be included as part of the mature peptide. Phylogenetic analyses of the pre-propeptide forms (containing the C-peptide, normally excised during post-translational modification, see Figure B1) of INSL4 and other RLN-locus duplicates, suggest that catarrhine/ape INSL4 is sister group to more recently derived RLN molecules (Figure 2.5), leading to further evidence that Hoffman and Opazo's (2011) hypothesis is incorrect. Lastly, if Hoffman and Opazo's hypothesis were correct, *INSL6* and *INSL4* would have duplicated and diverged both at ~ 100 MYA and we would expect them to be approximately equally divergent from the *RLN* locus; however this is clearly not the case (see Figure 2.6).

In contrast to *INSL4*, *INSL6* was present in most of the placental mammals examined, and is relatively well conserved suggesting that after duplication and divergence from the progenitor RLN molecule, it acquired its new function relatively quickly. Experimental evidence shows that *INSL4* is highly expressed only in placenta (Bieche et al. 2003), while *INSL6* is highly expressed exclusively in testis (Ivell and Grutzner 2009). On the other hand, *RLN* has more general expression in reproductive tissues with high expression patterns in prostate (Samuel et al. 2003). That *INSL4* and *INSL6* were subject to neofunctionalization related to the appearance of novel reproductive functions in placentals is further suggested by their switch to new, and still undiscovered, receptors (Kong et al. 2010). Given that *RLN* in placental mammals is highly divergent from its ortholog in teleosts (see Chapter 3), it appears that the entire locus containing *RLN* and its local duplicates has been the target of selection.

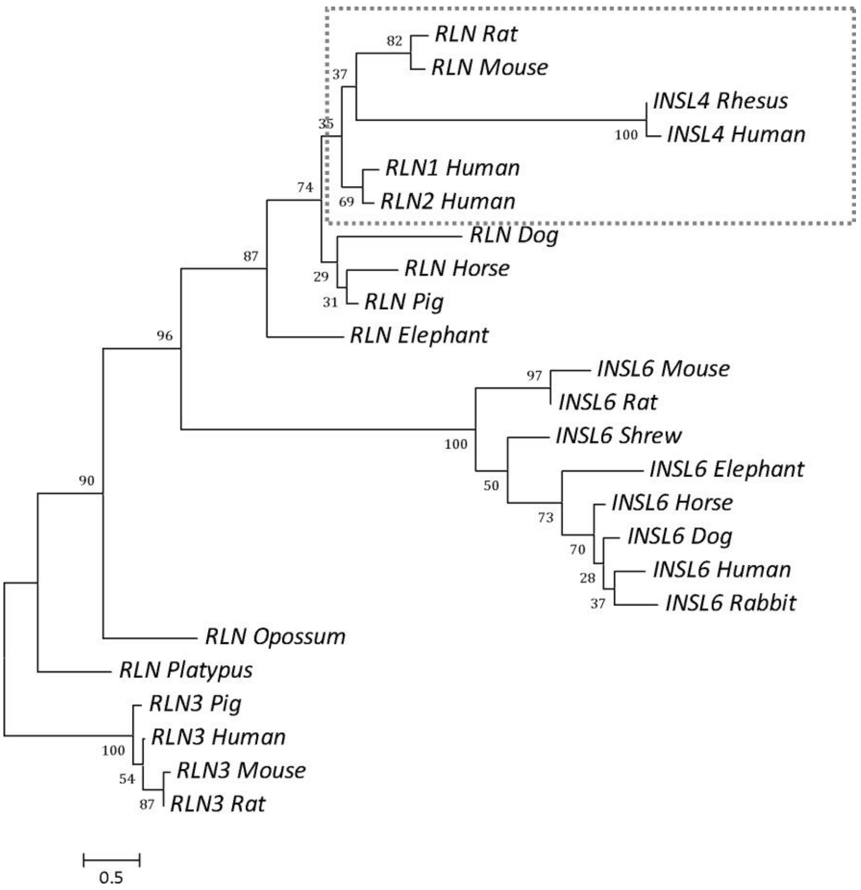


Figure 2.6. Phylogeny of the RLN peptides in mammals including sequences for RLN (together with RLN1/2 in humans), INSL4 (only present in monkey and primates), and INSL6.
RLN3 was used as outgroup.

Gene sets in lamprey and elephant shark.

To gain insight into the status of the relaxin ligand-receptor systems in early diverging vertebrates, such as agnathans and cartilaginous fish, I searched the sequenced genomes of lamprey and elephant shark using traditional database search strategies. These searches led to the discovery of two *RLN/RLN3*-like (ligand) genes in both lamprey and shark and four *RXFP*-like (three RXFP3/4-like and one RXFP1/2-like) genes in lamprey. Owing to the still poor assembly of the lamprey database and unassembled nature of the elephant shark genome, it seems unlikely that the sets of *rln/insl* and *rxfp* genes obtained for lamprey and shark are complete. The timing of divergence of lampreys and hagfish (collectively known as jawless fish or agnathans) in respect to the WGD events has been debated and presently agnathans are believed to have diverged post-2R (Kasahara 2007), which implies that these organisms must possess a set of *rln/insl* and *rxfp* genes of similar size (but not necessarily identical, because of lineage-specific gene loss and gain) to that of gnathostomes.

The novel *RXFP3*-like genes identified in lamprey, *rxfp3-L1* and *rxfp3-L2*, cluster together outside the vertebrate *RXFP3-1* and *RXFP3-2* clades, while lamprey *rxfp3-L3* is unplaced at the base of the tree (Figure D1a, Appendix D). It is possible that lamprey rxfp3-L1 and rxfp3-L2, which appear very similar structurally, are products of the same gene (e.g. they could be products of two alleles of a single gene; alternatively, if they are two distinct genes, gene conversion may have taken place- Dr. Campbell, personal communication). The lamprey RXFP1/2-like protein groups with RXFP2 and RXFP2-like sequences from other vertebrates (Figure D1b, Appendix D). The two novel lamprey rln/insl peptides that I identified cluster closely with rln from non-mammals and RLN3 providing good support for the orthologous relationship of these genes.

Two of the three RLN/INSL-like sequences experimentally identified by Schwabe's research group, from dogfish and sand tiger shark (Reinig et al. 1981, Steinetz et al. 1998), cluster outside the RLN/RLN3 clade, whereas another sequence from little skate branches off the INSL5 clade. At the same time the peptide from atlantic stingray, somewhat surprisingly, groups in the placental RLN clade. The elephant shark sequences showed a slightly different grouping: while one of them clustered in the RLN3 clade, the other clustered outside the RLN/RLN3 clade (Figure 2.7). In summary, it is clear that further studies must be conducted to more accurately define the identities of the novel genes from elephant shark and lamprey. Thus the constantly updated lamprey database (which has recently been added to the rapidly growing list of Ensembl genomes) holds promise to achieve enough chromosome coverage to allow the identification of lamprey rln/insl and rxfp genes through syntenic analyses.

Figure 2.7. Reconstruction of phylogenetic relationships among insulin-relaxin superfamily peptides of both deuterostomes (vertebrates and invertebrates) and protostomes (fruitfly).
Blue squares show novel relaxin family peptide-like (rfpl) sequences from lamprey and cartilaginous fish. Red squares surround anomalous tetrapod sequences which branch differently from their orthologs in other organisms. See a less collapsed version of this tree in Chapter 3.

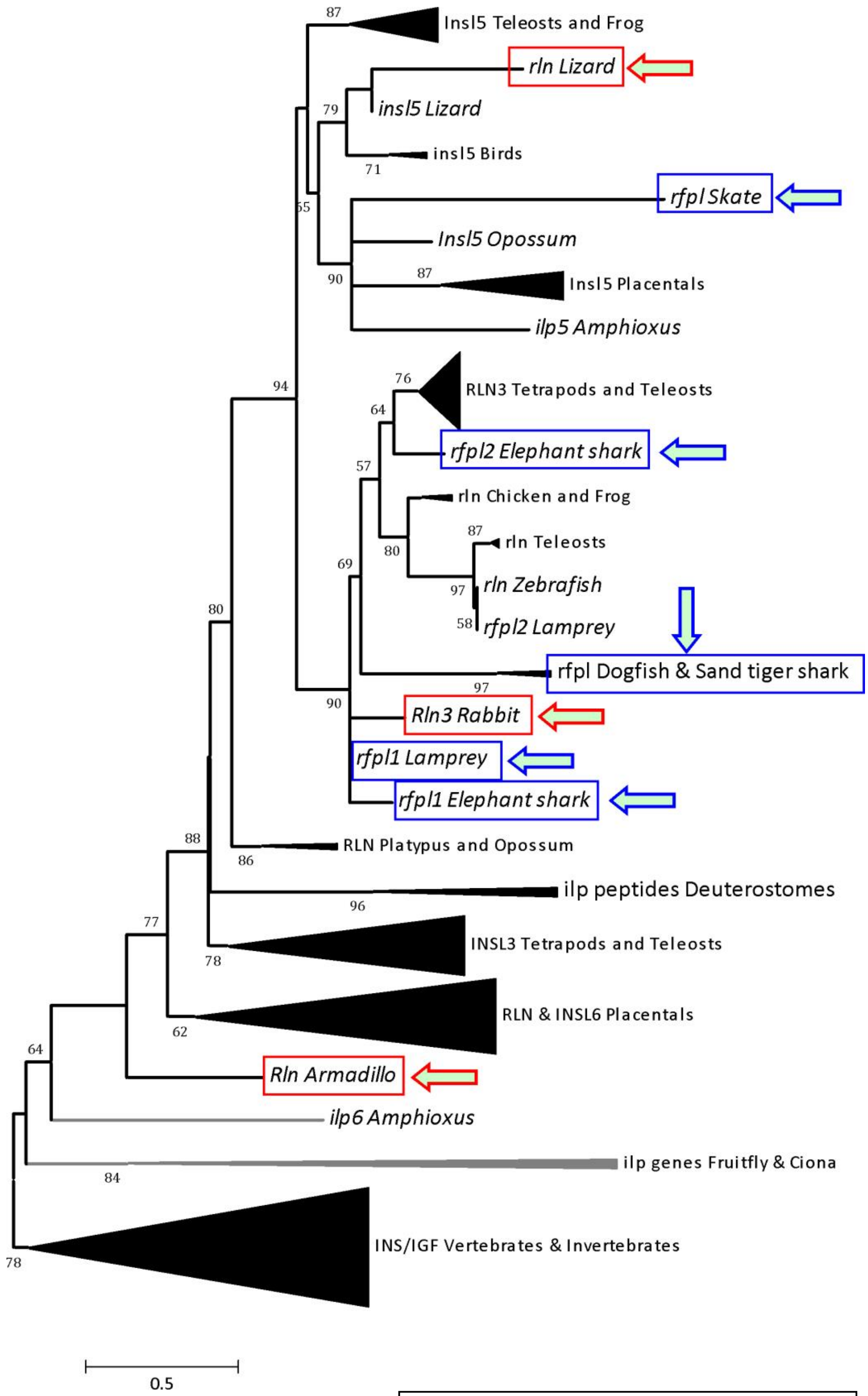


Figure 2.7 (Legend on previous page)

Part 2- The subfunctionalization of RLN/INSL and RXFP genes: hypothetical models

Subfunctionalization in a tripartite receptor-ligand-receptor system

Sub- and neofunctionalization as processes that are involved in the evolution of duplicated genes were introduced in Background using a scenario for a two-component system encoding a receptor and its ligand (Figure B5). There are at least two possible scenarios for the subfunctionalization of a more complex ligand-receptor system composed of one ligand and two receptors (Figure 2.8). The presence of two receptors in such a system implies that the sole ligand has a dual function. Thus it may regulate one kind of physiological process by signaling via one kind of receptor (e.g. receptor "*R*", Figure 2.8), and at the same time this same ligand may perform a different function using the other receptor (e.g. receptor "*S*", Figure 2.8). The duplication of both ligand and receptor genes will produce three pairs of daughter genes (*R'*-*R"*, *S'*-*S"* and *L'*-*L"*). Since the two daughter ligands are structurally identical, each of them can potentially function via 4 receptors. Subsequently, based on the needs of the genome and associated selection pressures (and assuming there is no receptor loss), the two ligands can form two daughter three-component signaling systems. Depending on whether there is a need to retain a dually functioning signaling system controlled by a ligand and two different kinds of receptors, the daughter systems may contain receptors of the same or different kinds (Figure 2.8).

The duplication model of *RLN/INSL* and *RXFP* genes (Chapter 1) proposes that the ancestral pre-duplication system was represented by a dually functioning ligand (AncRln-like) that used two receptors (AncRxfp3/4 and AncRxfp1/2). Furthermore, the modern vertebrate RLN/INSL-RXFP systems are the products of the two duplication events which amplified the genes for both ligands and receptors. Hence, if the evolution of the vertebrate relaxin ligand-receptor systems occurred via the processes of duplication and subsequent subfunctionalization, which of the two scenarios presented in Figure 2.8 did it most likely follow?

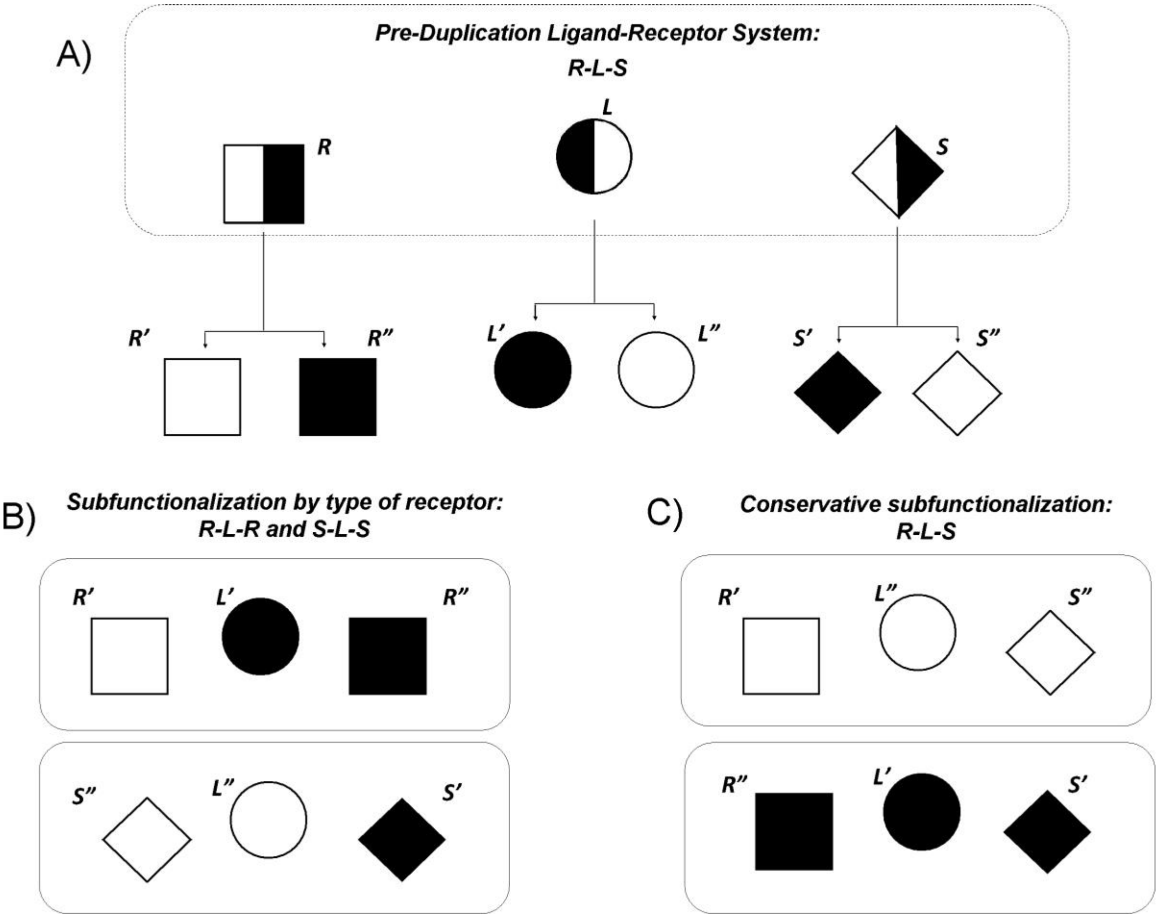

A)
Pre-Duplication Ligand-Receptor System:
R-L-S
R
L
S
R'
R"
L'
L"
S'
S"
B)
Subfunctionalization by type of receptor:
R-L-R and S-L-S
R'
L'
R"
S"
L"
S'
C)
Conservative subfunctionalization:
R-L-S
R'
L"
S"
R"
L'
S'

Figure 2.8. The two possible outcomes (B and C) of a subfunctionalization process taking place among the post-duplication descendants of a three-component receptor ligand system.
A) An ancestral system encoded by one ligand gene (*L*) and two receptor genes (*R* and *S*). After duplication, there are a total of six genes, which can subfunctionalize in at least two different ways: B) Each daughter ligand subfunctionalizes to work with only one type of daughter receptors (either *R* or *S*); or C) both ligands retain the ability to function via both types of receptors.

Using the theory of subfunctionalization to explain the functional diversification of the RLN/INSL-RXFP ligand-receptor system.

Our current knowledge about the endogenous receptor-ligand pairing of relaxin-like peptides and their receptors dictates that in mammals two of the four ohnologous peptides (RLN and INSL3) function via one kind of relaxin receptor (RXFP1/2-type, RXFP1 and RXFP2), while the other two ligands (RLN3 and INSL5) function via another kind of relaxin receptors (RXFP3/4, RXFP3 and RXFP4). The two kinds of relaxin peptide receptors are only distantly related and the exact mechanism by which the recruitment of such diverse receptors occurred has remained unclear. Taking into account that both ligand gene pairs, *RLN/INSL3* and *RLN3/INSL5*, are 2R products of two different post-1R ancestral genes, *AncRln-I* and *AncRln-II* (Figure 2.1), it can be hypothesized that the post-WGD daughter *RLN/INSL* genes subfunctionalized by specializing on one type of receptor (Figure 2.8B). This hypothesis further implies that relaxin peptides became distinctly paired with two different receptors already following 1R, when AncRln-I and AncRln-II subfunctionalized to interact with AncRxfp1/AncRxfp2 and AncRxfp3-I/AncRxfp3-II receptors (Figure 2.9A). The subfunctionalization of the post-1R system may have been triggered by the need to separate the two functions performed by the ancestral tripartite signaling system to produce two more systems specialized in reproductive (AncRln-I and AncRxfp1/2) and neuroendocrine regulation (AncRln-II and AncRxfp3-I/II).

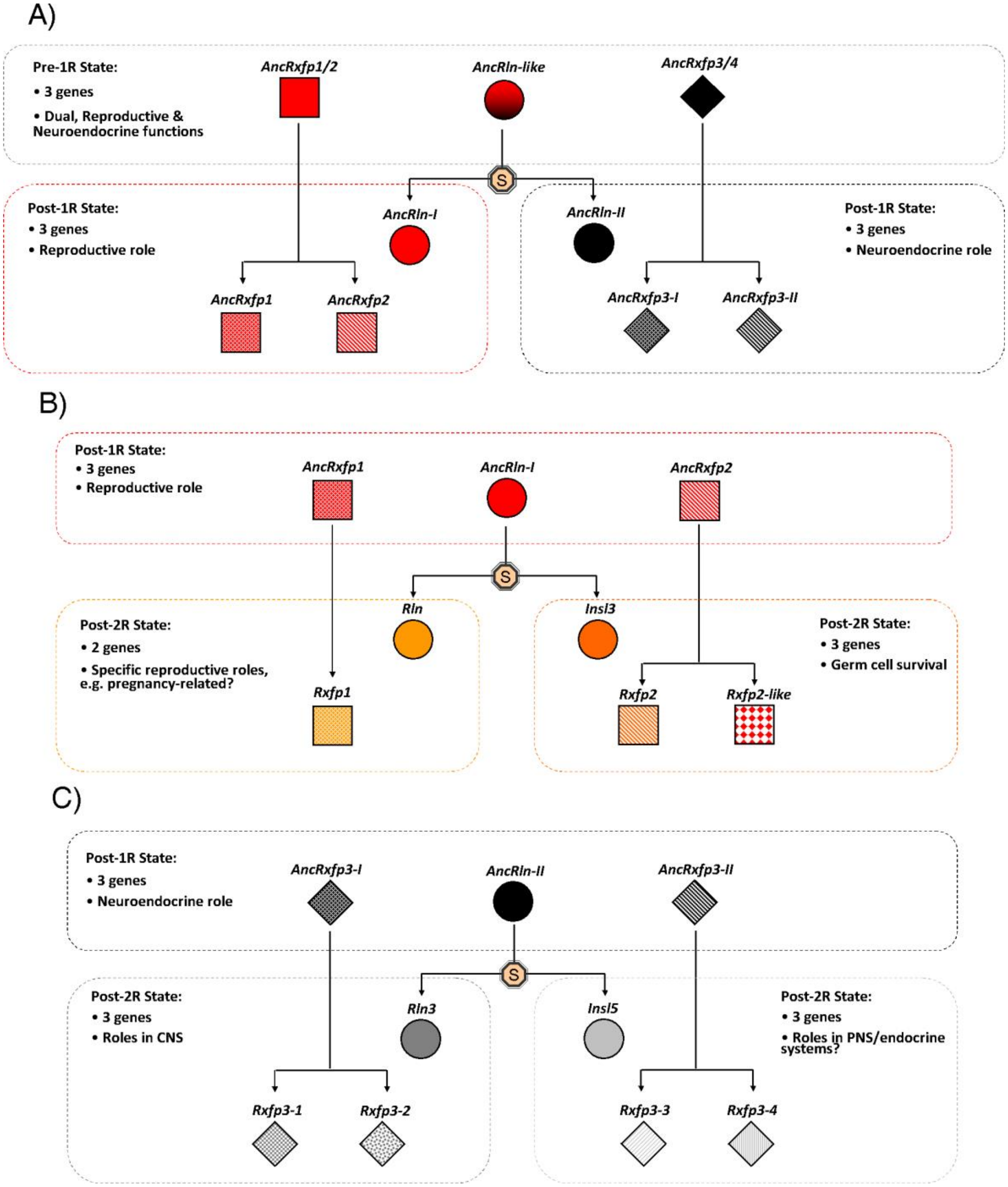
The second round of WGD completed the process of formation of modern vertebrate relaxin peptide gene sets by duplicating *AncRln-I* to produce *RLN* and *INSL3* and *AncRln-II* to produce *RLN3* and *INSL5*. While RLN3 became coupled with RXFP3-1 and RXFP3-2, INSL5 became coupled with RXFP3-3 and RXFP3-4. Notably, while most tetrapods lost *RXFP3-2* and *RXFP3-3*, all four *RXFP3/4* receptor ohnologs (and thus the three-component nature of the systems) were preserved in the teleost ancestor. At the same time, the evolution of the 2R products of AncRln-I ligand and AncRxfp1/2 receptors has taken very similar routes in both tetrapods and teleosts. In fact in all vertebrates there is a single *RXFP1* gene and in most vertebrates there is only one *RXFP2* gene, which indicates that the tripartite nature was not favored for these genes in modern vertebrates.

Figure 2.9. The diversification of the *RLN/INSL* and *RXFP* genes in the ancestor of jawless and jawed vertebrates.

A) The pre-1R three gene system duplicates to give rise to two ligands and two pairs of receptors. After duplication, both ligands and receptors are structurally and functionally identical, which is favorable for promiscuous ligand-receptor interactions. Such unobstructed ligand-receptor interaction in combination with certain selective pressures may have triggered the subfunctionalization of ligand genes, favoring the establishment of AncRln-I-AncRxfp1 and 2 and AncRln-II-AncRxfp3-I and -II as ligand-receptor pairs. Note that here I stress that the two post-1R *AncRln* (ligand) genes subfunctionalize to work with the two different types of receptors. Alternatively, each of the daughter ligand genes could have formed a system which would imitate the ancestral receptor-ligand system in that each ligand would still work with two different types of receptors.

B) Further subfunctionalization of the AncRln-I 2R-products: RLN, which subfunctionalizes to work with RXFP1 and INSL3, whose physiological target becomes RXFP2. On the basis of proposed relatedness of RXFP2-like to RXFP2, I hypothesize that RXFP2-like could, at least shortly after 2R, function as a receptor of INSL3.

C) Subfunctionalization of post-2R *AncRln-II* duplicates resulting in RLN3 and INSL5 which subfunctionalize to function via RXFP3-1/3-2 and RXFP3-3/3-4 receptors respectively. Since all tetrapods lost RXFP3-2 and most of them also lost RXFP3-3, their ligand-receptor pairs lost their ancestral three-component nature and became two-component, i.e. RLN3-RXFP3-1 and INSL5-RXFP4. Teleosts, on the other hand retained all post-2R RXFP3/4 receptors and seem to have experienced further subfunctionalization with the formation of complex ligand-receptor relationships.



Subfunctionalization of the system in teleosts

Teleost fish are different from other vertebrates in that they experienced an additional round of WGD, which amplified the gene sets established in the post-2R gnathostome ancestor. Post-3R selective gene loss and lineage-specific duplications determined the look of modern teleost genomes. The implication of this is that studying the gene systems of teleosts and comparing them to their tetrapod counterparts is illustrative for understanding the mechanisms involved in post-WGD evolution of genes and signaling networks.

There are 6 *rln/insl* genes in teleosts: *rln*, *insl3*, *rln3a*, *rln3b*, *insl5a* and *insl5b,* two thirds of which (i.e. *rln3a/b* and *insl5a/b*) are the products of 3R. The relaxin peptides that technically were lost after 3R (rln and insl3) are counterparts of tetrapod RLN and INSL3 which are involved in reproduction and whose receptors are RXFP1 and RXFP2 respectively. Both rxfp1 and rxfp2 receptors in most teleosts (like their putative ligands rln and insl3) have no 3R paralogs and are hence found as single copies. The latter implies that, like in tetrapods, rln-rxfp1 and insl3-rxfp2 are also endogenous ligand-receptor pairs in teleosts, which is supported by emerging experimental evidence from zebrafish (Dr. J. Bogerd, personal communication).

The situation with the counterparts of tetrapod RLN3/INSL5 and their RXFP3/4 receptors in teleosts is rather complex, because: 1) teleosts retained all rln3/insl5 3R duplicates and half of the rxfp3/4 receptor duplicates, and 2) post-3R lineage specific small-scale duplications further increased the number of *rxfp3/4* (receptor) genes without affecting the number of ligand genes. Thanks to these changes in the number of genes, the ligand-to-receptor ratio of relaxin peptide systems in teleosts is intermediate between that of tetrapods (approximately 1:1 in most) and the pre-3R teleost ancestor (1:2, Figure 2.9). Interestingly, the intensive diversification of rxfp3-2 and rxfp3-3 receptors observed in teleosts is the opposite of almost complete loss of these receptors in tetrapods, which implies more intricate relationships among teleost rxfp3/4-receptor systems.

Experimental studies performed in zebrafish (Donizetti et al. 2009) and eel (Hu et al. 2011) indicate that *rln3a* and *rln3b* exhibit the signs of spatial subfunctionalization, in that one of the paralogs (*rln3a*) is expressed in a broader range of tissues than the other (*rln3b*). Whether *rln3a* and *rln3b* subfunctionalized to function with different receptors *in vivo* has yet to be determined, but ongoing expression studies of teleost rxfp (receptor) genes already imply that this could be the case (Dr. Jan Bogerd, personal communication). Here I present a hypothetical model for the functional subfunctionalization of rln3 paralogs to work with rxfp3-1 and rxfp3-2 receptors (Figure 2.9A). Similar to the models previously derived for the subfunctionalization of RLN/INSL hormones and their receptors in the common ancestor of teleosts and tetrapods (Figure 2.8), it is assumed here that Rln3 peptide and Rxfp3-1 and Rxfp3-2 receptors form a tripartite ancestral teleost ligand receptor system, which is duplicated by means of 3R. Due to post-duplication loss of one *rxfp3-1* paralog, but retention of both duplicates of *rxfp3-2*, there are a total of 3 receptors which potentially function with rln3a and rln3b. Taking into account that in one of the percomorphan teleosts, in tetraodon, the loss of rln3b coincides with the pseudogenization of rxfp3-1 (see above), I propose that rln3b is a cognate ligand of rxfp3-1, while rln3a has specialized to function with two receptors, rxfp3-2a and rxfp3-2b (Figure 2.9A).

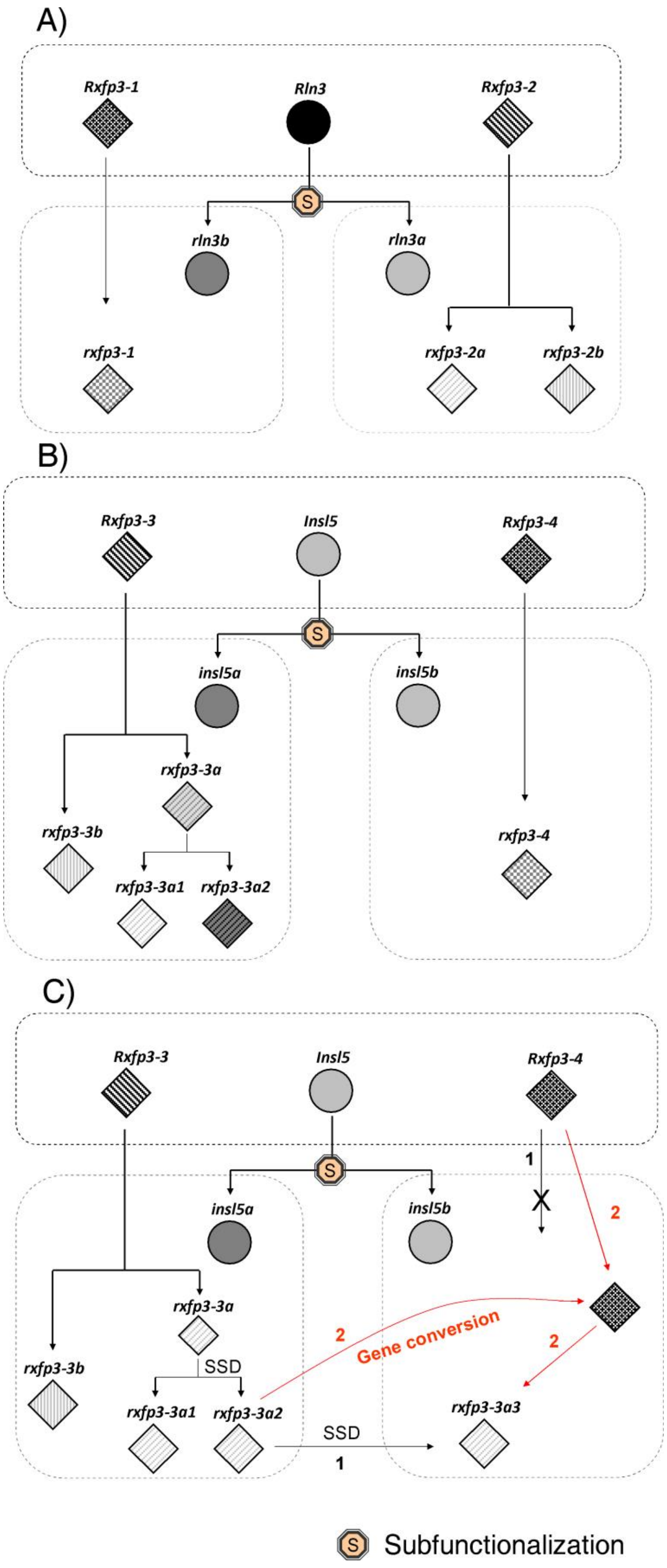

A)
Rxfp3-1
Rln3
Rxfp3-2
S
rln3b
rln3a
rxfp3-1
rxfp3-2a
rxfp3-2b
B)
Rxfp3-3
Insl5
Rxfp3-4
S
insl5a
insl5b
rxfp3-3a
rxfp3-3b
rxfp3-4
rxfp3-3a1
rxfp3-3a2
C)
Rxfp3-3
Insl5
Rxfp3-4
1
S
X
2
insl5a
insl5b
rxfp3-3a
2
Gene conversion
2
SSD
rxfp3-3b
rxfp3-3a1
rxfp3-3a2
rxfp3-3a3
SSD
1
S
Subfunctionalization

Figure 2.10. The diversification of rln3 and insl5 signaling systems in teleosts. Pre-3R teleost ancestor had two receptor-ligand-receptor trio systems, Rln3-Rxfp3-1/3-2 and Insl5-Rxfp3-3/3-4. Note that in both systems 3R ligand duplicates were completely retained, whereas only a half of receptor duplicates was kept. The receptor paralogs that were retained are descendants of *rxfp3-2* and *rxfp3-3*. A) Applying the principles outlined earlier in this chapter, one can hypothesize a functional specialization of the two *rln3* paralogs to work with *rxfp3-1* (*rln3a*) and two *rxfp3-2* genes. B) rxfp3-3 and rxfp3-4 receptors in percomorphans C) Zebrafish has lost its rxfp3-4 (i.e. rxfp4) receptor but has an extra copy of rxfp3-3a3, which may imply that the receptor of insl5b is rxfp3-3a3. There are two mutually exclusive ways (SSD, small-scale duplication, shown with number 1 (black) and gene conversion, shown as “2” [in red]) through which rxfp3-3a3 may have arisen. Note that in B) and C) insl5 paralogs are chosen arbitrarily and their interaction with receptors can be reversed, i.e. insl5a may function with rxfp3-4 and insl5b may interact with rxfp3-3 receptors.

At the same time the high structural similarity of rln3a and rln3b is probably associated with their equal ability to bind both kinds of receptors *in vitro* in the absence of spatial limitations characteristic to *in vivo* systems.

The story of insl5 paralogs and rxfp3-3 and rxfp3-4 receptors (Figure 2.10B) seems startlingly similar to that of rln3 and its putative receptors (Figure 2.10A) in that the 3R duplicates of only one class of receptor (rxfp3-3) survive gene loss (and go through additional lineage-specific duplications), while the other receptor (rxfp3-4) is retained in one copy in Percomorphans (Figure 2.9B) and is completely lost (according to the SSD scenario) or converted into a rxfp3-3-like gene (according to the gene conversion scenario) in zebrafish (Figure 2.9C). Thus here I hypothesize that while the endogenous receptor of one of the insl5 paralogs is rxfp3-4, the other insl5 gene uses three rxfp3-3 receptors.

CHAPTER 3: *Evidence of co-evolution between ligand-receptor pairs: analysis of the types and strengths of selection on RLN/INSL and RXFP genes in mammals versus teleosts*

INTRODUCTION

The ultimate sources of gene evolution are mutation and recombination.When mutations occur in protein coding genes, they may have negative, neutral or positive effects. When mutations occur at so-called degenerate sites, they do not cause changes to the peptide sequence, and are called silent or synonymous changes. Since these mutations are not “perceived” by the organisms, they tend to evolve at a rate largely dictated by the rate of mutation/substitution, i.e. they are neutral, although for proteins exhibiting high functional constraint even synonymous substitutions can be selected against(Nei and Kumar 2000). On the other hand, when mutations cause amino acid changes, so called non-synonymous changes, they may result in conservative or radical amino acid changes, and this can have negative or occasionally positive effects. Usually, these nonsynonymous changes have negative impacts on protein structure and/or function and they are selected against, a process known as purifying selection. However, sometimes these mutations may lead to amino acid changes that are favoured by the organisms in which they exist and this leads to a rapid fixation of such substitutions as the result of positive selection. Since ligands and receptors co-evolve, it is assumed that if an amion acid change is selected in one member of the ligand-receptor pair, a concomitant change will occur in the other member of the pair(Nei and Kumar 2000). One of the tenets of co-evolutionary theory of ligand-receptor pairs, is that they should exhibit similar types and rates of evolution and selection(Fraser et al. 2002), since a radical change in the active site of one member of the pair should be mirrored by a concomitant change in the other member..

Given this, a variety of authors have proposed that a test for co-evolution of ligand-receptors pairs is to calculate the evolutionary distance among putative ligand and receptor pairs for a suite of taxa . Potentially co-interacting pairs should exhibit similar rates of evolution, and thus have a correlation coefficient close to one, as has been shown for several co-evolving ligand-receptor pairs(Cyranoski 2009, Prasad et al. 2008). One of the assumptions of this test is that, aside from being assured that one is comparing orthologous genes and true ligand-receptor pairs(Braasch et al. 2009), is that the same selective forces operate among all taxa included in the analyzed group. A second potential caveat is that it assumes that a similar proportion of amino acids are subject to selection in each ligand-receptor pair, which may not be true, particularly if an entire binding pocket is the unit of selection in a receptor whereas only a few key residues are points of selection for the ligand. Nevertheless, tests of positive correlation among ligand and receptor pairs have been fruitfully employed in several studies. The correlation coefficient employed in these cases is normally calculated as the average amino acid distance among putative ligand-receptor pairs for the taxa included in the analysis (Goh et al. 2000), however, the test ould be employed for other parameters, such as the proportion of sites under purifying, neutral or positive selection in ligands and receptors respectively.

Additionally, sometimes ligand-receptor pairs undergo lineage-specific selection. For example, the teleostean peptide rln is highly similar to rln3 (in teleosts), while its mammalian counterpart (RLN) is highly divergent. This strongly suggests that the hormone RLN underwent positive selection in early mammalian history. To look for evidence of differential strengths of selection across genes or lineages, one can compare the ratio of the mean number of non-synonymous (d_N=amino acid replacing) to synonymous (d_S=silent sites that do cause a change in the amino acid) changes across the entire peptide. For proteins that are subject to strong purifying selection, this ratio is typically close to 0 ($0<d_{N/}\ d_S<0.2$, while for proteins subject to strong positive selection it is greater than 1.0 (i.e. $d_{N/}\ d_S>1$), with values between 0 and 1 indicating intermediate levels of selection (Nei and Kumar 2000). Because different parts of a protein are typically subject to different forms of selection, and because the active sites of proteins may be subject to novel forms of selection in distinct lineages, a better test of lineage specific selection is to look for evidence that specific amino acid exhibit a ratio of $d_{N/}\ d_S$ >1. Zhang et al. (Zhang et al.) developed a test look for this, called the branch-site test of positive selection, in which one looks for evidence of codon-specific positive selection within monophyletic clades as specified on a phylogenetic tree.

In this chapter, I will look at the role of selection by comparing the relative proportion of sites under purifying, neutral or positive selection in relaxin family peptide and receptor genes in both tetrapods and teleosts and examine if there is evidence of codon-specific positive selection in distinct lineages. The goal of the first part of the analyses is two-fold: by comparing the relative strength of purifying, neutral and positive selection in the focal genes in teleosts and tetrapods, I can assess whether orthologous ligand and receptor genes have experienced similar selective pressures in the two lineages. If they have, this suggests that the ligands and receptors may play similar roles in the two groups. Second, I calculate the correlation coefficient of putative ligand-receptor pairs separately for tetrapods and teleosts, to test whether putative ligand-receptor pairs have undergone similar levels of selection in the two lineages. If they have, this again, suggests that the proposed ligands and receptors function together and that their main roles have been preserved in the two lineages. Because the number of focal genes in teleosts is higher, this analysis is restricted to the orthologs of the teleost genes present in tetrapods, and I also assume that the probable ligand-receptor pairings in all teleosts and tetrapods are those that predominate in humans (i.e. RLN-RXFP1; INSL3-RXFP2; RLN3-RXFP3; INSL-RXFP4). Third, I present a phylogenetic reconstruction of the relationship among all relaxin family ligands to illustrate the evolutionary relationship among the sequences in the light of the results of the analyses of the selective forces that have shaped the appearance of modern genes. Lastly, I perform tests of codon-specific positive selection on the receptor (this chapter) and ligand genes (Chapter 4) to determine the amino acid positions and regions of the receptor genes that have undergone selection luding those genes that are teleost specific.

METHODS

Given the assumption that ligand and receptor pairs experience similar kinds and levels of selection, I calculated the proportion of amino acids in ligand and receptor pairs estimated to be subject to purifying, neutral or positive selection. These data were used to: 1) graph the proportion of sites under each kind of selection in all tetrapod and teleost genes and 2) plot the proportion of sites under each kind of selection for the four orthologs hypothesized to be ligand and receptor pairs in both tetrapods and teleosts. Second, to assess whether teleoestean ligand or receptor genes have been subject to lineage specific positive selection, I estimated the number and position of amino acids subject to positive selection in all genes. Both of these analyses were performed using the branch-site model A (Zhang et al. 2005), which tests whether the members of a user-defined clade (branch) on a phylogenetic tree exhibit evidence of codon-specific selection for the gene under study. The application of this model requires that the user specify *a priori* which branch is being tested for evidence of positive selection, the so-called foreground branch, while the remaining groups are defined as background branches. Tests of positive selection were made by comparing the branch-site model A in which (dn/ds) >1 (alternative hypothesis) to the model A in which dn/ds = 1 fixed (null hypothesis) and setting the foreground branch to the base of the clade containing the relaxin family ortholog in teleosts and the background branch to the same ortholog in mammals or tetrapods (depending on the tree structure) or vice versa. Analysis of the branch-site model A was done using CODEML from the PAML package (PAML v. 4.2); models were compared using the Likelihood Ratio Test with 1 degree of freedom and, where significant, the posterior probability that a codon was under positive selection was estimated using the Bayes empirical Bayes (BEB) procedure (Zhang et al. 2005).

Additionally, to further illustrate how selection has modified the perceived similarity of the relaxin family peptides, a phylogenetic tree was reconstructed based on the alignment of all relaxin family peptides (except insl4, see Appendix B), and including insulin and IGF–like peptides from diverse metazoan taxa to root the tree. The alignment of peptide sequences was performed using the algorithm MUSCLE as implemented in MEGA 5.01 (Tamura et al. 2011).The best model of sequence evolution was chosen using Maximum Likelihood inference as implemented in the program PROTEST (Abascal et al. 2005) and the RLN/INSL phylogeny was inferred using Maximum Likelihood methods as implemented in MEGA v. 5.01 (Tamura et al. 2011) using the JTT + Γ model of sequence evolution.

RESULTS

Using the proposed ligand-receptor pairings presented in Chapter 2, putative ligand-receptor pairs were graphed adjacent to one another, with the exception of the receptor rxfp2-like, whose ligand is unknown. If ligands and receptors co-evolve, we expect to observe a correlation of the rates and types of selection on ligand-receptor pairs, which is easily visualized on a histogram (Figure 3.1). Similarity between selection at ligand and receptor genes is observed for some loci but not others (Figure 3.1). For example, the patterns of selection for *RLN3-RXFP3-1*, *INSL3-RXFP2*, and *INSL5-RXFP4* in mammals are broadly similar: *RLN3-RXFP3-1* both evolve slowly and are characterized by purifying selection, *INSL3-RXFP2* evolve somewhat faster but have very similar selection profiles, but mammalian *INSL5* has more neutrally evolving sites than *RXFP4*, and the overall INSL5-RXFP4 system exhibits relaxed evolutionary pressures compared to RLN3-RXFP3-1 and INSL3-RXFP2 (Figure 3-1A). While these three ligand-receptor pairs have similar selection profiles, the same cannot be said of the mammalian RLN-RXFP1 system: RLN has more selected sites than any other gene, while RXFP1 exhibits a similar rate of evolution to RXFP2 and RXFP4.

The analysis of co-evolution of *rln/insl* and *rxfp* genes in teleosts, including those arising during 3R, was somewhat inconclusive (Figure 3-1B). In teleosts, the ligand rln was found to have a high number of neutrally evolving sites, although this may be an artifact of it being compared to mammalian rln (see Discussion). On the other hand, the number of selected sites in the proposed ligand-receptor pairs insl3-rxfp2, and rln3a/rln3b-rfp3-1 and rxfp3-2b look similar. In Chapter 2, I proposed that rxfp3-2a and rxfp3-2b are both potential receptors for rln3a and rln3b, and based on the selection analyses presented here it appears that the selection profile of rxfp3-2b parallels that of both ligands (rln3a/rln3b), while that for rxfp3-2a has more selected sites, haring a selection profile more like that of insl3.

In teleosts, as in mammals, insl5 exhibits higher rates of neutral evolution than the other ligands. However, none of the proposed receptors for insl5 exhibit the same high rate of neutral evolution: rxfp4 is the fastest evolving potential receptor, while all of the rxfp3-3 receptors are quite slowly evolving (Figure 3-1B). Thus, although teleost insl5 and rxfp4 genes have similar selection profiles to those of mammals, suggesting a conserved function between the two lineages, the other three proposed receptors for insl5 exhibit strong purifying selection and do not closely parallel the selection profile of teleost insl5.

In the second analysis, I plotted the relationship between the proportion of sites under purifying, neutral or positive selection in ligand versus receptor pairs using the following pairing rules: rln-rxfp1, insl3-rxfp2, rln3-rxfp3-1, and insl5-rxfp4 in both tetrapods and teleosts. Those values falling along the (0,0; 1,1) plane of the XY-plot exhibit similar kinds and strengths of selection between ligand-receptor pairs. Observation of a similar X,Y value for the same gene for mammals and teleosts for the same gene, further suggests that the pair may play similar roles in the two lineages. As Figure 3.2A clearly shows, the extent of purifying selection has been highly similar between mammals and teleosts for all RLN/INSL and RXFP genes. Moreover, the values of purifying selection are highly similar for both the ligand and receptor genes for RLN3-RXFP3 and INSL3-RXFP2 suggesting co-evolution, while for the remaining two genes, RLN-RXFP1 and INSL5-RXFP4, the proportion of sites under purifying selection is higher for the receptor genes (between 0.7 and 0.92), than the ligands (ranging from 0.4 to 0.95), suggesting a more diffuse co-evolution (or no co-evolution)

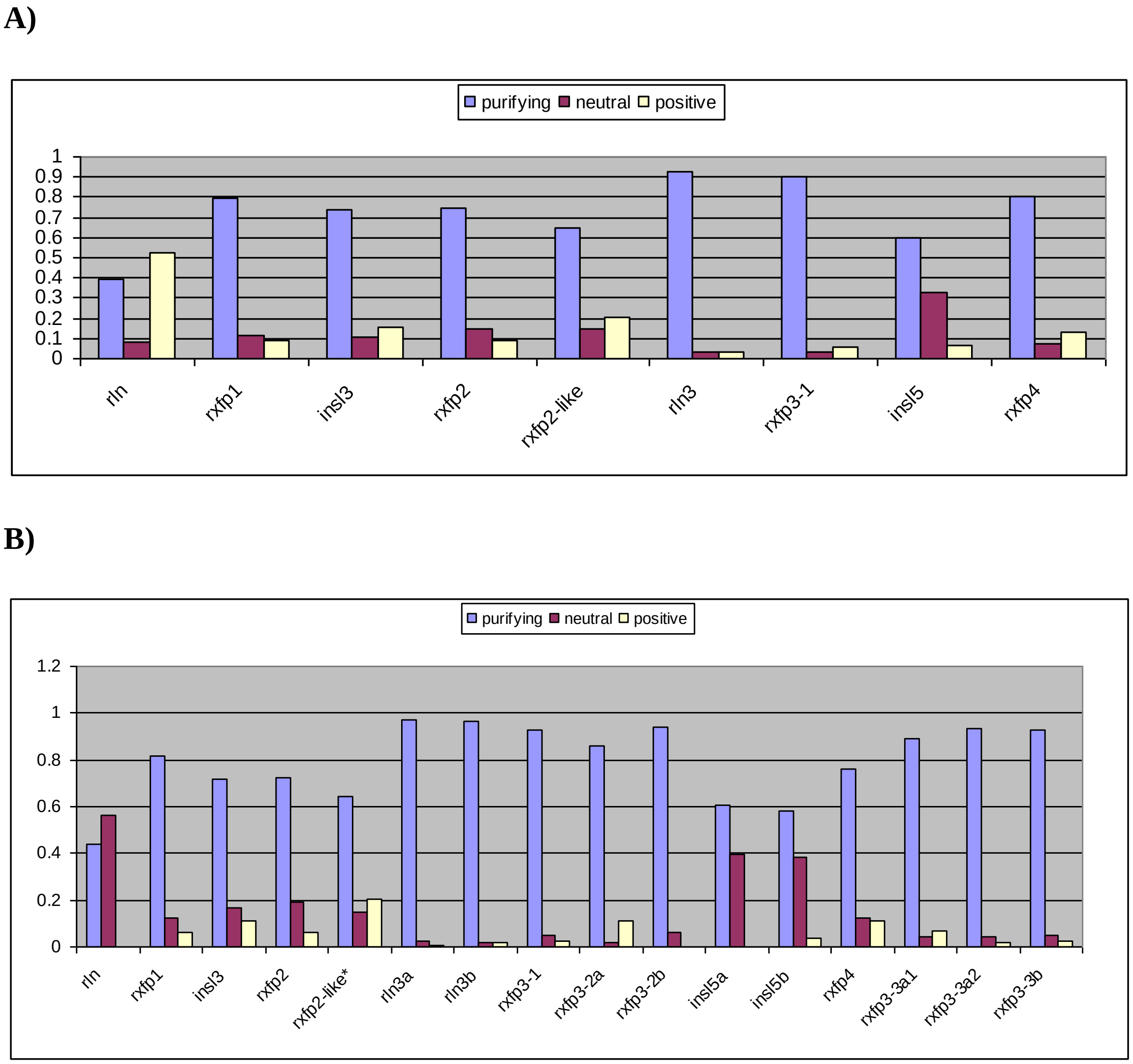
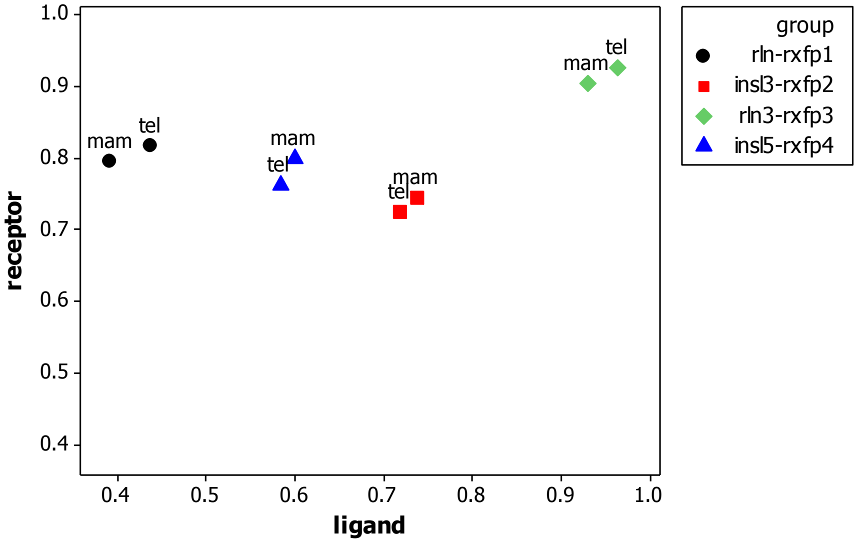


Figure 3.1. The proportion of sites in ligand and receptor genes subjected to different kinds of selection in mammals (panel A) and teleosts (panel B).

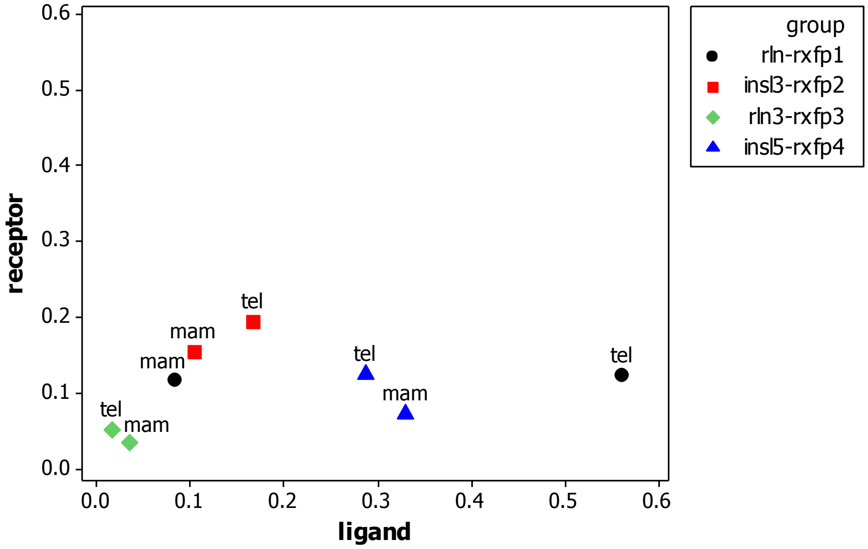
Selection types: purifying (light purple), neutral (dark purple) and positive (yellow).

On the other hand, there are significantly fewer sites, not surprisingly, which are evolving neutrally (Figure 3.2B) or are subject to positive selection (Figure 3.2C). For the receptor genes, from 3 to 20% of the sites were found to be evolving neutrally (Figure 2.3B), and from 2 to 13% were subject to positive selection; RXFP3 exhibits the fewest neutral or positively selected sites, RXFP4 has the highest proportion of sites under neutral evolution and RXFP2 exhibits the highest proportion under positive selection. Due largely to the anomalous nature of asymmetric selection on the RLN-RXFP1 ligand-receptor system in mammals, the extent of neutral and positive selection among ligand genes varied more widely, primarily because teleost rln was found to have a large number of sites evolving neutrally, whereas mammalian RLN has a large proportion of sites subject to positive selection (Figure 3.2B and 3.2C respectively). Thus, with the exception of the RLN-RXFP1 system, teleost and mammalian ligand-receptor pairs continue to reveal similar levels of neutral and positive selection suggesting similarity in their function.

A) Proportion of sites under purifying selection



B) Proportion of sites under neutral evolution



C) Proportion of sites under positive selection

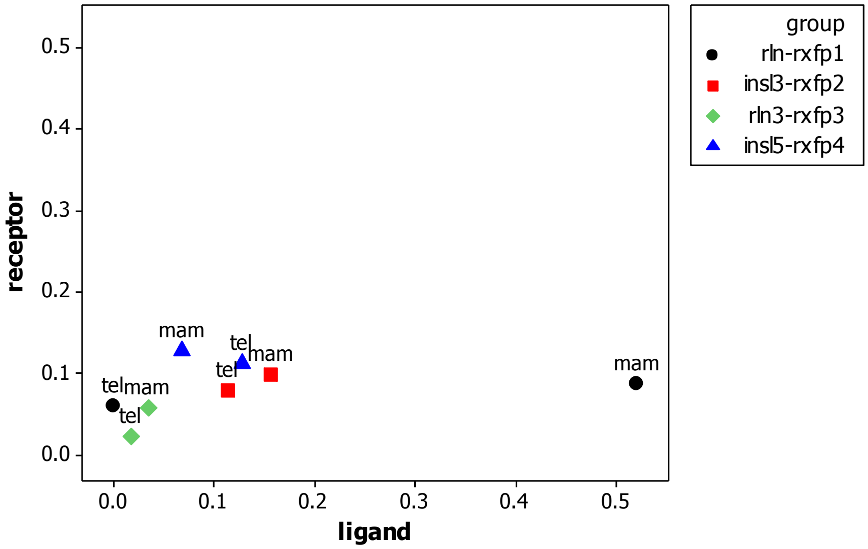


Figure 3.2. Estimated proportion of sites in the ligand (X-axis) and receptor (Y-axis) evolving under a) purifying b) neutral and c) positive selection in putative ligand-receptor pairs of the RLN/INSL-RXFP system in mammals and teleosts.
Those values falling along the XY: 0,0:1,1 plane represent pairs in which the same proportion of sites are observed to be under selection in both the ligand and receptor.

Imprint of selection on the phylogeny of relaxin family peptides

The signature of the different kinds and degrees of selection acting on the relaxin family peptides can be seen in a phylogenetic reconstruction of the relationship among this family of peptides (Figure 3.3). At the top of the tree, all of the INSL5 genes group together, with moderate support, into three well supported subclades (mammals + amphioxus, skate, teleosts, frog and reptiles). The branch lengths within each subclade are long, indicating that there is considerable variation among sequences, predominantly caused by neutral divergence of INSL5 genes as shown above. Subtending this clade, is a clade containing the RLN3 sequences of all vertebrates included in the analysis. This clade has high bootstrap support and exhibits the shortest branch lengths of any relaxin family peptide, a characteristic of genes subject to strong purifying selection. Clustering as a sister group to vertebrate RLN3, is a clade containing teleostean, amphibian (frog) and bird (chicken) rln. As described more fully in Chapter 4, rln sequences of non-mammals are highly similar to panvertebrate RLN3, and, as shown here, are characterized by purifying and neutral evolution. While these teleost and early vertebrate rln sequences are highly rln3-like, marsupial and monotreme RLNs are divergent from their teleost counterparts, and placental RLN is so divergent that it falls into its own clade, as a sister group to INSL6 (Figure 2.7). The clade containing placental RLN does not group with its teleostean, or marsupial/monotreme, orthologs because of the action of positive selection which has caused it to diverge so significantly that it appears as an independent gene. The long branches that characterize the clade containing mammalian RLNs is also caused by the high sequence divergence of the peptides which, in this case, is caused by positive, rather than neutral, evolution (see Chapter 4). Lastly, the action of moderate levels of positive selection on the INSL3 peptides is responsible for the loose clustering of the peptides from teleosts, mammals, monotremes and amphibians, although each subclade is characterised by long branches as expected for the divergent sequences included in them.

Thus, collectively, the relaxin family peptides are an interesting example of the diverse forms of selection that can act on peptides. They illustrate a classic example of the difference between gene trees versus species trees, since without the aid of synteny data, it would be very difficult to determine which genes are orthologs versus paralogs based on the phylogenetic reconstruction alone (discussed in Good-Avila et al., 2009). .

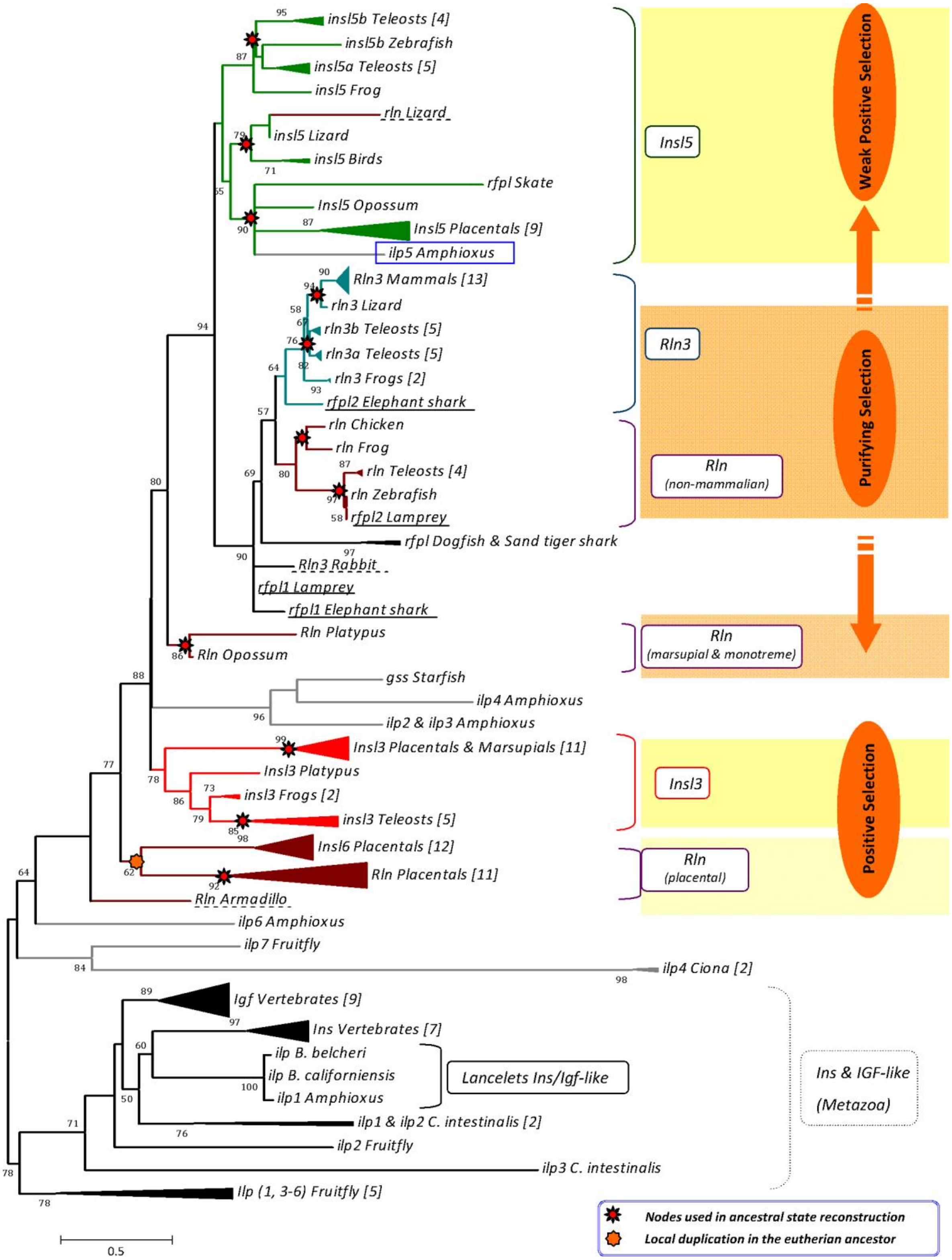


Figure 3.3. Phylogenetic reconstruction of the evolutionary relationships among the insulin-relaxin superfamily peptides.
Nodes shown (blue box) pertain to the ancestral reconstruction performed in Chapter 4.

Codon-specific positive selection in receptors

The test for lineage-specific codon selection revealed that the number of amino acids subject to positive selection varied considerably according to the region in which the amino acids were located. When all tetrapod RXFP1/2 genes were compared, the proportion of positively selected sites was found to be, in general, higher in LDLa (Low density lipoprotein-a) and the LRR (Leucine-rich repeat) than in the 7TM (seven transmembrane) domains. Intracellular loop 3 (ICL3) exhibited the highest proportion of sites under positive selection with 50% of the sites in the domain showing evidence of positive selection in across vertebrate lineages (Figure 3.4)

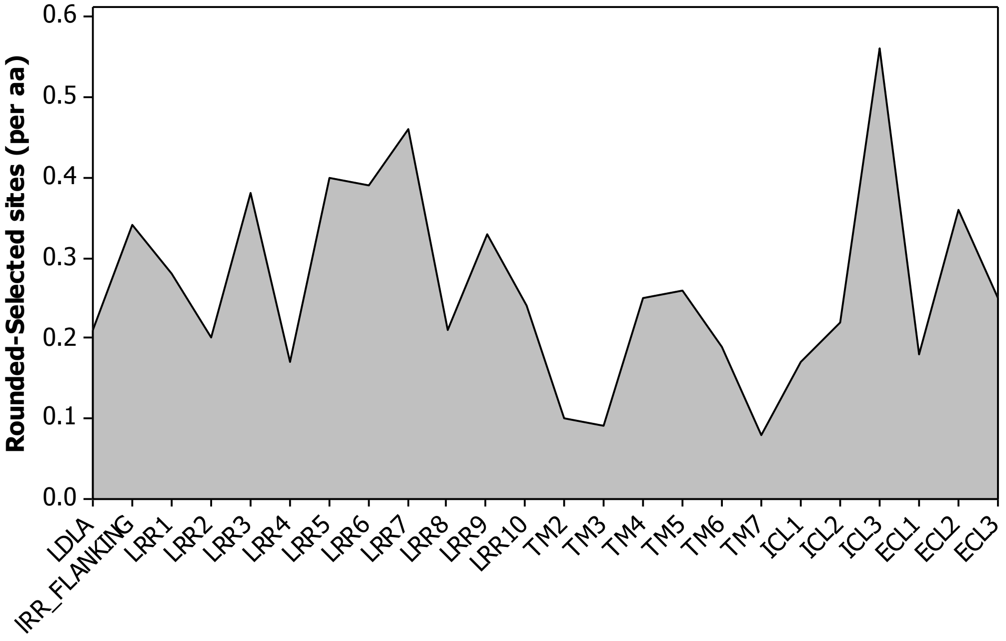


Figure 3.4. Area plot of the number of amino acids per region of the RXFP1/2-type receptors that showed evidence of positive selection when comparing teleost and tetrapod RXFP1/2 proteins.
The specific amino acids estimated to be under selection are shown in Figure 3.7.

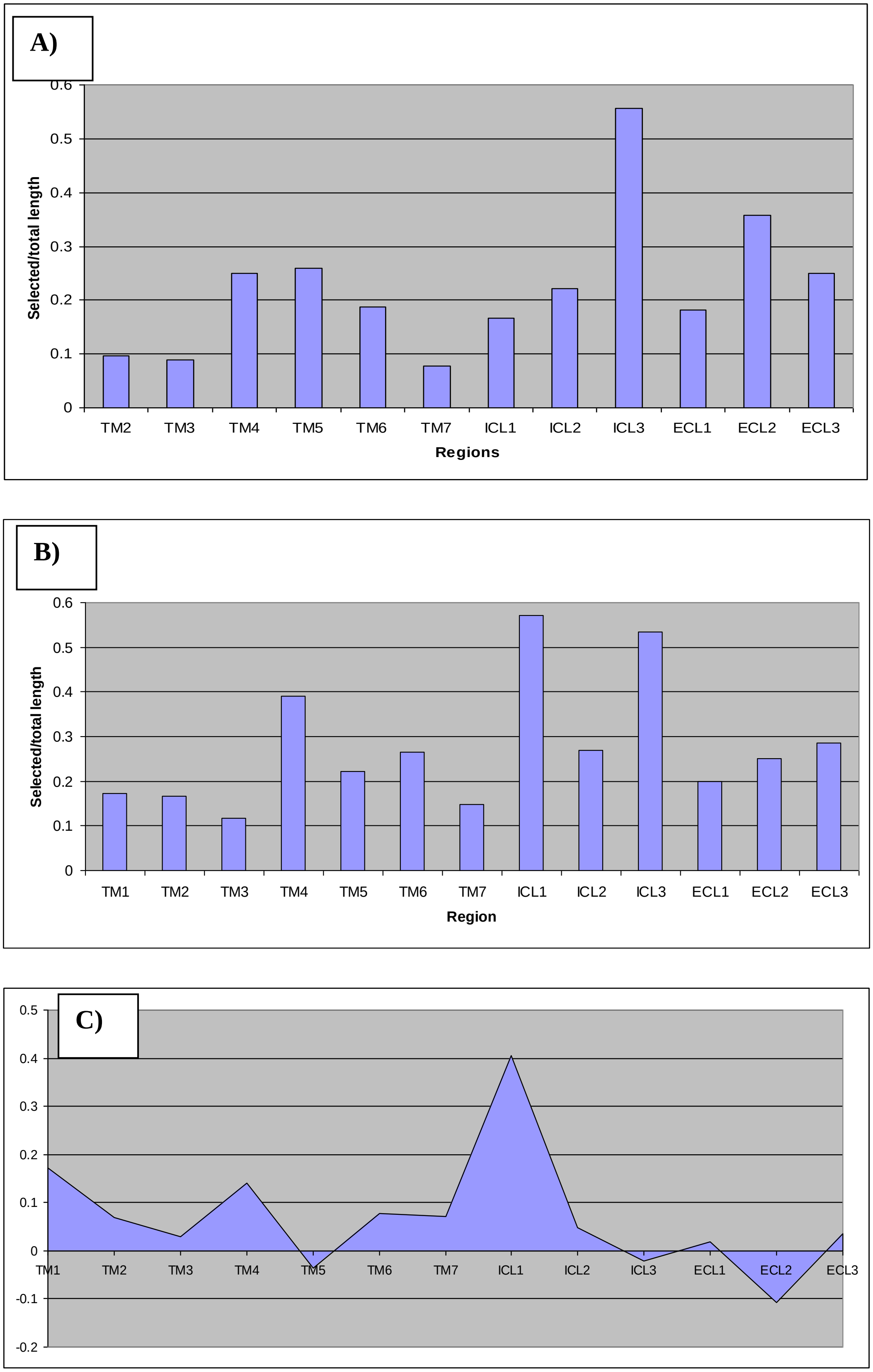

A)
Selected/total length
0
0.1
0.2
0.3
0.4
0.5
0.6
TM2
TM3
TM4
TM5
TM6
TM7
ICL1
ICL2
ICL3
ECL1
ECL2
ECL3
Regions
B)
Selected/total length
0
0.1
0.2
0.3
0.4
0.5
0.6
TM1
TM2
TM3
TM4
TM5
TM6
TM7
ICL1
ICL2
ICL3
ECL1
ECL2
ECL3
Region
C)
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
TM1
TM2
TM3
TM4
TM5
TM6
TM7
ICL1
ICL2
ICL3
ECL1
ECL2
ECL3

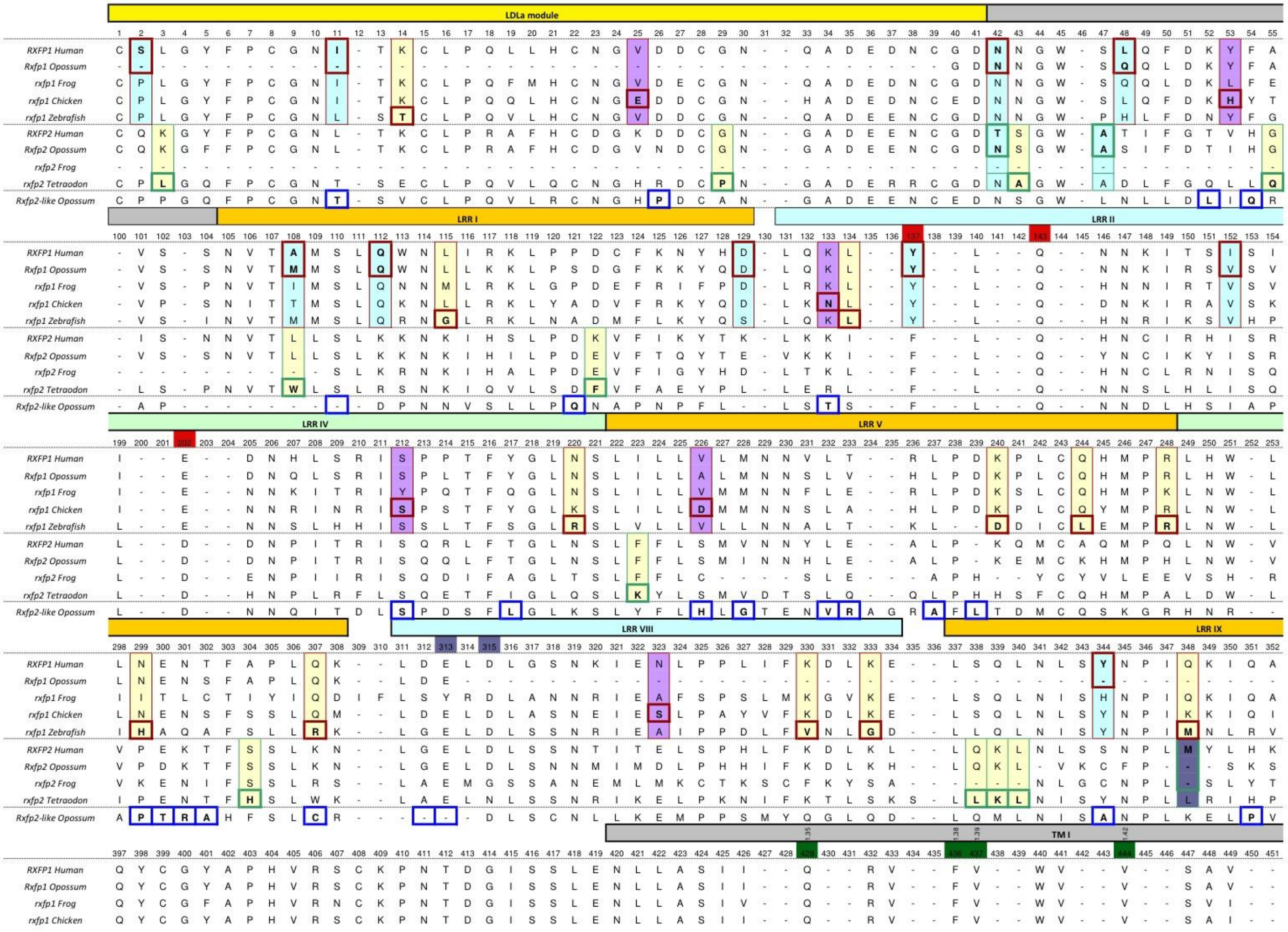
Figure 3.5. Histogram of the proportion of amino acids per region of the RXFP1/2 and RXFP3/4 receptors that showed evidence of positive selection for the branch-site test of positive selection comparing teleost and tetrapod genes.
Only those regions shared between RXFP1/2 (A) and RXFP3/4-type (B) receptors are included (e.g. the LRR region unique to RXFP1/2's was excluded). The difference between the proportions of amino acids selected per region is shown in panel C. The specific amino acids estimated to be under selection are given in Figure 3.6.

To compare the number of amino acids selected in RXFP1/2 versus RXFP3/4 type genes, I plotted the number of amino acids subject to positive selection among vertebrate lineages for only those domains common to the two receptor types (TM, ICL and ECL-Extracellular loop). I additionally subtracted the number of positively selected sites per domain in RXFP3/4 (which generally exhibited more sites under selection) from those in RXFP1/2 (Figure 3.5). In summary Figure 3.5 indicates that:

- ICL3 is the domain subject to the most positive selection among lineages for both receptor types (Figure 3.5A and 3.5B);
- ICL1 has a high proportion of selected sites for RXFP3/4 type receptors (Figure 3.5B), which is particularly evident when the receptor types are compared (Figure 3.5C).
- The only domain for which RXFP1/2 type genes appear to have more selection than RXFP3/4 type genes is ECL2, in which ~10% more amino acids are under selection for RXFP1/2 type genes (Figure 3.5C).

Of the 883 amino acids in the RXFP1/2 type molecules, 156 sites were found to have evidence of positive selection in one or more lineages of the vertebrate tree (Figure 3.6). Interestingly, of these selected amino acids, 29 were found to be selected in more than one receptor (i.e. RXFP1 and 2 or RXFP 1 and 2-like) within distinct lineages. Additionally, frequently selected amino acid positions were found adjacent to each other collectively suggesting that there are definite cold and hot spots of selection (Figure 3.5) In total, 41, 43 and 43 amino acids were found to be positively selected within either RXFP1, RXFP2 or RXFP2-like respectively (Figure 3.6). This indicates that similar levels of selection have taken place in the three RXFP1/2 type genes, and that there are common sites/areas of the proteins that are the targets of selection. Comparing the number of positively selected sites among lineages, slightly more positively selected sites were observed in teleosts for both RXFP1 (30) and RXFP2 (27) compared to mammals (22 in each).

For the RXFP3/4 genes, slightly more differences were observed among genes and lineages. Of the 297 amino acids in the RXFP3/4 alignment, 19 showed evidence of positive selection for RXFP3, 41 for RXFP4 and 26 for RXFP3-3 genes (Figure 3.6). Of these sites, mammalian RXFP3 showed evidence of positive selection at 12 amino acids compared to only 4 selected sites in teleosts (and 3 in chicken) while 23 amino acids showed evidence of selection in teleosts for RXFP4 but only 18 sites in mammalian RXFP4. Thus, overall there has been more selection in RXFP4 than RXFP3, although mammalian RXFP3 showed some evidence of selection, with five of the changes occurring in ICL3, an important domain as shown above.


LDLa module
LRR I
LRR II
LRR IV
LRR V
LRR VIII
LRR IX
TM I
1.35
1.38
1.39
1.42

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55
RXFP1 Human	C	S	L	G	Y	F	P	C	G	N	I	-	T	K	C	L	P	Q	L	L	H	C	N	G	V	D	D	C	G	N	-	-	Q	A	D	E	D	N	C	G	D	N	N	G	W	-	S	L	Q	F	D	K	Y	F	A
Rxfp1 Opossum	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	D	N	N	G	W	-	S	Q	Q	L	D	K	Y	F	A
rxfp1 Frog	C	P	L	G	Y	F	P	C	G	N	I	-	T	K	C	L	P	Q	F	M	H	C	N	G	V	D	E	C	G	N	-	-	Q	A	D	E	D	N	C	G	D	N	N	G	W	-	S	Q	Q	L	D	K	L	F	E
rxfp1 Chicken	C	P	L	G	Y	F	P	C	G	N	I	-	T	K	C	L	P	Q	Q	L	H	C	N	G	E	D	D	C	G	N	-	-	H	A	D	E	D	N	C	E	D	N	N	G	W	-	S	L	Q	F	D	K	H	Y	T
rxfp1 Zebrafish	C	P	L	G	Y	F	P	C	G	N	L	-	S	T	C	L	P	Q	V	L	H	C	N	G	V	D	D	C	G	N	-	-	Q	A	D	E	E	N	C	G	D	N	N	G	W	-	P	H	L	F	D	N	Y	F	G
RXFP2 Human	C	Q	K	G	Y	F	P	C	G	N	L	-	T	K	C	L	P	R	A	F	H	C	D	G	K	D	D	C	G	N	-	-	G	A	D	E	E	N	C	G	D	T	S	G	W	-	A	T	I	F	G	T	V	H	G
Rxfp2 Opossum	C	Q	K	G	F	F	P	C	G	N	L	-	T	K	C	L	P	R	A	F	H	C	D	G	V	N	D	C	G	N	-	-	G	A	D	E	E	N	C	G	D	N	S	G	W	-	A	S	I	F	D	T	I	H	G
rxfp2 Frog	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rxfp2 Tetraodon	C	P	L	G	Q	F	P	C	G	N	T	-	S	E	C	L	P	Q	V	L	Q	C	N	G	H	R	D	C	P	N	-	-	G	A	D	E	R	R	C	G	D	N	A	G	W	-	A	D	L	F	G	Q	L	L	Q
Rxfp2-like Opossum	C	P	P	G	Q	F	P	C	G	N	T	-	S	V	C	L	P	Q	V	L	R	C	N	G	H	P	D	C	A	N	-	-	G	A	D	E	E	N	C	E	D	N	S	G	W	-	L	N	L	L	D	L	I	Q	R

	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154
RXFP1 Human	-	V	S	-	S	N	V	T	A	M	S	L	Q	W	N	L	I	R	K	L	P	P	D	C	F	K	N	Y	H	D	-	L	Q	K	L	-	-	Y	-	-	L	-	-	Q	-	-	N	N	K	I	T	S	I	S	I
Rxfp1 Opossum	-	V	S	-	S	N	V	T	M	M	S	L	Q	W	N	L	L	K	K	L	P	S	D	G	F	K	K	Y	Q	D	-	L	Q	K	L	-	-	Y	-	-	L	-	-	Q	-	-	N	N	K	I	R	S	V	S	V
rxfp1 Frog	-	V	S	-	P	N	V	T	I	M	S	L	Q	N	N	M	L	R	K	L	G	P	D	E	F	R	I	F	P	D	-	L	R	K	L	-	-	Y	-	-	L	-	-	Q	-	-	H	N	N	I	R	T	V	S	V
rxfp1 Chicken	-	V	P	-	S	N	I	T	T	M	S	L	Q	K	N	L	L	R	K	L	Y	A	D	V	F	R	K	Y	Q	D	-	L	K	N	L	-	-	Y	-	-	L	-	-	Q	-	-	D	N	K	I	R	A	V	S	K
rxfp1 Zebrafish	-	V	S	-	I	N	V	T	M	M	S	L	Q	R	N	G	L	R	K	L	N	A	D	M	F	L	K	Y	Q	S	-	L	Q	K	L	-	-	Y	-	-	L	-	-	Q	-	-	H	N	R	I	K	S	V	H	P
RXFP2 Human	-	I	S	-	N	N	V	T	L	L	S	L	K	K	N	K	I	H	S	L	P	D	K	V	F	I	K	Y	T	K	-	L	K	K	I	-	-	F	-	-	L	-	-	Q	-	-	H	N	C	I	R	H	I	S	R
Rxfp2 Opossum	-	V	S	-	S	N	V	T	L	L	S	L	K	K	N	K	I	H	I	L	P	D	E	V	F	T	Q	Y	T	E	-	V	K	K	I	-	-	F	-	-	L	-	-	Q	-	-	Y	N	C	I	K	Y	I	S	R
rxfp2 Frog	-	-	-	-	-	-	-	-	-	-	S	L	K	R	N	K	I	H	A	L	P	D	E	V	F	I	G	Y	H	D	-	L	T	K	L	-	-	F	-	-	L	-	-	Q	-	-	H	N	C	L	R	N	I	S	Q
rxfp2 Tetraodon	-	L	S	-	P	N	V	T	W	L	S	L	R	S	N	K	I	Q	V	L	S	D	F	V	F	A	E	Y	P	L	-	L	E	R	L	-	-	F	-	-	L	-	-	Q	-	-	N	N	S	L	H	L	I	S	Q
Rxfp2-like Opossum	-	A	P	-	-	-	-	-	-	-	-	-	D	P	N	N	V	S	L	L	P	Q	N	A	P	N	P	F	L	-	-	L	S	T	S	-	-	F	-	-	L	-	-	Q	-	-	N	N	D	L	H	S	I	A	P

	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253
RXFP1 Human	I	-	-	E	-	-	D	N	H	L	S	R	I	S	P	P	T	F	Y	G	L	N	S	L	I	L	L	V	L	M	N	N	V	L	T	-	-	R	L	P	D	K	P	L	C	Q	H	M	P	R	L	H	W	-	L
Rxfp1 Opossum	I	-	-	E	-	-	D	N	Q	L	S	R	I	S	P	L	T	F	Y	G	L	N	S	L	I	L	L	A	L	M	N	N	S	L	V	-	-	H	L	P	D	K	P	L	C	Q	H	M	P	R	L	H	W	-	L
rxfp1 Frog	I	-	-	E	-	-	N	N	K	I	T	R	I	Y	P	Q	T	F	Q	G	L	N	S	L	I	L	L	V	M	M	N	N	F	L	E	-	-	R	L	P	D	K	S	L	C	Q	H	M	P	K	L	N	W	-	L
rxfp1 Chicken	I	-	-	E	-	-	N	N	R	I	N	R	I	S	P	S	T	F	Y	G	L	K	S	L	I	L	L	D	M	M	N	N	S	L	A	-	-	H	L	P	D	K	P	L	C	Q	Y	M	P	R	L	N	W	-	L
rxfp1 Zebrafish	L	-	-	E	-	-	N	N	S	L	H	H	I	S	S	L	T	F	S	G	L	R	S	L	V	L	L	V	L	L	N	N	A	L	T	-	-	K	L	-	-	D	D	I	C	L	E	M	P	R	L	N	W	-	L
RXFP2 Human	L	-	-	D	-	-	D	N	P	I	T	R	I	S	Q	R	L	F	T	G	L	N	S	L	F	F	L	S	M	V	N	N	Y	L	E	-	-	A	L	P	-	K	Q	M	C	A	Q	M	P	Q	L	N	W	-	V
Rxfp2 Opossum	L	-	-	D	-	-	D	N	P	I	T	R	I	S	Q	Q	L	F	T	G	L	N	S	L	F	F	L	S	M	I	N	N	H	L	E	-	-	A	L	P	-	K	E	M	C	K	H	M	P	H	L	N	W	-	V
rxfp2 Frog	L	-	-	D	-	-	E	N	P	I	I	R	I	S	Q	D	I	F	A	G	L	T	S	L	F	F	L	C	-	-	-	-	S	L	E	-	-	-	A	P	H	-	-	Y	C	Y	V	L	E	E	V	S	H	-	R
rxfp2 Tetraodon	L	-	-	D	-	-	H	N	P	L	R	F	L	S	Q	E	T	F	I	G	L	Q	S	L	K	Y	L	S	M	V	D	T	S	L	Q	-	-	Q	L	P	H	H	S	F	C	Q	H	M	P	A	L	D	W	-	L
Rxfp2-like Opossum	L	-	-	D	-	-	N	N	Q	I	T	D	L	S	P	D	S	F	L	G	L	K	S	L	Y	F	L	H	L	G	T	E	N	V	R	A	G	R	A	F	L	T	D	M	C	Q	S	K	G	R	H	N	R	-	-

	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352
RXFP1 Human	L	N	E	N	T	F	A	P	L	Q	K	-	-	L	D	E	L	D	L	G	S	N	K	I	E	N	L	P	P	L	I	F	K	D	L	K	E	-	-	L	S	Q	L	N	L	S	Y	N	P	I	Q	K	I	Q	A
Rxfp1 Opossum	L	N	E	N	S	F	A	P	L	Q	K	-	-	L	D	E	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rxfp1 Frog	I	I	T	L	C	T	I	Y	I	Q	D	I	F	L	S	Y	R	D	L	A	N	N	R	I	E	A	F	S	P	S	L	M	K	G	V	K	E	-	-	L	S	Q	L	N	I	S	H	N	P	I	Q	K	I	Q	A
rxfp1 Chicken	L	N	E	N	S	F	S	S	L	Q	M	-	-	L	D	E	L	D	L	A	S	N	E	I	E	S	L	P	A	Y	V	F	K	D	L	K	E	-	-	L	S	Q	L	N	L	S	Y	N	P	I	K	K	I	Q	I
rxfp1 Zebrafish	I	H	A	Q	A	F	S	L	L	R	K	-	-	L	G	E	L	D	L	S	S	N	R	I	E	A	I	P	P	D	L	F	V	N	L	G	D	-	-	L	L	Q	L	N	I	S	Y	N	P	I	M	N	L	R	V
RXFP2 Human	V	P	E	K	T	F	S	S	L	K	N	-	-	L	G	E	L	D	L	S	S	N	T	I	T	E	L	S	P	H	L	F	K	D	L	K	L	-	-	L	Q	K	L	N	L	S	S	N	P	L	M	Y	L	H	K
Rxfp2 Opossum	V	P	D	K	T	F	S	S	L	K	N	-	-	L	G	E	L	D	L	S	N	N	M	I	M	D	L	P	H	H	I	F	K	D	L	K	H	-	-	L	Q	K	L	-	V	K	C	F	P	-	-	-	S	K	S
rxfp2 Frog	V	K	E	N	I	F	S	S	L	R	S	-	-	L	A	E	M	G	S	S	A	N	E	M	L	M	K	C	T	K	S	C	F	K	Y	S	A	-	-	-	-	-	-	N	L	G	C	N	P	L	-	S	L	Y	T
rxfp2 Tetraodon	I	P	E	N	T	F	H	S	L	W	K	-	-	L	A	E	L	N	L	S	S	N	R	I	K	E	L	P	K	N	I	F	K	T	L	S	K	S	-	L	L	K	L	N	I	S	Y	N	P	L	L	R	I	H	P
Rxfp2-like Opossum	A	P	T	R	A	H	F	S	L	C	R	-	-	-	-	-	-	D	L	S	C	N	L	L	K	E	M	P	P	S	M	Y	Q	G	L	Q	D	-	-	L	Q	M	L	N	I	S	A	N	P	L	K	E	L	P	V

	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451
RXFP1 Human	Q	Y	C	G	Y	A	P	H	V	R	S	C	K	P	N	T	D	G	I	S	S	L	E	N	L	L	A	S	I	I	-	-	Q	-	-	R	V	-	-	F	V	-	-	W	V	-	-	V	-	-	S	A	V	-	-
Rxfp1 Opossum	Q	Y	C	G	Y	A	P	H	V	R	S	C	K	P	N	T	D	G	I	S	S	L	E	N	L	L	A	S	I	I	-	-	Q	-	-	R	V	-	-	F	V	-	-	W	V	-	-	V	-	-	S	A	V	-	-
rxfp1 Frog	Q	Y	C	G	F	A	P	H	V	R	N	C	K	P	N	T	D	G	I	S	S	L	E	N	L	L	A	S	I	V	-	-	Q	-	-	R	V	-	-	F	V	-	-	W	V	-	-	V	-	-	S	V	I	-	-
rxfp1 Chicken	Q	Y	C	G	Y	A	P	H	V	R	S	C	K	P	N	T	D	G	I	S	S	L	E	N	L	L	A	S	I	I	-	-	Q	-	-	R	V	-	-	F	V	-	-	W	V	-	-	V	-	-	S	A	I	-	-

rxfp1 Zebrafish	Q	Y	C	G	Y	A	P	Y	V	R	S	C	K	P	N	T	D	G	I	S	S	F	E	D	L	L	A	N	I	V	-	-	L	-	-	R	V	-	-	F	V	-	-	W	A	-	-	V	-	-	S	A	T	-	-
RXFP2 Human	R	Y	C	S	Y	A	P	H	V	R	I	C	M	P	L	T	D	G	I	S	S	F	E	D	L	L	A	N	N	I	-	-	L	-	-	R	I	-	-	F	V	-	-	W	V	-	-	I	-	-	A	F	I	-	-
Rxfp2 Opossum	R	Y	C	S	Y	A	P	H	V	R	I	C	I	P	L	T	D	G	I	S	S	F	E	D	L	L	A	N	N	I	-	-	L	-	-	R	I	-	-	F	V	-	-	W	V	-	-	I	-	-	A	F	I	-	-
rxfp2 Frog	R	Y	C	S	Y	A	P	H	V	R	V	C	T	P	L	T	D	G	I	S	S	F	E	N	L	L	A	N	T	I	-	-	L	-	-	R	V	-	-	F	V	-	-	W	V	-	-	I	-	-	A	C	I	-	-
rxfp2 Tetraodon	Q	Y	C	S	Y	A	P	H	V	R	S	C	K	P	N	T	D	G	I	S	S	F	E	D	L	L	A	N	M	V	-	-	L	-	-	R	V	-	-	S	V	-	-	W	V	-	-	I	-	-	A	F	I	-	-
Rxfp2-like Opossum	Q	Y	C	S	Y	V	P	H	V	R	N	C	Q	P	N	T	D	G	I	S	S	L	E	N	L	L	A	N	L	I	-	-	L	-	-	R	I	-	-	F	V	-	-	W	V	-	-	I	-	-	A	C	T	-	-

	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550
																															ECL I																								
	2.53	2.54							2.57	2.58						2.60	2.61	2.62						2.64	2.65																							3.25	3.26						3.28
RXFP1 Human	M	G	-	-	I	Y	-	-	L	F	-	-	V	-	-	I	G	G	-	-	F	-	-	D	L	-	-	K	F	R	G	E	Y	N	K	H	A	Q	L	W	M	E	S	T	H	-	-	C	Q	-	-	L	-	-	V
Rxfp1 Opossum	M	G	-	-	I	Y	-	-	L	F	-	-	V	-	-	I	G	A	-	-	F	-	-	D	L	-	-	K	Y	R	G	E	Y	N	K	H	A	Q	L	W	M	D	S	T	Y	-	-	C	Q	-	-	L	-	-	V
rxfp1 Frog	M	G	-	-	V	Y	-	-	L	F	-	-	V	-	-	I	G	Y	-	-	F	-	-	D	L	-	-	K	Y	R	G	E	Y	N	E	N	A	Q	A	W	M	D	S	T	Q	-	-	C	R	-	-	L	-	-	V
rxfp1 Chicken	M	G	-	-	I	Y	-	-	L	F	-	-	V	-	-	I	G	A	-	-	F	-	-	D	L	-	-	K	Y	R	G	E	Y	N	K	H	A	Q	L	W	M	D	S	I	H	-	-	C	Q	-	-	L	-	-	V
rxfp1 Zebrafish	M	G	-	-	V	Y	-	-	L	F	-	-	M	-	-	I	G	A	-	-	Y	-	-	D	L	-	-	K	F	R	G	E	Y	N	R	H	A	Q	A	G	M	D	S	E	A	-	-	C	Q	-	-	V	-	-	I
RXFP2 Human	M	G	-	-	V	Y	-	-	L	F	-	-	F	-	-	V	G	I	-	-	F	-	-	D	I	-	-	K	Y	R	G	Q	Y	Q	K	Y	A	L	L	W	M	E	S	V	Q	-	-	C	R	-	-	L	-	-	M
Rxfp2 Opossum	M	G	-	-	V	Y	-	-	L	F	-	-	F	-	-	V	G	F	-	-	F	-	-	D	I	-	-	K	Y	R	G	Q	Y	Q	K	Y	A	L	L	W	M	E	S	L	Q	-	-	C	R	-	-	L	-	-	M
rxfp2 Frog	M	G	-	-	I	Y	-	-	L	F	-	-	F	-	-	I	G	V	-	-	F	-	-	D	V	-	-	K	Y	R	G	Q	Y	K	K	Y	A	L	L	W	M	E	S	L	Q	-	-	C	R	-	-	S	-	-	L
rxfp2 Tetraodon	M	G	-	-	V	Y	-	-	L	F	-	-	F	-	-	V	G	V	-	-	F	-	-	D	V	-	-	K	F	R	G	E	Y	N	R	H	A	L	L	W	M	E	S	V	E	-	-	C	R	-	-	T	-	-	I
Rxfp2-like Opossum	M	G	-	-	V	Y	-	-	L	F	-	-	F	-	-	I	G	A	-	-	S	-	-	D	L	-	-	R	Y	A	G	E	Y	N	K	H	A	Q	A	W	M	A	S	P	Q	-	-	C	Q	-	-	A	-	-	G

	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649
	ICL2										TM IV																																												
																									4.50							4.53							4.56	4.57							4.60	4.61							
RXFP1 Human	V	Y	P	F	R	C	V	R	P	-	G	K	C	R	T	I	T	V	L	I	L	I	-	-	W	-	-	I	T	-	-	G	-	-	F	I	-	-	V	A	-	-	F	I	-	-	P	L	-	-	S	N	K	-	-
Rxfp1 Opossum	V	Y	P	F	R	C	L	K	P	-	G	K	C	R	T	I	T	I	L	I	L	I	-	-	W	-	-	I	I	-	-	G	-	-	F	V	-	-	V	A	-	-	F	I	-	-	P	L	-	-	S	N	K	-	-
rxfp1 Frog	V	Y	P	F	R	C	L	K	P	-	G	K	C	R	T	I	T	T	L	I	L	I	-	-	W	-	-	I	I	-	-	G	-	-	F	V	-	-	I	A	-	-	F	I	-	-	P	L	-	-	S	N	Q	-	-
rxfp1 Chicken	V	Y	P	F	R	C	L	K	P	-	R	K	C	R	T	I	S	I	L	V	L	I	-	-	W	-	-	V	I	-	-	G	-	-	F	A	-	-	V	A	-	-	F	I	-	-	P	L	-	-	S	N	K	-	-
rxfp1 Zebrafish	V	Y	P	F	R	Y	L	T	L	-	G	R	R	R	T	V	T	I	L	V	V	I	-	-	W	-	-	V	L	-	-	G	-	-	F	I	-	-	I	A	-	-	F	L	-	-	P	L	-	-	L	F	K	-	-
RXFP2 Human	V	F	P	F	S	N	I	R	P	-	G	K	R	Q	T	S	V	I	L	I	C	I	-	-	W	-	-	M	A	-	-	G	-	-	F	L	-	-	I	A	-	-	V	I	-	-	P	F	-	-	W	N	K	-	-
Rxfp2 Opossum	V	F	P	F	S	N	I	R	P	-	G	K	H	Q	T	L	I	I	L	V	C	I	-	-	W	-	-	I	A	-	-	G	-	-	F	L	-	-	I	A	-	-	I	I	-	-	P	F	-	-	W	S	E	-	-
rxfp2 Frog	V	F	P	F	S	N	I	R	P	-	G	K	R	Q	T	L	I	I	L	I	S	L	-	-	W	-	-	A	V	-	-	G	-	-	F	I	-	-	I	A	-	-	I	V	-	-	P	F	-	-	W	N	E	-	-
rxfp2 Tetraodon	V	F	P	F	S	N	L	R	P	-	G	K	L	L	T	G	V	V	L	A	S	I	-	-	W	-	-	L	L	-	-	G	-	-	V	I	-	-	I	A	-	-	A	V	-	-	P	L	-	-	M	N	E	-	-
Rxfp2-like Opossum	V	F	P	F	S	H	Y	R	A	-	G	R	R	Q	T	L	A	T	L	V	G	I	-	-	W	-	-	L	V	-	-	G	-	-	F	T	-	-	I	A	-	-	V	V	-	-	P	F	-	-	W	S	R	-	-

	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748
																											ICL3																												
					5.46	5.47							5.50	5.51																																									
RXFP1 Human	L	G	-	-	I	N	-	-	L	A	-	-	A	F	-	-	I	I	I	V	F	S	Y	G	S	M	F	Y	S	V	H	Q	S	A	I	T	A	T	E	I	R	N	Q	V	K	K	-	E	M	I	L	A	K	R	F
Rxfp1 Opossum	L	G	-	-	V	N	-	-	L	A	-	-	A	F	-	-	I	I	I	V	F	S	Y	G	S	M	F	Y	S	V	H	Q	T	A	V	T	A	T	E	I	R	N	H	V	K	K	-	E	V	T	L	A	K	R	F
rxfp1 Frog	L	G	-	-	V	N	-	-	L	A	-	-	A	F	-	-	I	I	I	V	F	S	Y	S	S	M	F	Y	S	I	H	R	T	A	I	M	A	T	E	I	H	N	H	I	K	K	-	E	M	T	L	A	K	R	F
rxfp1 Chicken	L	G	-	-	V	N	-	-	L	A	-	-	A	F	-	-	L	I	I	V	F	S	Y	G	S	M	F	Y	S	V	H	Q	T	A	I	M	A	T	E	I	Q	N	H	I	K	K	-	E	M	I	L	A	K	R	F
rxfp1 Zebrafish	L	G	-	-	L	N	-	-	L	V	-	-	A	F	-	-	L	I	I	V	L	S	Y	G	S	M	F	Y	N	I	Q	R	T	G	T	Q	T	T	K	Y	S	N	H	I	K	K	-	E	L	T	I	A	K	R	F
RXFP2 Human	L	G	-	-	V	N	-	-	L	L	-	-	A	F	-	-	L	I	I	V	F	S	Y	I	T	M	F	C	S	I	Q	K	T	A	L	Q	T	T	E	V	R	N	C	F	G	R	-	E	V	A	V	A	N	R	F
Rxfp2 Opossum	L	G	-	-	V	N	-	-	L	L	-	-	A	F	-	-	L	I	I	V	F	S	Y	I	S	M	F	C	S	I	Q	K	T	A	L	Q	T	S	D	M	R	I	P	I	R	R	-	D	V	A	L	A	N	R	F
rxfp2 Frog	L	G	-	-	V	N	-	-	L	L	-	-	A	F	-	-	I	I	I	V	F	S	Y	I	S	M	F	C	S	I	Q	K	T	A	L	R	T	S	E	V	N	S	H	I	H	T	-	D	V	A	V	A	N	R	F
rxfp2 Tetraodon	L	G	-	-	L	N	-	-	L	V	-	-	A	F	-	-	L	I	I	V	L	S	Y	S	S	M	F	Y	S	I	Y	K	T	G	I	N	A	T	D	L	R	S	R	L	H	R	-	D	V	A	V	A	N	R	F
Rxfp2-like Opossum	L	G	-	-	V	N	-	-	L	I	-	-	A	F	-	-	I	T	M	V	V	A	Y	S	S	M	F	H	A	V	R	V	S	R	A	R	S	A	Q	-	-	-	H	L	S	R	-	E	V	S	I	A	K	R	F

	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847
	ECL II				TM VII																																																		
															7.33	7.34	7.35	7.36						7.38	7.39	7.40							7.43						7.45	7.46							7.49							7.52	
RXFP1 Human	L	Q	V	E	I	-	-	P	-	-	G	T	-	-	I	T	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
Rxfp1 Opossum	L	Q	V	E	I	-	-	P	-	-	G	T	-	-	I	T	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
rxfp1 Frog	L	Q	V	E	I	-	-	P	-	-	G	S	-	-	I	S	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
rxfp1 Chicken	L	Q	V	E	I	-	-	P	-	-	G	T	-	-	I	T	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	A	-	N	S	-	-	A	L	-	-	N	-	-	P	L	-	-	L	-
rxfp1 Zebrafish	M	E	V	E	I	-	-	P	-	-	G	T	-	-	I	S	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
RXFP2 Human	F	R	V	E	I	-	-	P	-	-	D	T	-	-	M	T	S	W	-	-	I	-	-	V	I	F	-	-	F	-	-	L	P	-	-	V	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
Rxfp2 Opossum	F	Q	V	E	I	-	-	P	-	-	D	T	-	-	I	T	S	W	-	-	I	-	-	V	I	F	-	-	F	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
rxfp2 Frog	F	R	V	E	I	-	-	P	-	-	G	T	-	-	V	T	S	W	-	-	I	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
rxfp2 Tetraodon	L	E	V	E	I	-	-	P	-	-	G	T	-	-	I	S	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	I	-	-	L	-
Rxfp2-like Opossum	L	Q	L	E	I	-	-	P	-	-	G	A	-	-	V	T	S	W	-	-	V	-	-	V	I	F	-	-	I	-	-	L	P	-	-	I	-	-	N	S	-	-	A	L	-	-	N	-	-	P	V	-	-	L	-

	LRR flanking region																																											
	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
RXFP1 Human	S	Y	-	-	-	P	F	E	A	E	T	P	E	C	V	V	G	S	V	P	V	Q	C	L	C	-	Q	G	L	E	L	D	C	D	E	-	T	N	L	R	A	V	P	S
Rxfp1 Opossum	N	N	-	-	-	P	F	E	T	E	T	S	E	C	L	V	E	S	V	P	V	K	C	L	C	-	Q	G	L	E	L	D	C	D	E	-	A	N	L	R	A	V	P	S
rxfp1 Frog	K	H	-	-	-	D	L	E	M	K	P	S	E	C	T	L	G	P	V	P	T	Q	C	L	C	-	R	G	L	E	L	D	C	D	G	-	A	K	L	R	T	V	P	S
rxfp1 Chicken	L	D	-	-	-	A	F	Q	T	K	T	P	E	C	L	A	G	A	V	P	V	E	C	K	C	-	Q	G	L	E	V	F	C	D	A	-	A	K	L	R	D	V	P	L
rxfp1 Zebrafish	I	S	N	N	L	G	N	K	S	D	A	-	-	C	L	L	G	T	V	P	A	E	C	Q	C	-	R	D	L	E	L	D	C	D	G	-	A	H	F	K	D	V	P	M
RXFP2 Human	N	A	-	-	-	N	S	V	A	L	T	Q	E	C	F	L	K	Q	Y	P	Q	C	C	D	C	-	K	E	T	E	L	E	C	V	N	-	G	D	L	K	S	V	P	M
Rxfp2 Opossum	N	P	-	-	-	N	N	M	D	L	S	Q	E	C	Y	L	Q	Q	Y	P	Q	C	C	A	C	-	K	E	T	E	L	E	C	I	N	-	V	N	L	K	S	V	P	L
rxfp2 Frog	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rxfp2 Tetraodon	N	H	V	D	Q	D	P	L	N	-	-	-	D	C	-	L	Q	D	Y	P	E	S	C	G	C	-	V	Q	T	D	V	A	C	I	Q	-	V	D	L	Q	D	V	P	L
Rxfp2-like Opossum	T	K	-	-	-	K	T	R	E	Q	L	K	D	C	S	L	G	L	Y	P	K	G	C	K	C	-	S	W	R	V	V	D	C	T	A	-	Q	G	L	S	G	V	P	P

											LRR III																								LRR IV									
	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198
RXFP1 Human	Y	A	F	R	G	L	N	S	-	-	L	T	K	L	Y	L	S	H	N	R	I	T	F	L	K	P	G	V	F	E	D	L	H	R	L	E	-	W	-	L	-	-	I	-
Rxfp1 Opossum	Y	A	F	R	G	L	H	N	-	-	L	T	K	L	Y	L	S	H	N	K	I	T	V	L	K	P	G	V	F	E	D	L	H	R	L	E	-	W	-	L	-	-	I	-
rxfp1 Frog	H	A	F	K	G	L	Y	N	-	-	L	T	K	L	Y	L	S	H	N	E	I	T	T	L	K	P	G	V	F	E	D	L	H	R	L	E	-	W	-	L	-	-	I	-
rxfp1 Chicken	H	A	F	K	G	L	Y	N	-	-	L	T	K	L	Y	L	S	N	N	K	I	T	N	L	K	P	R	V	F	E	D	L	H	K	L	E	-	W	-	L	-	-	I	-
rxfp1 Zebrafish	Q	A	F	R	G	L	Y	N	-	-	L	T	R	L	Y	L	S	Y	N	R	I	T	T	L	L	P	D	V	F	Q	D	L	H	K	L	E	-	W	-	L	-	-	I	-
RXFP2 Human	K	A	F	F	G	L	C	N	-	-	L	Q	I	L	Y	L	N	H	N	C	I	T	T	L	R	P	G	I	F	K	D	L	H	Q	L	T	-	W	-	L	-	-	I	-
Rxfp2 Opossum	K	A	F	I	G	L	H	K	-	-	L	Q	M	L	Y	L	S	H	N	C	I	T	S	L	R	P	G	V	F	K	D	L	H	E	L	S	-	W	-	L	-	-	I	-
rxfp2 Frog	K	A	F	F	G	L	Y	H	-	-	L	Q	R	L	Y	L	S	N	N	C	I	S	Y	L	Q	Q	G	I	F	S	H	L	R	E	L	K	-	W	-	L	-	-	I	-
rxfp2 Tetraodon	H	A	F	S	G	L	R	I	-	-	L	K	R	L	F	L	S	E	N	L	I	S	S	L	S	P	G	V	F	K	D	L	H	Q	L	Q	-	W	-	L	W	-	L	-
Rxfp2-like Opossum	G	A	F	A	G	L	S	K	-	-	L	R	K	L	F	L	S	H	N	R	I	R	S	L	P	P	R	L	F	Q	D	L	F	Q	L	E	-	W	-	L	-	-	M	-

	LRR VI																													LRR VII														
	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297
RXFP1 Human	-	-	D	-	-	L	-	-	-	E	-	-	G	N	H	I	H	N	L	R	N	L	T	F	I	S	C	-	S	N	L	T	V	L	V	M	R	K	-	N	K	I	N	H
Rxfp1 Opossum	-	-	D	-	-	F	-	-	-	E	-	-	G	N	H	I	H	N	L	R	N	M	T	F	I	S	C	-	S	T	L	T	V	L	V	M	R	K	-	N	E	I	N	H
rxfp1 Frog	-	-	D	-	-	F	-	-	-	E	-	-	G	N	N	V	Q	S	L	T	N	T	T	F	I	S	C	-	A	T	L	T	V	L	M	-	R	K	E	N	E	Q	G	H
rxfp1 Chicken	-	-	D	-	-	L	-	-	-	E	-	-	G	N	H	I	H	H	L	R	N	V	T	F	I	S	C	-	S	T	L	T	V	L	V	M	R	Q	-	N	K	I	S	S
rxfp1 Zebrafish	-	-	D	-	-	I	-	-	-	E	-	-	G	N	K	M	E	T	V	G	N	V	T	F	R	S	C	-	N	M	L	T	V	L	V	L	Q	R	-	N	R	I	S	R
RXFP2 Human	-	-	D	-	-	L	-	-	-	E	-	-	G	N	R	I	K	Y	L	T	N	S	T	F	L	S	C	-	D	S	L	T	V	L	F	L	P	R	-	N	Q	I	G	F
Rxfp2 Opossum	-	-	D	-	-	L	-	-	-	E	-	-	G	N	Q	I	K	S	L	M	N	S	T	F	L	A	C	-	D	E	L	T	V	L	F	L	P	R	-	N	Q	I	D	F
rxfp2 Frog	-	-	D	-	-	F	-	-	-	E	-	-	G	N	R	I	K	T	L	E	S	S	S	F	V	T	C	-	N	E	L	T	V	L	F	L	R	G	-	N	Q	I	S	L
rxfp2 Tetraodon	-	-	D	-	-	L	-	-	-	E	-	-	G	N	Q	I	Q	T	L	N	Y	S	I	L	K	T	C	-	S	K	L	E	V	L	L	L	M	N	-	N	R	I	Q	R
Rxfp2-like Opossum	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	P	C	-	Q	S	G	S	Q	L	T	F	P	T	-	P	K	V	G	L

									LRR X																																			
	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396
RXFP1 Human	N	Q	F	D	Y	L	V	K	L	K	S	L	-	S	L	E	G	I	E	I	S	N	-	-	I	Q	Q	R	M	F	R	P	L	M	N	L	S	H	I	Y	F	K	K	F
Rxfp1 Opossum	-	-	-	-	-	-	-	-	-	-	-	F	-	S	L	E	G	I	E	I	S	N	-	-	I	Q	R	R	M	F	A	P	L	R	N	L	S	H	I	Y	F	K	K	F
rxfp1 Frog	D	Q	F	D	Y	V	I	K	L	K	S	L	-	S	L	E	G	I	E	I	P	N	-	-	I	Q	R	R	M	F	M	P	L	K	N	L	T	H	I	Y	F	K	K	F
rxfp1 Chicken	D	Q	F	D	F	L	T	K	L	K	S	L	-	S	L	E	G	I	E	I	A	N	-	-	I	Q	R	R	M	F	I	P	L	R	N	L	S	H	I	Y	F	K	K	F
rxfp1 Zebrafish	D	H	F	D	K	L	H	K	L	K	S	L	-	S	I	E	G	I	E	I	G	N	-	-	I	H	R	R	M	F	E	P	L	K	N	L	T	H	I	Y	F	K	K	F
RXFP2 Human	N	Q	F	E	S	L	K	Q	L	Q	S	L	-	D	L	E	R	I	E	I	P	N	-	-	I	N	T	R	M	F	Q	P	M	K	N	L	S	H	I	Y	F	K	N	F
Rxfp2 Opossum	H	S	F	T	F	M	L	P	I	F	S	R	-	D	L	E	K	I	E	I	P	N	-	-	I	N	T	R	M	F	Q	H	M	R	N	L	S	Y	I	Y	F	K	N	F
rxfp2 Frog	E	Q	F	E	S	L	Q	H	L	Q	S	L	-	D	L	E	K	I	E	I	P	N	-	-	I	S	T	R	M	F	Q	P	M	K	N	L	S	H	I	Y	F	K	K	F
rxfp2 Tetraodon	S	H	F	T	H	L	I	H	L	Q	S	L	W	A	L	E	G	I	E	I	P	E	-	-	I	E	T	K	M	F	L	P	M	K	N	L	S	H	I	Y	F	K	T	F
Rxfp2-like Opossum	D	H	F	D	S	L	P	Y	L	Q	S	L	-	S	M	E	G	M	E	I	S	N	-	-	I	E	N	R	M	F	Q	K	L	T	N	L	S	Y	V	Y	F	G	H	F

	1.46																ICL I						TM II															2.50						2.52
	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495
RXFP1 Human	T	-	-	C	F	G	N	I	F	V	I	C	M	R	P	Y	I	R	S	E	N	K	L	-	Y	A	M	S	I	I	S	L	C	C	-	-	A	D	-	-	C	-	-	L
Rxfp1 Opossum	T	-	-	C	F	G	N	I	F	V	I	C	M	R	P	Y	I	R	S	E	N	K	L	-	H	A	L	S	I	I	S	L	C	C	-	-	A	D	-	-	C	-	-	L
rxfp1 Frog	T	-	-	C	F	G	N	I	F	V	I	C	T	R	P	Y	I	R	S	E	N	K	L	-	H	A	M	S	I	I	S	L	C	C	-	-	A	D	-	-	C	-	-	L
rxfp1 Chicken	T	-	-	C	F	G	N	I	F	V	I	C	M	R	P	Y	I	R	S	E	N	K	L	-	H	A	I	S	I	M	S	L	C	C	-	-	A	D	-	-	C	-	-	L

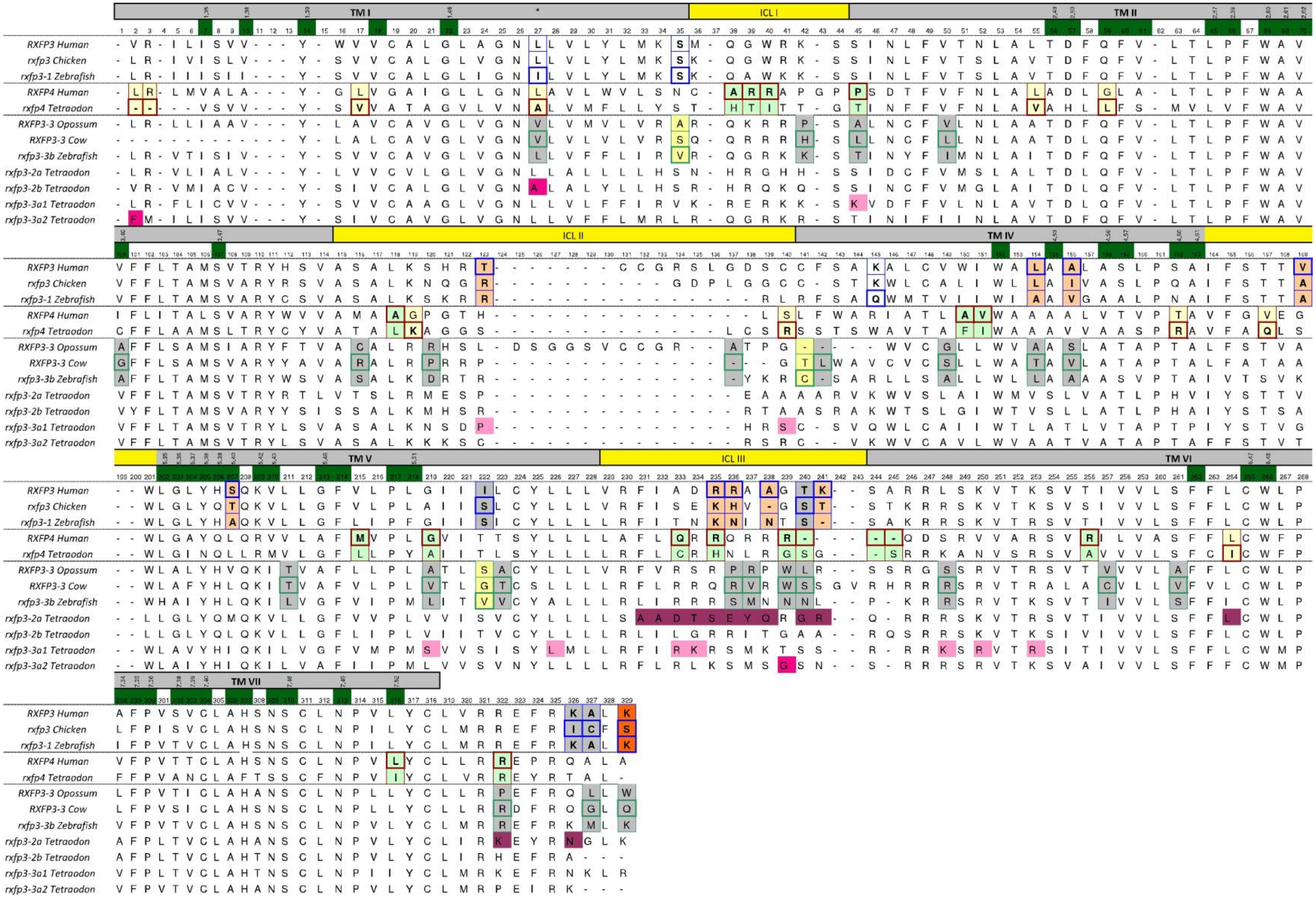
rxfp1 Zebrafish	T	-	-	C	F	G	N	I	F	V	I	C	M	R	S	Y	I	R	S	E	N	K	L	-	H	A	M	C	I	I	S	L	C	C	-	-	A	D	-	-	G	-	-	L
RXFP2 Human	T	-	-	C	F	G	N	L	F	V	I	G	M	R	S	F	I	K	A	E	N	T	T	-	H	A	M	S	I	K	I	L	C	C	-	-	A	D	-	-	C	-	-	L
Rxfp2 Opossum	T	-	-	C	F	G	N	L	F	V	I	C	M	R	T	F	I	R	A	E	N	K	T	-	H	A	M	S	I	K	I	L	C	C	-	-	A	D	-	-	C	-	-	L
rxfp2 Frog	T	-	-	C	F	G	N	I	F	V	I	G	M	R	S	C	I	Q	S	E	N	K	T	-	H	T	M	S	I	K	V	L	C	C	-	-	A	D	-	-	C	-	-	L
rxfp2 Tetraodon	T	-	-	C	F	G	N	L	L	V	I	F	M	R	S	L	I	R	A	E	N	N	**L**	-	H	A	V	C	I	K	V	L	C	C	-	-	A	D	-	-	C	-	-	L
Rxfp2-like Opossum	T	-	-	C	L	G	N	L	L	V	V	C	M	R	S	L	I	V	P	E	N	**F**	**Q**	-	H	A	M	A	I	K	C	L	C	C	-	-	A	D	-	-	F	-	-	L

	3.29							3.32	3.33							TM III							3.39	3.40											3.47									
	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594
RXFP1 Human	G	-	-	S	L	-	-	A	I	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	F	L	-	-	T	-	-	L	E	K	Y	I	C	I
Rxfp1 Opossum	G	-	-	S	L	-	-	A	I	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	F	L	-	-	T	-	-	L	E	K	Y	I	C	I
rxfp1 Frog	G	-	-	S	L	-	-	A	I	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	L	-	-	T	-	-	L	E	K	Y	I	C	I
rxfp1 Chicken	G	-	-	S	L	-	-	A	I	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	L	-	-	T	-	-	L	E	K	Y	I	C	I
rxfp1 Zebrafish	G	-	-	S	L	-	-	A	M	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	L	-	-	T	-	-	L	E	K	Y	I	C	I
RXFP2 Human	G	-	-	F	L	-	-	A	M	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	L	-	-	T	-	-	L	E	K	F	L	V	I
Rxfp2 Opossum	G	-	-	F	L	-	-	A	M	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	L	-	-	T	-	-	L	E	K	Y	V	V	I
rxfp2 Frog	G	-	-	F	L	-	-	A	M	-	-	L	-	-	S	T	E	-	-	V	-	-	S	V	-	-	L	L	L	T	F	L	-	-	T	-	-	L	E	K	Y	L	A	I
rxfp2 Tetraodon	G	-	-	F	L	-	-	A	M	-	-	L	-	-	S	S	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	L	-	-	T	-	-	L	E	K	F	L	V	I
Rxfp2-like Opossum	G	-	-	**C**	L	-	-	A	M	-	-	L	-	-	S	S	E	-	-	V	-	-	S	V	-	-	L	L	L	T	Y	M	-	-	T	-	-	L	E	K	Y	V	G	I

														ECL2																5.35	5.36	5.37	5.38	5.39			TM V				5.42	5.43		
	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693
RXFP1 Human	E	F	F	K	N	-	Y	Y	G	T	-	N	G	V	C	F	P	L	H	S	E	D	T	E	S	I	G	-	-	A	Q	I	Y	S	V	-	-	A	-	-	I	F	-	-
Rxfp1 Opossum	E	F	F	K	N	-	Y	Y	G	T	-	N	G	V	C	F	P	L	H	S	E	Q	G	E	S	M	G	-	-	A	Q	I	Y	S	V	-	-	V	-	-	I	F	-	-
rxfp1 Frog	T	F	F	H	N	-	Y	Y	G	T	-	N	G	V	C	F	P	L	H	S	E	Q	P	E	S	T	A	-	-	A	Q	I	Y	S	V	-	-	I	-	-	I	F	-	-
rxfp1 Chicken	E	F	F	R	N	-	Y	Y	G	T	-	N	G	V	C	F	P	L	H	S	E	Q	S	E	S	S	G	-	-	**S**	Q	I	Y	S	V	-	-	V	-	-	I	F	-	-
rxfp1 Zebrafish	G	**V**	F	R	N	-	F	Y	G	T	-	N	G	V	C	F	P	L	H	S	E	Q	P	E	T	L	G	-	-	A	Q	I	Y	S	I	-	-	V	-	-	I	F	-	-
RXFP2 Human	D	Y	F	G	N	-	F	Y	G	K	-	N	0	V	C	F	P	L	Y	Y	D	Q	T	E	D	I	G	-	-	S	K	**G**	Y	S	L	-	-	G	-	-	I	F	-	-
Rxfp2 Opossum	E	F	F	G	N	-	F	Y	G	K	-	N	G	V	C	F	P	L	Y	-	D	Q	T	E	D	F	R	-	-	S	K	**G**	Y	S	L	-	-	G	-	-	I	F	-	-
rxfp2 Frog	D	F	F	G	N	-	Y	Y	G	K	-	N	G	V	C	F	P	L	Y	P	D	Q	T	E	E	A	G	-	-	G	Q	G	Y	S	L	-	-	G	-	-	V	F	-	-
rxfp2 Tetraodon	G	**V**	F	G	N	-	Y	Y	G	**H**	-	N	G	V	C	F	P	L	H	S	D	R	**Q**	E	K	P	T	-	-	A	K	G	Y	S	**T**	-	-	G	-	-	V	F	-	-
Rxfp2-like Opossum	G	P	**F**	G	N	-	Y	Y	G	T	-	N	G	V	C	F	P	L	H	F	D	**T**	S	E	**S**	A	**M**	-	-	A	Q	H	Y	S	T	-	-	**A**	-	-	I	F	-	-

									6.44							TM VI									6.51	6.52						6.54	6.55								6.58	6.59		
	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792
RXFP1 Human	F	F	I	V	F	T	-	-	D	-	-	A	L	-	-	C	-	W	-	-	I	P	-	-	I	F	-	-	V	-	-	V	K	-	-	F	-	L	-	-	S	L	-	-
Rxfp1 Opossum	F	F	I	V	F	T	-	-	D	-	-	A	L	-	-	C	-	W	-	-	I	P	-	-	I	F	-	-	I	-	-	L	K	-	-	F	-	L	-	-	S	L	-	-
rxfp1 Frog	F	F	I	V	F	T	-	-	D	-	-	A	L	-	-	C	-	W	-	-	I	P	-	-	I	F	-	-	I	-	-	L	K	-	-	L	-	L	-	-	S	L	-	-
rxfp1 Chicken	F	F	I	V	F	T	-	-	D	-	-	A	L	-	-	C	-	W	-	-	I	P	-	-	I	F	-	-	I	-	-	L	K	-	-	L	-	L	-	-	S	L	-	-
rxfp1 Zebrafish	F	S	I	V	I	T	-	-	D	-	-	S	L	-	-	C	-	W	-	-	I	P	-	-	I	F	-	-	I	-	-	L	K	-	-	T	-	L	-	-	S	L	-	-
RXFP2 Human	F	F	I	V	F	S	-	-	D	-	-	A	I	-	-	C	-	W	-	-	I	P	-	-	V	F	-	-	V	-	-	V	K	-	-	I	-	L	-	-	S	L	-	-
Rxfp2 Opossum	F	F	I	V	F	S	-	-	D	-	-	A	I	-	-	C	-	W	-	-	I	P	-	-	V	F	-	-	V	-	-	I	K	-	-	T	-	L	-	-	S	L	-	-
rxfp2 Frog	F	F	I	V	F	S	-	-	D	-	-	A	V	-	-	C	-	W	-	-	I	P	-	-	V	F	-	-	L	-	-	L	K	-	-	I	-	L	-	-	S	L	-	-
rxfp2 Tetraodon	F	F	I	V	F	S	-	-	D	-	-	A	L	-	-	C	-	W	-	-	I	P	-	-	I	F	-	-	L	-	-	V	K	-	-	I	-	L	-	-	S	L	-	-
Rxfp2-like Opossum	F	F	I	V	F	T	-	-	D	-	-	A	L	-	-	C	-	W	-	-	L	P	-	-	I	F	-	-	L	-	-	L	K	-	-	**L**	-	L	-	-	S	L	-	-

	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883
RXFP1 Human	-	Y	T	L	T	T	R	P	F	K	E	M	I	-	H	R	F	W	Y	N	Y	R	Q	R	K	S	M	D	S	K	G	Q	K	T	Y	-
Rxfp1 Opossum	-	Y	T	L	T	T	R	P	F	K	E	M	I	-	R	Q	V	W	S	N	Y	K	Q	R	R	S	I	E	S	K	G	S	Q	K	T	Y
rxfp1 Frog	-	Y	T	I	T	T	R	P	F	K	E	M	I	-	G	Q	I	W	S	N	Y	K	Q	R	R	S	I	G	N	R	N	S	H	K	A	C
rxfp1 Chicken	-	Y	T	L	T	T	R	P	F	K	E	M	I	-	H	Q	F	W	Y	N	Y	R	Q	R	R	S	K	R	**G**	K	G	S	Q	K	A	Y
rxfp1 Zebrafish	-	Y	T	L	T	T	R	P	F	K	E	T	L	-	L	Q	V	W	S	N	Y	R	Q	R	R	P	L	F	S	**S**	R	P	P	H	L	**P**
RXFP2 Human	-	Y	T	L	T	T	**N**	F	F	K	D	K	L	-	K	Q	L	L	H	K	H	Q	R	K	S	I	F	K	I	K	K	K	S	**L**	**S**	T
Rxfp2 Opossum	-	Y	T	L	T	T	**S**	F	F	K	D	K	L	-	K	Q	L	L	H	K	H	R	R	R	S	I	F	K	S	E	K	K	S	**L**	**S**	T
rxfp2 Frog	-	Y	T	L	T	T	S	F	F	K	E	K	L	-	K	Q	L	L	H	R	Q	R	R	R	S	V	F	R	N	E	R	K	S	**L**	S	T
rxfp2 Tetraodon	-	Y	T	L	T	T	S	F	F	R	E	Q	V	-	E	**V**	L	L	C	R	W	Q	R	R	Q	**V**	L	K	K	D	G	K	S	L	T	S
Rxfp2-like Opossum	-	Y	T	I	T	T	**S**	P	F	Q	E	R	L	-	K	Q	C	L	Q	R	**E**	G	R	**S**	V	P	S	S	Q	**S**	A	-	-	-	-	-

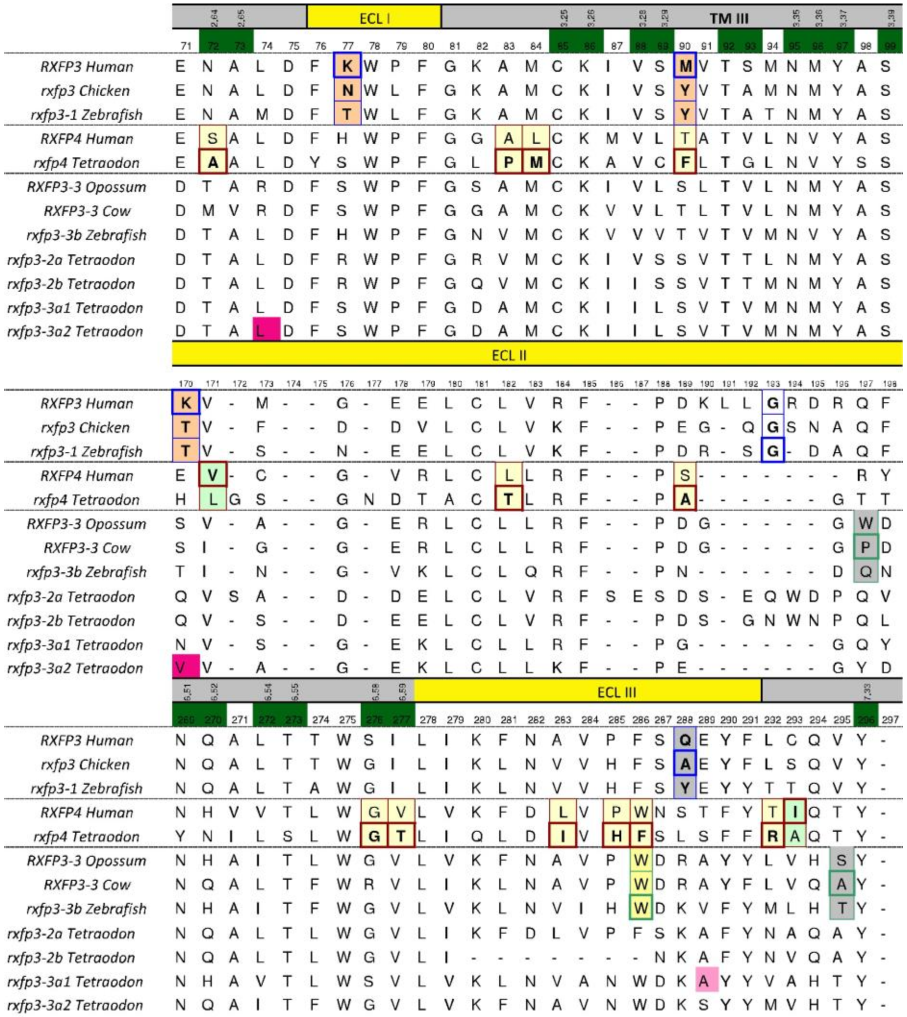


```
                     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
RXFP3 Human          - V R - I L I S V V - - - Y - W V V C A L G L A G N L L V L Y L M K S M - Q G W R K - S S I N L F V T N L A L T D F Q F V - L T L P F W A V
rxfp3 Chicken        - L R - I V I S L V - - - Y - S V V C A L G L V G N L L V L Y L M K S K - Q G W R K - S S I N L F V T S L A V T D F Q F V - L T L P F W A V
rxfp3-1 Zebrafish    - L R - I I I S I I - - - Y - S V V C A L G L I G N I L V L Y L M K S K - Q A W K K - S S I N L F V T S L A V T D F Q F V - L T L P F W A V
RXFP4 Human          - L R - L M V A L A - - - Y - G L V G A I G L L G N L A V L W V L S N C - A R R A P G P P S D T F V F N L A L A D L G L A - L T L P F W A A
rxfp4 Tetraodon      - - - - - - V S V V - - - Y - S V V A T A G V L V N A L V M F L L Y S T - H T I T T - G T I N F F V F N L A V A H L L F S - M V L V F W A V
RXFP3-3 Opossum      - L R - L L I A A V - - - Y - L A V C A V G L V G N V L V M V L V R A R - Q K R R P - S A L N C F V L N L A A T D F Q F V - L T L P F W A V
RXFP3-3 Cow          - - - - - - - - - - - - - Y - L A L C A V G L V G N V L V L V L V R S Q - Q R R R H - S L L N C F L L N L A A T D L Q F V - L T L P F W A V
rxfp3-3b Zebrafish   - L R - V T I S I V - - - Y - S V V C A V G L V G N L L V F F L I R V R - Q G R K K - S T I N Y F I M N L A I T D F Q F V - L T L P F W A V
rxfp3-2a Tetraodon   - L R - V L I A L V - - - Y - L V V C V L G L V G N L L A L L L L H S N - H R G H H - S S I D C F V M S L A L T D L Q F V - L T L P F W A V
rxfp3-2b Tetraodon   - V R - V M I A C V - - - Y - S I V C A L G L V G N A L A L Y L L H S R - H R Q K Q - S S I N C F V M G L A I T D L Q F V - L T L P F W A V
rxfp3-3a1 Tetraodon  - L R - F L I C V V - - - Y - S V V C A A G L V G N L L V L F F I R V K - R E R K K - S K V D F F V L N L A V T D L Q F V - L T L P F W A V
rxfp3-3a2 Tetraodon  - F M - I L I S V V - - - Y - S I V C A V G L V G N L L V F F L M R L R - Q G R K R - S T I N I F I I N L A V T D F Q F V - L T L P F W A V

                     100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169
RXFP3 Human          V F F L T A M S V T R Y H S V A S A L K S H R T - - - - - - - C C G R S L G D S C C F S A K A L C V W I W A L A A L A S L P S A I F S T T V
rxfp3 Chicken        V F F L T A M S V A R Y R S V A S A L K N Q G R - - - - - - - - - - G D P L G G C C - S T K W L C A L I W L L A I V A S L P Q A I F S T T A
rxfp3-1 Zebrafish    V F F L T A M S V A R Y C S V A S A L K S K R R - - - - - - - - - - - - - - - R L R F S A Q W M T V I I W I A A V G A A L P N A I F S T T A
RXFP4 Human          I F L I T A L S V A R Y W V V A M A A G P G T H - - - - - - - - - - - - - - - L S L F W A R I A T L A V W A A A A L V T V P T A V F G V E G
rxfp4 Tetraodon      C F F L A A M S L T R Y C Y V A T A L K A G G S - - - - - - - - - - - - - L C S R S S T S W A V T A F I W A A A V V A A S P R A V F A Q L S
RXFP3-3 Opossum      A F F L S A M S I A R Y F T V A C A L R R H S L - D S G G S V C C G R - - A T P G - - - - - W V C G L L W V A A S L A T A P T A L F S T V A
RXFP3-3 Cow          G F F L S A M S V A R Y Y A V A R A L R P R R P - - - - - - - - - - - - - - - - G T L W A V C V C S L L W A T A V L A T A P T A L F A T A A
rxfp3-3b Zebrafish   A F F L T A M S V T R Y W S V A S A L K D R T R - - - - - - - - - - - - - - Y K R C - S A R L L S A L L W L L A A A A S V P T A I V T S V K
rxfp3-2a Tetraodon   V F F L T A M S V T R Y R T L V T S L R M E S P - - - - - - - - - - - - - - E A A A A R V K W V S L A I W M V S L V A T L P H V I Y S T T V
rxfp3-2b Tetraodon   V Y F L T A M S V A R Y Y S I S S A L K M H S R - - - - - - - - - - - - - - R T A A S R A K W T S L G I W T V S L L A T L P H A I Y S T S A
rxfp3-3a1 Tetraodon  V F F L T A M S V T R Y L S V A S A L K N S D P - - - - - - - - - - - - - - H R S C - S V Q W L C A I I W T L A T L V T A P T P I Y S T V G
rxfp3-3a2 Tetraodon  V F F L T A M S V T R Y L S V A S A L K K K S C - - - - - - - - - - - - - - R S R C - - V K W V C A V L W V A A T V A T A P T A F F S T V T

                     199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268
RXFP3 Human          - - W L G L Y H S Q K V L L G F V L P L G I I I L C Y L L L V R F I A D R R A A G T K - - S A R R L S K V T K S V T I V V L S F F L C W L P
rxfp3 Chicken        - - W L G L Y Q T Q K V L L G F V L P L A I I S L C Y L L L V R F I S E K H V - G S T - - S T K R R S K V T K S V S I V V L S F F L C W L P
rxfp3-1 Zebrafish    - - W L G L Y H A Q K V L L G F L I P F G I I S I C Y L L L L R F I T N K N I N T S - - - S A K R R S K V T R S V T I V V L S F F L C W L P
RXFP4 Human          - - W L G A Y Q L Q R V V L A F M V P L G V I T T S Y L L L L A F L Q R R Q R R R - - - - - - Q D S R V V A R S V R I L V A S F F L C W F P
rxfp4 Tetraodon      - - W L G I N Q L L R M V L G F L L P Y A I I T L S Y L L L L R F L C R H N L R G S G - - - S R R K A N V S R S V A V V V L S F C I C W F P
RXFP3-3 Opossum      - - W L A L Y H V Q K I T V A F L L P L A T L S A C Y L L L V R F V R S R P R P W L R - - S S R G S S R V T R S V T V V V L A F F L C W L P
RXFP3-3 Cow          - - W L A F Y H L Q K I T V A F V L P L V T L G T C S L L L L R F L R R Q R V R W S S G V R H R R R S R V T R A L A C V L L V F V L C W L P
rxfp3-3b Zebrafish   - - W H A I Y H L Q K I L V G F V I P M L I T V V C Y A L L L R L I R R R S M N N N L - - P - K R R S R V T K S V T I V V L S F F I C W L P
rxfp3-2a Tetraodon   - - L L G L Y Q M Q K V L L G F V V P L V V I S V C Y L L L L S A A D T S E Y Q R G R - - Q - R R R S K V T R S V T I V V L S F F L C W L P
rxfp3-2b Tetraodon   - - L L G L Y Q L Q K V L L G F L I P L V I I T V C Y L L L L R L I L G R R I T G A A - - R Q S R R S K V T K S I V I V V L S F F L C W L P
rxfp3-3a1 Tetraodon  - - W L A V Y H I Q K I L V G F V M P M S V V S I S Y L M L L R F I R K R S M K T S S - - R - R R K S R V T R S I T I V V L S F F L C W M P
rxfp3-3a2 Tetraodon  - - W L A I Y H I Q K I L V A F I I P M L V V S V N Y L L L L R F L R L K S M S G S N - - S - R R R S R V T K S V A I V V L S F F F C W M P

                     298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329
RXFP3 Human          A F P V S V C L A H S N S C L N P V L Y C L V R R E F R K A L K
rxfp3 Chicken        L F P I S V C L A H S N S C L N P I L Y C L M R R E F R I C F S
rxfp3-1 Zebrafish    I F P V T V C L A H S N S C L N P I L Y C L M R R E F R K A L K
RXFP4 Human          V F P V T T C L A H S N S C L N P V L Y C L L R R E P R Q A L A
rxfp4 Tetraodon      F F P V A N C L A F T S S C F N P V I Y C L V R R E Y R T A L -
RXFP3-3 Opossum      L F P V T I C L A H A N S C L N P L L Y C L L R P E F R Q L L W
RXFP3-3 Cow          L F P V S I C L A H S N S C L N P L L Y C L L R R D F R Q G L Q
rxfp3-3b Zebrafish   V F P V T V C L A H S N S C L N P V L Y C L M R R E F R K M L K
rxfp3-2a Tetraodon   A F P L T V C L A H A N S C L N P V L Y C L I R K E Y R N G L K
rxfp3-2b Tetraodon   A F P L T V C L A H T N S C L N P V L Y C L I R H E F R A - - -
rxfp3-3a1 Tetraodon  V F P L T V C L A H T N S C L N P I I Y C L M R K E F R N K L R
rxfp3-3a2 Tetraodon  V F P V T V C L A H A N S C L N P V L Y C L M R P E I R K - - -
```



						ECL I																TM III							
		2.64	2.65												3.25	3.26		3.28	3.29						3.35	3.36	3.37		3.39
	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
RXFP3 Human	E	N	A	L	D	F	**K**	W	P	F	G	K	A	M	C	K	I	V	S	**M**	V	T	S	M	N	M	Y	A	S
rxfp3 Chicken	E	N	A	L	D	F	**N**	W	L	F	G	K	A	M	C	K	I	V	S	**Y**	V	T	A	M	N	M	Y	A	S
rxfp3-1 Zebrafish	E	N	A	M	D	F	**T**	W	L	F	G	K	A	M	C	K	I	V	S	**Y**	V	T	A	T	N	M	Y	A	S
RXFP4 Human	E	S	A	L	D	F	H	W	P	F	G	G	A	L	C	K	M	V	L	T	A	T	V	L	N	V	Y	A	S
rxfp4 Tetraodon	E	**A**	A	L	D	Y	S	W	P	F	G	L	**P**	**M**	C	K	A	V	C	**F**	L	T	G	L	N	V	Y	S	S
RXFP3-3 Opossum	D	T	A	R	D	F	S	W	P	F	G	S	A	M	C	K	I	V	L	S	L	T	V	L	N	M	Y	A	S
RXFP3-3 Cow	D	M	V	R	D	F	S	W	P	F	G	G	A	M	C	K	V	V	L	T	L	T	V	L	N	M	Y	A	S
rxfp3-3b Zebrafish	D	T	A	L	D	F	H	W	P	F	G	N	V	M	C	K	V	V	V	T	V	T	V	M	N	V	Y	A	S
rxfp3-2a Tetraodon	D	T	A	L	D	F	R	W	P	F	G	R	V	M	C	K	I	V	S	S	V	T	T	L	N	M	Y	A	S
rxfp3-2b Tetraodon	D	T	A	L	D	F	R	W	P	F	G	Q	V	M	C	K	I	I	S	S	V	T	T	M	N	M	Y	A	S
rxfp3-3a1 Tetraodon	D	T	A	L	D	F	S	W	P	F	G	D	A	M	C	K	I	I	L	S	V	T	V	M	N	M	Y	A	S
rxfp3-3a2 Tetraodon	D	T	A	L	D	F	S	W	P	F	G	D	A	M	C	K	I	I	L	S	V	T	V	M	N	M	Y	A	S

	ECL II																												
	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198
RXFP3 Human	**K**	V	-	M	-	-	G	-	E	E	L	C	L	V	R	F	-	-	P	D	K	L	L	**G**	R	D	R	Q	F
rxfp3 Chicken	**T**	V	-	F	-	-	D	-	D	V	L	C	L	V	K	F	-	-	P	E	G	-	Q	**G**	S	N	A	Q	F
rxfp3-1 Zebrafish	**T**	V	-	S	-	-	N	-	E	E	L	C	L	V	K	F	-	-	P	D	R	-	S	**G**	-	D	A	Q	F
RXFP4 Human	E	**V**	-	C	-	-	G	-	V	R	L	C	L	L	R	F	-	-	P	S	-	-	-	-	-	-	-	R	Y
rxfp4 Tetraodon	H	L	G	S	-	-	G	N	D	T	A	C	**T**	L	R	F	-	-	P	**A**	-	-	-	-	-	-	G	T	T
RXFP3-3 Opossum	S	V	-	A	-	-	G	-	E	R	L	C	L	L	R	F	-	-	P	D	G	-	-	-	-	-	G	W	D
RXFP3-3 Cow	S	I	-	G	-	-	G	-	E	R	L	C	L	L	R	F	-	-	P	D	G	-	-	-	-	-	G	P	D
rxfp3-3b Zebrafish	T	I	-	N	-	-	G	-	V	K	L	C	L	Q	R	F	-	-	P	N	-	-	-	-	-	-	D	Q	N
rxfp3-2a Tetraodon	Q	V	S	A	-	-	D	-	D	E	L	C	L	V	R	F	S	E	S	D	S	-	E	Q	W	D	P	Q	V
rxfp3-2b Tetraodon	Q	V	-	S	-	-	D	-	E	E	L	C	L	V	R	F	-	-	P	D	S	-	G	N	W	N	P	Q	L
rxfp3-3a1 Tetraodon	N	V	-	S	-	-	G	-	E	K	L	C	L	L	R	F	-	-	P	G	-	-	-	-	-	-	G	Q	Y
rxfp3-3a2 Tetraodon	V	V	-	A	-	-	G	-	E	K	L	C	L	L	K	F	-	-	P	E	-	-	-	-	-	-	G	Y	D

										ECL III																			
	6.51	6.52		6.54	6.55			6.58	6.59																			7.33	
	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297
RXFP3 Human	N	Q	A	L	T	T	W	S	I	L	I	K	F	N	A	V	P	F	S	**Q**	E	Y	F	L	C	Q	V	Y	-
rxfp3 Chicken	N	Q	A	L	T	T	W	G	I	L	I	K	L	N	V	V	H	F	S	**A**	E	Y	F	L	S	Q	V	Y	-
rxfp3-1 Zebrafish	N	Q	A	L	T	A	W	G	I	L	I	K	L	N	V	V	H	F	S	**Y**	E	Y	Y	T	T	Q	V	Y	-
RXFP4 Human	N	H	V	V	T	L	W	G	V	L	V	K	F	D	L	V	P	W	N	S	T	F	Y	T	**I**	Q	T	Y	-
rxfp4 Tetraodon	Y	N	I	L	S	L	W	**G**	**T**	L	I	Q	L	D	**I**	V	**H**	**F**	S	L	S	F	F	**R**	A	Q	T	Y	-
RXFP3-3 Opossum	N	H	A	I	T	L	W	G	V	L	V	K	F	N	A	V	P	W	D	R	A	Y	Y	L	V	H	S	Y	-
RXFP3-3 Cow	N	Q	A	L	T	F	W	R	V	L	I	K	L	N	A	V	P	W	D	R	A	Y	F	L	V	Q	A	Y	-
rxfp3-3b Zebrafish	N	H	A	I	T	F	W	G	V	L	V	K	L	N	V	I	H	W	D	K	V	F	Y	M	L	H	T	Y	-
rxfp3-2a Tetraodon	N	Q	A	L	T	L	W	G	V	L	I	K	F	D	L	V	P	F	S	K	A	F	Y	N	A	Q	A	Y	-
rxfp3-2b Tetraodon	N	Q	A	L	T	L	W	G	V	L	I	-	-	-	-	-	-	-	N	K	A	F	Y	N	V	Q	A	Y	-
rxfp3-3a1 Tetraodon	N	H	A	V	T	L	W	S	V	L	V	K	L	N	V	A	N	W	D	K	A	Y	Y	V	A	H	T	Y	-
rxfp3-3a2 Tetraodon	N	Q	A	I	T	F	W	G	V	L	V	K	F	N	A	V	N	W	D	K	S	Y	Y	M	V	H	T	Y	-

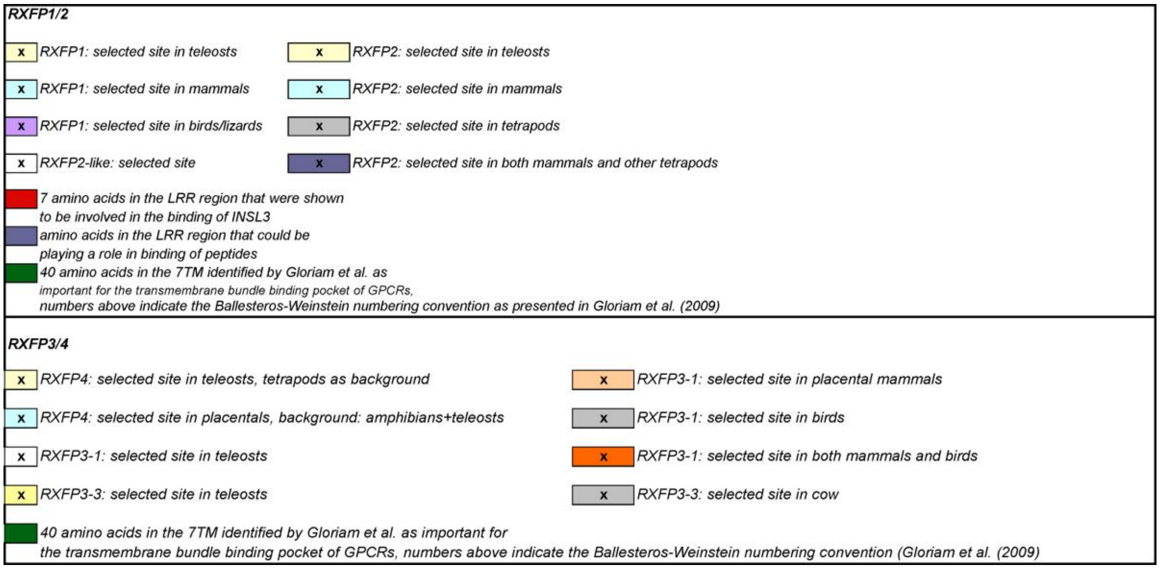

RXFP1/2
x RXFP1: selected site in teleosts
x RXFP2: selected site in teleosts
x RXFP1: selected site in mammals
x RXFP2: selected site in mammals
x RXFP1: selected site in birds/lizards
x RXFP2: selected site in tetrapods
x RXFP2-like: selected site
x RXFP2: selected site in both mammals and other tetrapods
7 amino acids in the LRR region that were shown to be involved in the binding of INSL3
amino acids in the LRR region that could be playing a role in binding of peptides
40 amino acids in the 7TM identified by Gloriam et al. as important for the transmembrane bundle binding pocket of GPCRs, numbers above indicate the Ballesteros-Weinstein numbering convention as presented in Gloriam et al. (2009)
RXFP3/4
x RXFP4: selected site in teleosts, tetrapods as background
x RXFP3-1: selected site in placental mammals
x RXFP4: selected site in placentals, background: amphibians+teleosts
x RXFP3-1: selected site in birds
x RXFP3-1: selected site in teleosts
x RXFP3-1: selected site in both mammals and birds
x RXFP3-3: selected site in teleosts
x RXFP3-3: selected site in cow
40 amino acids in the 7TM identified by Gloriam et al. as important for the transmembrane bundle binding pocket of GPCRs, numbers above indicate the Ballesteros-Weinstein numbering convention (Gloriam et al. (2009)

Figure 3.6. Amino acid positions found to be subject to positive selection in the foreground lineages when compared to those in the background lineage for RXFP1/2 and RXFP3/4 receptors.
Colored boxes surround the sites under positive selection (see description of symbols above). First four diagrams show RXFP1/2 receptors, last two diagrams show RXFP3/4 receptors.

DISCUSSION

The analyses in this Chapter showed that both *RLN/INSL* hormone and *RXFP* receptor genes exhibit every kind of selection: some of the genes are subject to strong purifying selection (RLN3), others are evolving relatively neutrally (INSL5), one ligand has been subject to strong positive selection (RLN), while another ligand has experienced more limited, but detectable, levels of positive selection (INSL3). Moreover, the analyses of the types and extent of selection operating on both ligand and receptor genes show that: 1) most, but not all, ligand-receptor pairs are evolving similarly in mammalian and teleostean lineages, 2) some ligand-receptor pairs show evidence of co-evolution, as assessed by a strong correlation in the proportion of sites under selection for ligand-receptor pairs, and 3) there is evidence of codon-specific positive selection for all receptor genes in most lineages and evidence that there is differential selection in mammalian and teleostean lineages.

I examined whether mammalian and teleostean orthologs of *RLN/INSL* and *RXFP* genes experienced similar kinds and degrees of selection by graphing the proportion of sites estimated to be under purifying, neutral or positive selection. I found that for all genes except relaxin (RLN), putative ligand-receptor pairs experienced similar levels of selection in both mammalian and teleostean lineages. This suggests that, in general, the genes may play similar roles in these two lineages. However, as will be discussed below, many amino acids exhibit differential selection in mammalian versus teleostean lineages suggesting that there have been different selective pressures in the two lineages.

The only gene pair for which there was a poor correlation in the nature of selection was for the comparison between mammalian and teleostean *RLN*: mammalian and teleost *RXFP1* have evolved in similar ways, but the relaxin gene has been subject to purifying and neutral evolution in teleosts, while it has been the target of strong positive selection in mammals (see Figure 3.2 and Figure 3.3). Approximately 50% of the amino acid positions in mammalian *RLN* show evidence of positive selection, whereas no sites in teleost *rln* do. The role of selection on the mammalian rln locus is more fully discussed in the next chapter.

Additionally, I examined the nature of selection in all teleostean genes to look for evidence of which ligands and receptors may function together based purely on the hypothesis that ligand-receptor pairs should exhibit the same kinds and extent of selection. While most of the putative teleostean orthologs of mammalian RLN-RXFP genes have evolved under similar forms of selection, many of the teleost 3R receptor genes are dominated by purifying selection. Earlier (see Chapter 2) I proposed that the two teleost 3R duplicates of rln3, namely rln3a and rln3b, function with rxfp3-1 (the ortholog of mammalian RXFP3), but also with the 3R paralogs, rxfp3-2a and rxfp3-2b. Both rln3a and rln3b are subject to strong purifying selection, and their proposed receptor genes have also predominantly evolved under purifying selection, with the exception of rln3-2b which shows a small proportion of sites that have been subject to positive selection. These findings are in agreement with previous studies (Wilkinson et al. 2005b)and further support the hypothesis about the highly conserved nature of the RLN3-RXFP3 system due to its neuroendocrine function.

On the other hand, I also proposed that the potential receptors for the two 3R-products of insl5, teleost insl5a and insl5b, are rxfp4, rxfp3-3a1, rxfp3-3a2 and rxfp3-3a3. As discussed, and as shown in a previous study from our lab (Good-Avila et al. 2009), *insl5* genes in both teleosts and mammals are evolving relatively neutrally. However, while the selection profile of rxfp4 matches that of its two 3R duplicated ligands in teleosts, all three rxfp3-3 receptors are dominated by purifying selection and have selection profiles similar to those of rln3. Thus, it is unclear which receptors are cognate to insl5a and insl5b based on the selection data alone.

The situation for teleost insl3-rxfp2 is simpler: their selection profiles are similar, suggesting a co-functioning of peptide and receptor. Additionally, rxfp2-like (which among teleosts is only present in zebrafish) also has a similar selection profile to insl3, suggesting that it may be a receptor for insl3 as well. Lastly, although the selection profile for teleost rln indicates that it has been subject predominantly to neutral evolution, this result may have been caused by the fact these values are calculated against a background of what has occurred to mammalian RLN which, as described, has been the subject to strong positive selection. The true selection profile for teleost rln is probably more similar to rln3 (as teleost rln3 and rln are structurally very similar) and is probably more closely defined by predominant purifying selection, with some neutral evolution as shown for its putative receptor, rxfp1.

The analysis of codon-specific positive selection revealed, not surprisingly, that some receptor domains are the targets of more selection than others. For this analysis, sites were deemed to be subject to codon-specific selection if, when comparing a particular branch of the phylogenetic tree for that gene, there was evidence that certain amino acids were selected to be different from those in the “background” lineage for the same gene. By analyzing the genes in this way, I found that for the RXFP1/2-type genes, the LDLa-LRR region of the N-terminus generally showed high levels of selection, which is perhaps not surprising because of the roles these domains play in receptor-ligand signaling (Halls et al. 2007). The LRR region is important for the binding of the cognate ligand, while the LDLa module is essential for cAMP accumulation which takes place after the ligand is recognized and bound (Halls et al. 2007). Approximately 20% of the amino acid sites in LDLa, LRR2, LRR4, LRR8, LRR8 and LRR11 were found to be subject to positive selection, and at least 30% of the sites were under selection for LRR flanking, LRR3, LRR5, LRR6, LRR7 and LRR9. Apart from these regions, the only other two regions which were identified as having more than 20% of the sites under selection for the RXFP1/2 type genes were ICL3 and ECL2.

In general, lineage-specific selection was higher for the RXFP3/4 type genes: all domains were found to have more than 20% of the amino acids subject to positive selection except for the TM1, TM2, TM3, TM7 and ECL1, the latter having exactly 20% of its sites estimated to be under selection across lineages. Of particular interest, is the fact that for the RXFP3/4 type genes, ICL1 is equally important as ICL3 in terms of selection. The finding that ICL3 (both receptor types) and ICL1 (RXFP3/4 type receptors) are targets of selection suggests that a major component of selection for the RXFP receptors concerns downstream receptor signaling rather than selection for ligand binding *per se*.

Ligand binding in the rhodopsin class GPCR receptors has been associated with the set of 40 amino acids composing the "transmembrane binding pocket (Gloriam et al. 2009). Interestingly, my analyses did not detect any significant difference in the number of selected sites between the binding pocket and other major domains of RXFP receptors. This may be due to high conservation of the amino acids involved in ligand binding, or it could also imply that the 40 amino acids may not necessarily be key to the specific ligand binding of RXFP receptors.

In addition to observing certain domains as the targets of selection, I also find that different domains have been the targets of selection in mammals versus teleosts. Although the numbers of amino acid changes were similar between mammalian and teleostean lineages, for all orthologs of the receptor genes, with the exception of RXFP3, teleosts were observed to have more amino acid sites under selection than mammals. For example, mammalian RXFP4 shows strong evidence of selection in both ICL1 and ICL3 suggesting that novel pathways for intracellular signaling may have been selected. On the other hand, selection in teleost rxfp4 occurs mostly in the TM domain and in ECL2 and ECL3 suggesting that, for teleosts, selection has operated mostly on ligand binding. In contrast, mammalian RXFP3 shows evidence of selection in both ICL3 and at several sites in the TM and ECL domains suggesting that both ligand binding and intracellular signaling have been targets of selection, while very few sites have been positively selected in teleost rxfp3, consistent with the highly conserved nature of rln3 paralogs (particularly rln3a) in teleosts.

The pattern of selection for RXFP1/2-type genes is less clear, but most of the amino acids selected in mammalian RXFP1 occurred at the N-terminus between the LDLa domain and LRR2, or in ICL3, while for teleostean rxfp1, the selected sites were scattered throughout the receptor domains. Lastly, the majority of selected amino acids for both teleostean and mammalian RXFP2 genes occurred in either the LRR flanking region, LRR6 or else were scattered throughout the TM/ICL and ECL domains. However, as for RXFP4, there was tendency for mammalian RXFP2 to have more selected sites in ICL3, while teleostean rxfp2 had no selected sites on ICL3, but several on the ECL domains. Collectively, these data suggest that despite overall similarity in many of the selective processes among teleostean and mammalian genes, there is evidence that distinct signaling pathways have been selected in different groups and that some lineages have been selected to modify ligand-binding while other lineages have had more changes in intracellular signaling pathways. It will be interesting to see whether experimental work supports these hypotheses.

CHAPTER 4: *Reconstruction of the structure of ancestral relaxin family peptides. Selection on RLN/INSL loci.*

INTRODUCTION

Thus far, I have looked at the duplication history and origins of the relaxin family peptides and their receptors (Chapter 1), discussed the main highlights of their diversification in different vertebrate lineages (Chapter 2) and established hypotheses to explain the functional specialization and co-evolutionary processes by which they evolved (Chapter 2). In addition, to test the hypotheses derived in Chapter 2, I looked at the effects of natural selection on the evolution of relaxin family peptides and their receptors in teleosts and human (Chapter 3). However, the story of a gene family's evolution could not be complete without a description of the structural changes that have taken place during the evolutionary history of the concerned molecules. Therefore in this Chapter, I use the established dataset of vertebrate RLN/INSL ligand sequences to reconstruct the primary structures of ancestral peptides at different periods of vertebrate evolution and in different vertebrate lineages to show how they have changed over time. The ancestral structure of relaxin peptide receptors is not discussed here (for reasons described further below), but it is suggested that the ancestral state reconstruction of both RXFP1/2 and RXFP3/4 be done in future studies.

Ancestral gene reconstruction is a method that allows, with some caveats, one to study the properties of long lost genes and their products from ancestral organisms. This method can be used, for example, to investigate and compare the functions of genes from modern species with their orthologs from more ancient organisms, or to study the functional characteristics of an ancestral gene that gave rise to a specific gene family (Thornton 2004). Major advances in the field of ancestral gene reconstruction have been made in recent years thanks to the progress in bioinformatics and nucleic acid synthesis. The procedure of resurrecting an ancestral gene principally consists of two stages (Thornton 2004): 1) the sequence of the hypothetical ancestral sequence is derived based on the set of genes/proteins available from extant taxa; 2) the predicted ancestral sequence is synthesized using oligonucleotide synthesis technologies with consequent use of bacterial or yeast cell culture to produce an intact protein (if such is needed). Whereas the second stage of the procedure requires the application of complex laboratory equipment and methodologies, the first step of ancestral resurrection is performed using bioinformatic algorithms and software. Here, for instance, the set of characterized vertebrate RLN/INSL peptides was analyzed using the ML algorithm. First, the relationship among the extant vertebrate peptide sequences was reconstructed using phylogenetic inference. Second, using the topology of the resulting phylogenetic tree, ML methods were employed. These ML methods used the individual amino acid sequences to estimate the most likely ancestral structures for each set of compared sequences. Third, at selected nodes in the tree (which represent ancestral clades on the generated phylogeny), the ancestral states of RLN/INSL peptides were reconstructed by choosing the most statistically supported structure given the model of sequence evolution employed (Figure 4.1). The statistical confidence with which a peptide sequence for a given ancestral clade is reconstructed is the product of the probabilities of each individually predicted ancestral amino acid state, which generally means that shorter peptides will be the reconstructed with statistically more reliable results. Conversely, the longer the reconstructed molecule, the lower is the probability that it actually existed. The size of the mature relaxin peptides (~ 60 aa) is ideal for ancestral gene reconstruction (Thornton 2004). However this is not the case with relaxin peptide receptors (400-800 aa long), whose size renders the reconstruction process statistically weak. For the purposes of co-evolutionary studies, it would be more practical to reconstruct the parts of RXFP receptors which directly participate in ligand binding. Since only a few of the functionally important sites in RXFP receptors have been identified to date, I leave the task of reconstructing their ancestral states to future studies.

As mentioned above, ancestral reconstruction has some caveats which should be considered with caution. Thus the results of a reconstruction can be influenced by the assumed model of evolution of genes and taxa, tree topology and orthologous/paralogous relationships among members of protein families. The effect of one of the above mentioned factors, the assumed model of gene duplication, will be discussed further in this chapter using Park et al.'s (2008) study focusing on the evolution of INSL3 as an example.

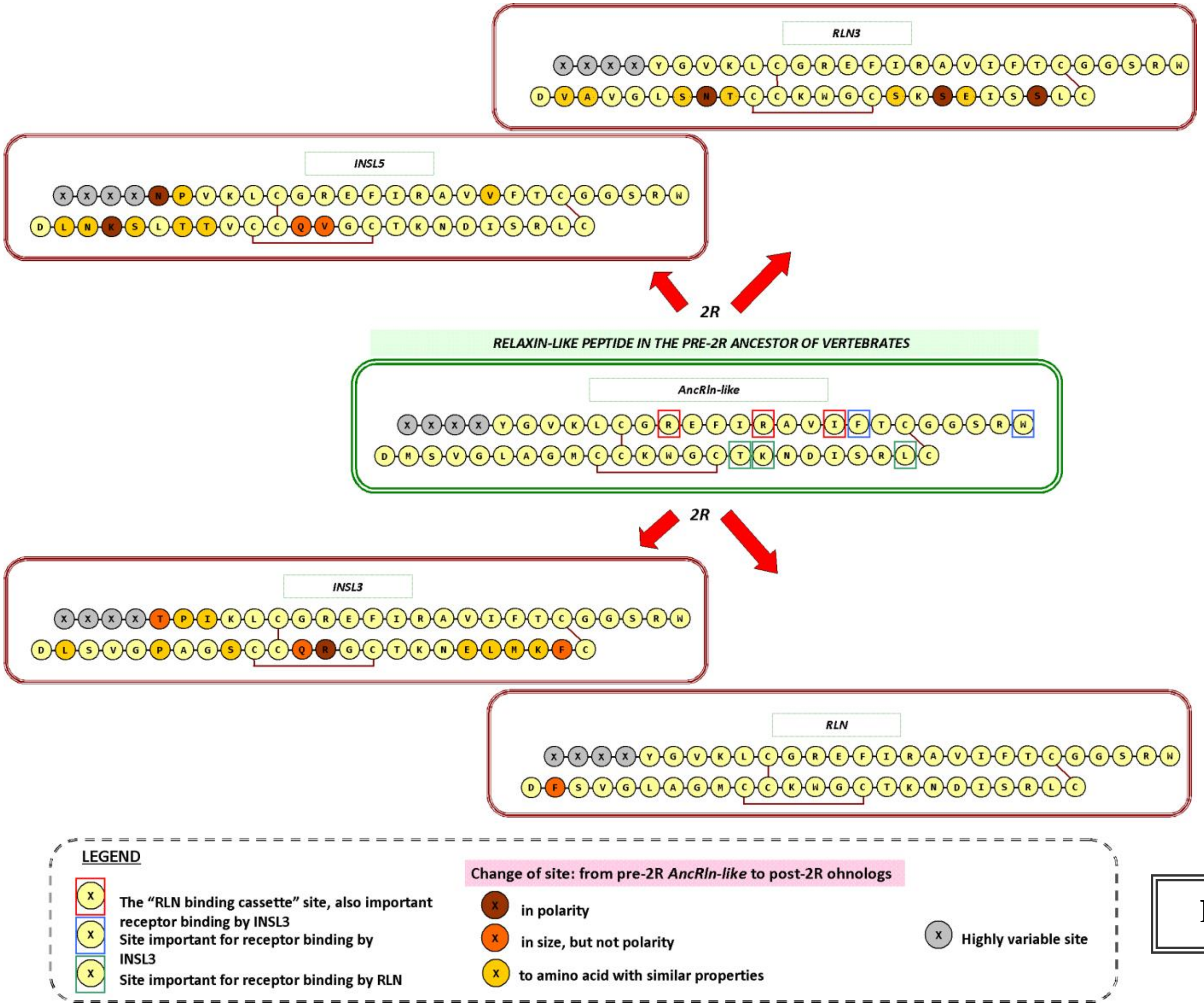

RLN3
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D V A V G L S N T C C K W G C S K S E I S S L C
INSL5
X X X X N P V K L C G R E F I R A V V F T C G G S R W
D L N K S L T T V C C Q V G C T K N D I S R L C
2R
RELAXIN-LIKE PEPTIDE IN THE PRE-2R ANCESTOR OF VERTEBRATES
AncRln-like
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D M S V G L A G M C C K W G C T K N D I S R L C
2R
INSL3
X X X X T P I K L C G R E F I R A V I F T C G G S R W
D L S V G P A G S C C Q R G C T K N E L M K F C
RLN
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D F S V G L A G M C C K W G C T K N D I S R L C
LEGEND
X The “RLN binding cassette” site, also important receptor binding by INSL3
X Site important for receptor binding by INSL3
X Site important for receptor binding by RLN
Change of site: from pre-2R AncRln-like to post-2R ohnologs
X in polarity
X in size, but not polarity
X to amino acid with similar properties
X Highly variable site

Figure 4.1

Figure 4.1. Ancestral state reconstruction of the four relaxin family peptide ohnologs which existed in the post-2R vertebrate organism (the ancestor of euteleostomi) before the diversification of modern vertebrate lineages.
It seems possible that *AncRln-like* may have acquired the classical relaxin family motifs, such as its receptor-binding cassette, just before the onset of 2R, because this motif is not found in the members of the insulin-relaxin superfamily in primitive deuterostomes (Chapter 1). Receptor binding sites of peptides (boxed) were obtained from the literature. Amino acids are shown as circles, amino acid changes (substitutions) that occurred in ohnologs are classifies as follows:

- change in polarity → from nonpolar (neutral) to polar and vice versa (brown)
- change in size → from relatively small to relatively large and vice versa (orange)
- change to amino acid with similar characteristics → change occurs within groups of amino acids of same polarity and similar size (yellow)

METHODS

Ancestral state reconstruction and evidence of codon-specific selection in the ancestral genes of RLN/INSL peptides

The ancestral states of specific nodes on the *RLN/INSL* phylogeny were inferred using Maximum Likelihood methods as implemented in MEGA v. 5.01 (Tamura et al. 2011) using the JTT+G matrix-based model of sequence evolution (the model chosen based on AIC criterion using ML inference as implemented in the program Prottest, (Abascal et al. 2005)) and after excluding highly divergent sequences (shown in Figure 4.4). The ancestral nodes selected for reconstruction are shown in Figure 4.4. As shown previously (Good-Avila et al. 2009, Park et al. 2008, Wilkinson et al. 2005b), *RLN/INSL* loci have been subject to diverse selection pressures and to further assess the role of selection during lineage specific diversification of the peptides, the amino acid sites subject to codon-specific selection were estimated using the branch-site model A on orthologous gene families from distinct vertebrate lineages. Tests of positive selection were made by comparing the branch-site model A in which ω (dn/ds) ≥ 1 (alternative hypothesis) to the model A in which ω = 1 fixed (null hypothesis) and setting the foreground branch to be that of a strongly supported vertebrate clade (teleost, marsupial/monotreme, birds/lizards and/or mammals depending on the gene) while the background branch was then left as the remaining vertebrate genes/clades. The analysis was done using the CODEML package from PAML (PAML v. 4.2); models were compared using the Likelihood Ratio Test with 1 degree of freedom and, where significant, the posterior probability that a codon was under positive selection was estimated using the Bayes empirical Bayes (BEB) procedure (Zhang et al. 2005)

RESULTS and DISCUSSION

Ancestral state reconstruction and evidence of codon-specific selection in the ancestral genes coding for RLN/INSL peptides

To elucidate the structure of the pre-2R Rln/Insl peptide and to address when and how the relaxin family hormones diverged over time and in different lineages, the following ancestral states of relaxin family peptides were reconstructed:

1) the pre-2R peptide (also named AncRln-like); 2) the four RLN/INSL 2R ohnologs (RLN, RLN3, INSL3 and INSL5) that existed shortly after 2R in the ancestor of all euteleostomi [the term “ancestor of euteleostomi” and not “gnathostome ancestor” is used here because jawless and cartilaginous fish relaxin peptide sequences were omitted from this ancestral reconstruction due to their incompleteness], and 3) the RLN/INSL ancestors in specific vertebrate clades supported by the phylogenetic tree.

Additionally, as described in Chapter 3, the branch-site test of positive selection was conducted on the ancestral ligand peptides. In this test, branches on the phylogenetic tree are specified *a priori* by the user to be “foreground” branches, and then the amino acids that show evidence of positive selection are identified in that lineage by comparing the amino acids changes in the lineage to those in a similarly selected background branch, which is chosen to consist of sequences of the same peptide but in a different branch. Thus, in a sense, this test is complementary to the ancestral state reconstruction, because it ultimately looks for evidence of positive selection in amino acid positions at the ancestral nodes of the phylogenetic tree.

The ancestral state reconstruction presented in Figure 4.1 shows that the AncRln-like peptide strongly resembles today’s vertebrate RLN3 and non-placental RLN peptides. Most of the changes in the structure of RLN3 were acquired in a period following 2R and before the divergence of multiple vertebrate taxa, after which RLN3 evolved in a remarkably conserved fashion in most vertebrates (for example, zebrafish rln3 is ~70 % similar to that of human). According to Wilkinson et al. (2005), the highly conserved nature of RLN3 across multiple vertebrate taxa mirrors its conserved role as a neuropeptide, which also implies that the function of RLN3 in the CNS was established early in vertebrate evolution.

Somewhat surprisingly, it appears that RLN exhibited a slower rate of evolution than RLN3 (Figure 4.2b), but acquired a few lineage specific mutations, predominantly in its A-chain in the ancestors of teleosts and most tetrapods (Figure 4.2b). Then, exceptionally, in placental mammals a burst of mutations in both the B- and A-chains occurs and a remarkable 23 amino acids show evidence of codon-specific positive selection including sites in the B-chain pro-hormone cleavage site (Figure 4.2b). This placental-specific sudden leap in the slow paced evolution of vertebrate RLN seems to have been coincident with: 1) the diversification of placentals as a group and 2) the massive local duplications of the RLN locus (see Chapter 2).

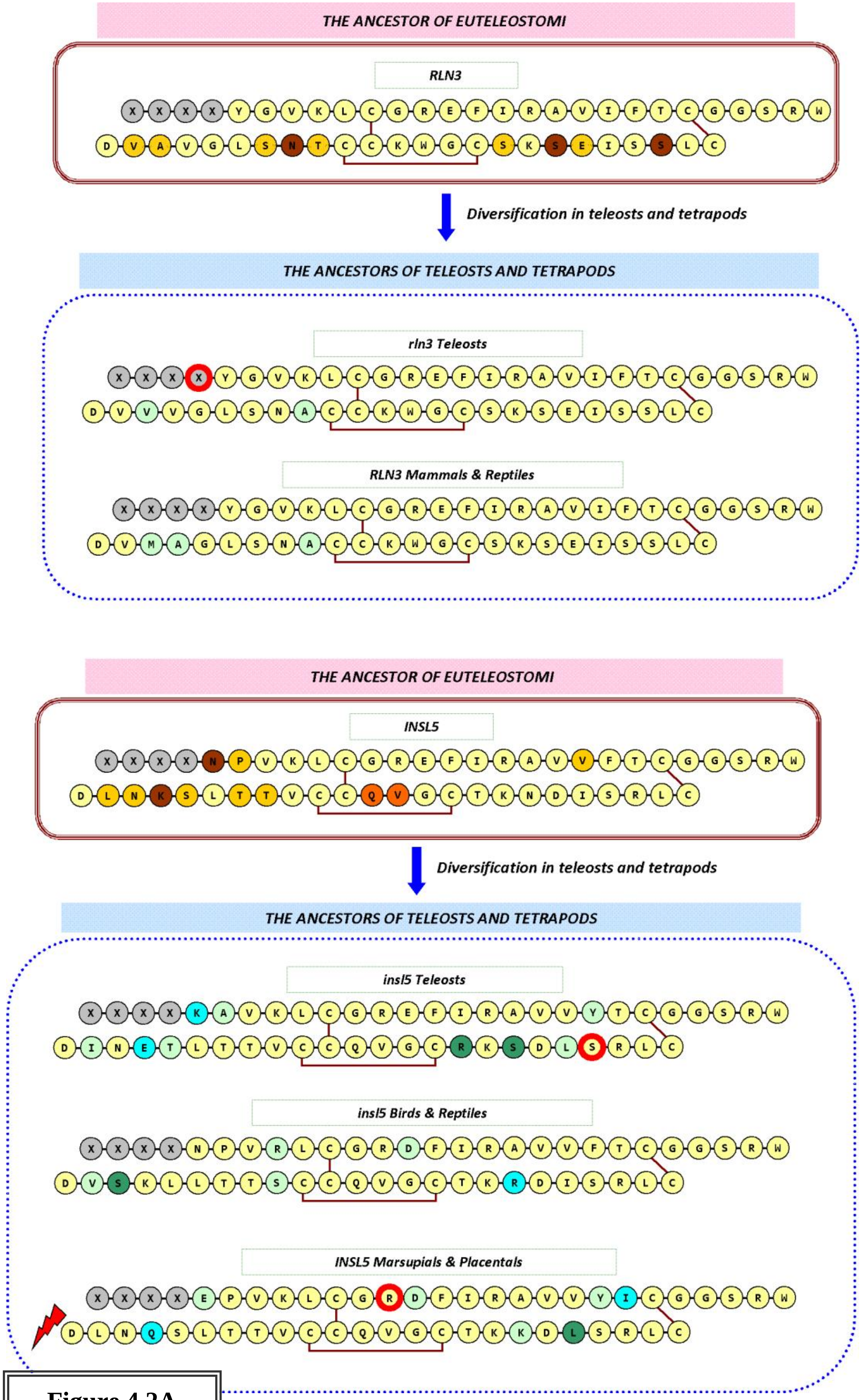

THE ANCESTOR OF EUTELEOSTOMI
RLN3
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D V A V G L S N T C C K W G C S K S E I S S L C
Diversification in teleosts and tetrapods
THE ANCESTORS OF TELEOSTS AND TETRAPODS
rln3 Teleosts
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D V V V G L S N A C C K W G C S K S E I S S L C
RLN3 Mammals & Reptiles
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D V M A G L S N A C C K W G C S K S E I S S L C
THE ANCESTOR OF EUTELEOSTOMI
INSL5
X X X X N P V K L C G R E F I R A V V F T C G G S R W
D L N K S L T T V C C Q V G C T K N D I S R L C
Diversification in teleosts and tetrapods
THE ANCESTORS OF TELEOSTS AND TETRAPODS
insl5 Teleosts
X X X X K A V K L C G R E F I R A V V Y T C G G S R W
D I N E T L T T V C C Q V G C R K S D L S R L C
insl5 Birds & Reptiles
X X X X N P V R L C G R D F I R A V V F T C G G S R W
D V S K L L T T S C C Q V G C T K R D I S R L C
INSL5 Marsupials & Placentals
X X X X E P V K L C G R D F I R A V V Y I C G G S R W
D L N Q S L T T V C C Q V G C T K K D L S R L C

Figure 4.2A

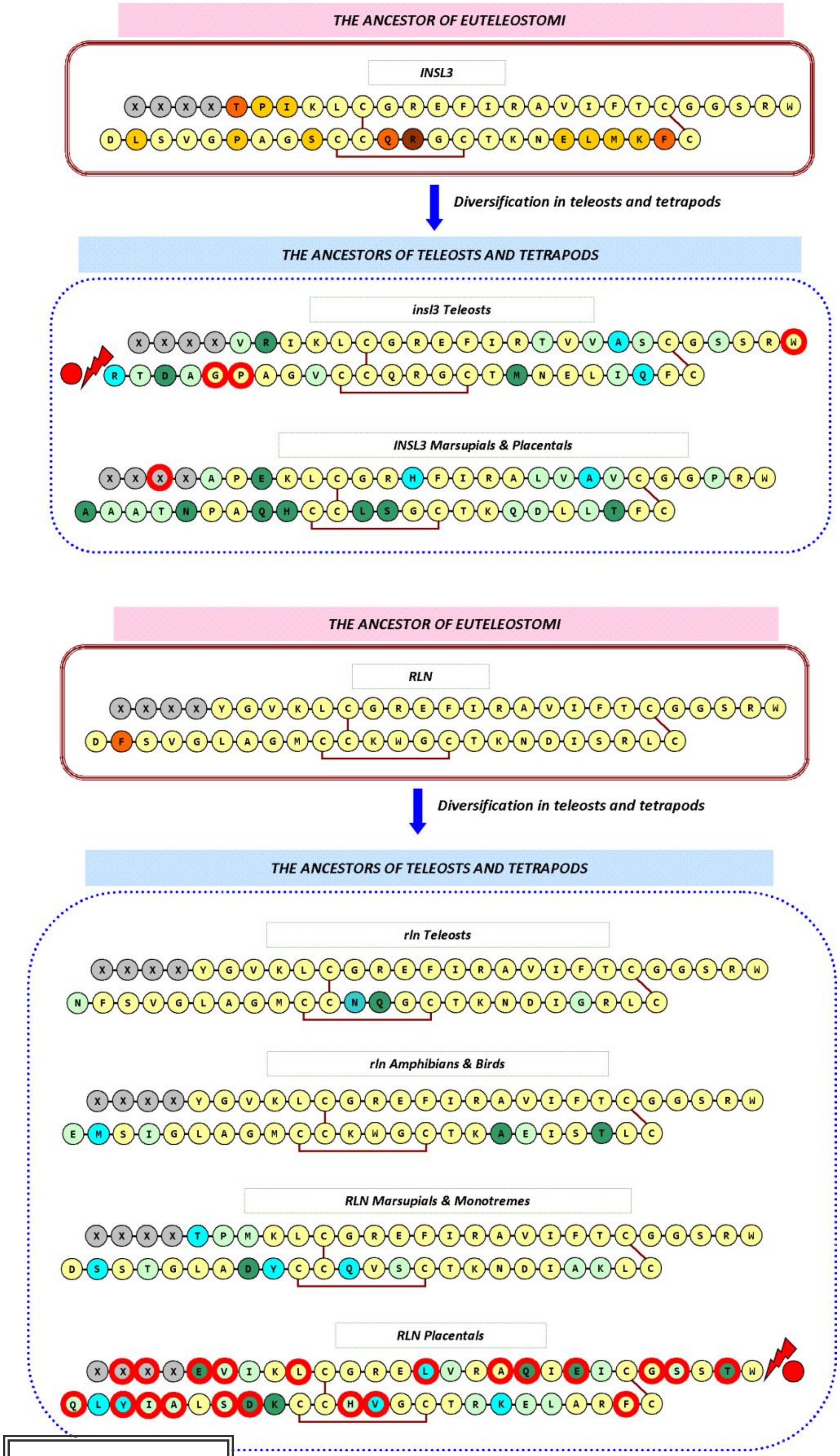

THE ANCESTOR OF EUTELEOSTOMI
INSL3
X X X X T P I K L C G R E F I R A V I F T C G G S R W
D L S V G P A G S C C Q R G C T K N E L M K F C
Diversification in teleosts and tetrapods
THE ANCESTORS OF TELEOSTS AND TETRAPODS
insl3 Teleosts
X X X X V R I K L C G R E F I R T V V A S C G S S R W
R T D A G P A G V C C Q R G C T M N E L I Q F C
INSL3 Marsupials & Placentals
X X X X A P E K L C G R H F I R A L V A V C G G P R W
A A A T N P A Q H C C L S G C T K Q D L L T F C
THE ANCESTOR OF EUTELEOSTOMI
RLN
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
D F S V G L A G M C C K W G C T K N D I S R L C
Diversification in teleosts and tetrapods
THE ANCESTORS OF TELEOSTS AND TETRAPODS
rln Teleosts
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
N F S V G L A G M C C N Q G C T K N D I G R L C
rln Amphibians & Birds
X X X X Y G V K L C G R E F I R A V I F T C G G S R W
E M S I G L A G M C C K W G C T K A E I S T L C
RLN Marsupials & Monotremes
X X X X T P M K L C G R E F I R A V I F T C G G S R W
D S S T G L A D Y C C Q V S C T K N D I A K L C
RLN Placentals
X X X X E V I K L C G R E L V R A Q I E I C G S S T W
Q L Y I A L S D K C C H V G C T R K E L A R F C

Figure 4.2B

LEGEND

Change of site: from *euteleostomian* to *teleost* and *tetrapod* forms

- ● x in polarity
- ● x in size, but not polarity
- ○ x to amino acid with similar properties
- ● x Highly variable site
- ◯ x Site under positive selection
- ● Positively selected site in the prohormone convertase cleavage region

Figure 4.2 and its legend. Ancestral state reconstruction of relaxin peptides in the ancestors of main vertebrate lineages.
A) RLN3 and INSL5; B) INSL3 and RLN. The amino acid symbols and classification of amino acid substitution are as in *Figure 4.1*.

Both *INSL5* and *INSL3* acquired several changes immediately following 2R but then seem to have evolved steadily in different lineages (Figure 4.2), with evidence of codon-specific positive selection in the ancestor of teleosts and marsupial/placental mammals. In the case of *INSL5* only one codon in each teleosts and marsupials/placentals (S and R respectively in Figure 4.2A) was selected, and the A-chain pro-hormone cleavage site was additionally found to be under selection in mammals. In the case of *INSL3*, one codon (the highly variable site “X” in the B-chain in Figure 4.2B) shows evidence of selection in marsupials/placentals, and three sites plus the A-chain pro-hormone cleavage site are subject to lineage-specific selection in teleosts (Figure 4.2).

Ancestral gene reconstruction requires clear understanding of evolutionary pathways: Park et al.’s study

The present study is not the first one to look at the ancestral states of relaxin family peptides. Although previous attempts to reveal the structure of the ancestral peptides were inevitably constrained by the absence of a unified and correct model for the duplication history of *RLN/INSL* genes, they still had a big impact at the understanding of the evolution of relaxin family peptides. One such study is the well-known work of Park et al. (2008), in which the authors aimed to delineate the evolutionary origins of INSL3-

mediated testicular descent in mammals. Park et al. (2008) proposed that *INSL3* emerged in the monotreme ancestor as a result of a small-scale duplication from an ancestral *RLN3* gene. The authors further hypothesized that ancestral RLN3 originally functioned via two kinds of receptors, RXFP1 and RXFP2. Subsequent to the duplication of the ancestral gene, they propose that its products, RLN3 and INSL3, subfunctionalized by type of receptors to produce RLN3-RXFP1 and INSL3-RXFP2 ligand-receptor pairs (Figure 4.3A).

To explore their hypotheses, Park et al (2008) used a set of vertebrate RLN3 and INSL3 sequences and reconstructed the structure of the hypothesized peptide ancestral to both RLN3 and INSL3. The functional analyses of the reconstructed peptide indicated that it was capable of activating both RXFP1 and RXFP2 human receptors, which the authors took as inarguable evidence in support of their hypotheses. But were they right?

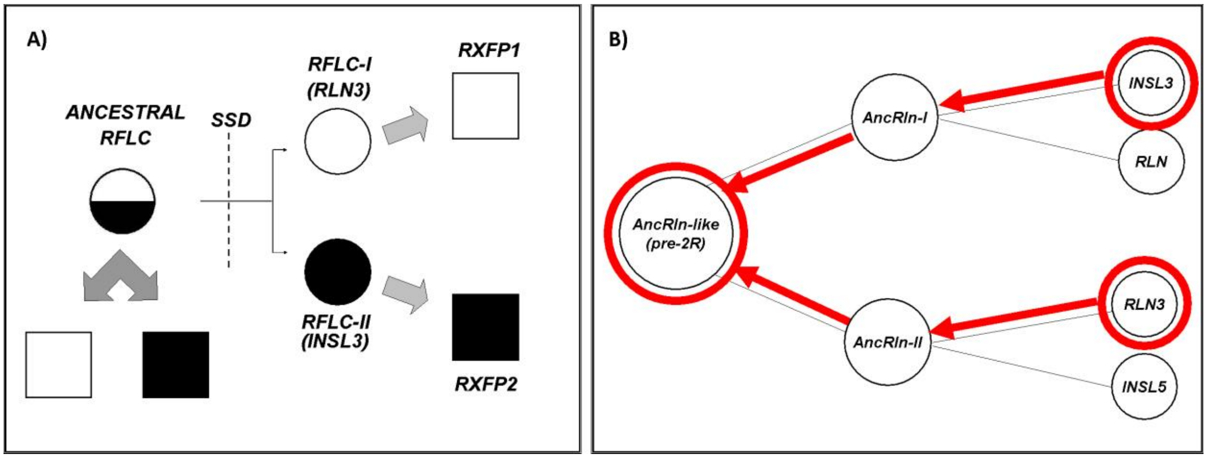


Figure 4.3. Diagrammatic representation of Park et al.'s hypotheses.
A) The hypotheses of Park et al.'s describing the origin of *RLN3* and *INSL3* genes from an ancestral *RLN3*-like gene and the subfunctionalization of RLN3 and INSL3 to function with RXFP1 and RXFP2 respectively. SSD: small-scale duplication; and B) the result of the ancestral gene reconstruction performed by Park et al. (2008). Red circles represent the genes that were used in the reconstruction (vertebrate *RLN3* and *INSL3*). Red arrows depict the evolutionary pathway to the gene ancestral to both *RLN3* and *INSL3*, which was reconstructed by Park et al.(2008) and which in fact is equivalent to *AncRln-like*.

Keeping in mind that the four vertebrate relaxin family loci arose ~550 MYA as a result of 2R and not as a consequence of a local duplication in the monotreme ancestor, it becomes obvious that Park et al.'s reconstructed ancestral peptide is nothing else but the common ancestor of all four relaxin peptides, equivalent to the AncRln-like peptide reconstructed here (Figure 4.1). Interestingly, the B-chain of Park et al.'s ancestral peptide is identical to that of AncRln-like. The major flaw in Park et al.'s work was hence the use of a wrong model of evolution for RLN3 and INSL3, which dated their emergence to the more recent history of RLN/INSL peptides.

An interesting ramification of this conclusion is that it means that the ancestral relaxin peptide, synthesized by Park et al's lab, and predicted to be the ancestor of all relaxin family genes here, is capable of binding both RXFP1/2- and RXFP3/4-type receptors. Such serendipitous support provided by Park et al.'s reconstructed peptide, lends further weight to the hypothesis presented earlier in this work about the dual functionality of AncRln-like and its ability to work via two distinct kinds of receptors (see Chapters 1 and 2).

CONCLUSIONS

In the last decade we have witnessed enormous progress in the fields of molecular biology and evolution. The relaxin field has not been excluded from this progress, although many of the aspects of molecular evolution of relaxin peptides and their receptors have to date been unclear and confusing to the scientific community. For instance, it has until very recently been believed that three of the four distinct relaxin family peptides (i.e. RLN, INSL3 and INSL5) are specific to mammals and have no orthologs in other vertebrates (Wilkinson et al. 2005), which was lately shown not to be the case (Good-Avila et al. 2009). In another instance, it was even once claimed that relaxin is not susceptible to evolution and that its structure has remained static for 500 MY(Georges and Schwabe 1999)!

This study provides evidence in support of the WGD-driven model for the origination of relaxin hormones and their two distinct classes of receptors in vertebrates. I postulate that the relaxin hormone-receptor signaling system in the pre-2R ancestor consisted of three components, one ligand and two receptors, and had a dual (reproductive and neuroendocrine regulatory) function. The genetic linkage of *RLN/INSL* and *RXFP3/4* genes, which is still highly conserved in teleosts, probably played a role in the original establishment of ligand-receptor interactions between ancestral RLN/INSL peptides and RXFP3/4 peptides in invertebrate deuterostomes. I show that most of the ligand and receptor genes duplicated during 2R (or 3R) and that, compared to tetrapods, teleosts have had significantly higher post-2R retention rates of *RXFP* genes.

Overall, this study highlights the utility of incorporating ancestral genome data into investigations of the origin, linkage relationship and duplication history of gene families. The methodology employed here (such as the use of ancestral genome reconstructions) will hopefully be useful in similar studies, where traditional approaches may fail to clearly resolve the origin and diversification of genes due to their small size, strong roles of selection or insufficient synteny data. Presently, however, a major drawback of the method is the absence of a unified scheme, which would avoid having to perform the time consuming and tedious manual inspection of multiple ancestral genome reconstruction models. In the future this problem could be resolved by designing appropriate computer software. Thus, rather than being viewed as a primarily heuristic tool for studying large scale genome evolution, ancestral genome reconstructions have a potential to form the basis of an instrument that could be routinely consulted to supplement traditional bioinformatic analyses.

Much of the current knowledge on relaxin family peptides and their receptors has come from the studies performed in rodents and humans. My searches of public databases indicated that both the ligands and receptors have had different fates throughout the evolution of vertebrates and that the human/rodent-derived properties of the family may not be applied to every mammal, not to mention other vertebrates.

Ligand-receptor signaling systems present interesting cases in which to study the evolution of genes, partly because they represent clearly defined sets of interacting molecules, whose origin and co-evolutionary dynamics can be investigated within a well-defined context. Both the relaxin family peptides, the RXFP3/4–type receptors and, to a lesser extent, the RXFP1/2-type receptors exhibit high rates of post-WGD retention. This finding is less surprising in light of the finding that GPCRs, in general, appear to have played an important role in the evolution of vertebrate signaling networks and were preferentially retained during the 2R (Semyonov et al. 2008). It is interesting that the RXFP1/2 receptors are more conserved than their RXFP3/4 counterparts and that the *RXFP1/2* genes differ from *RXFP3/4* genes in their rates of WGD-duplicated gene retention with RXFP3/4 paralogs having been retained more often.

As I elucidate in Chapter 3, analysis of the levels of functional constraint on RLN/INSL and RXFP genes suggests that 1) the RLN3-RXFP3-1 and INSL3-RXFP2 systems appear to be co-evolving based on the similarity of the selection profiles of ligand and receptor genes and 2) mammalian and teleost genes have somewhat similar roles in mammals and teleosts based on the observation of highly similar selection profiles for all focal genes with the notable exception of the RLN-RXFP1 system.

The analysis of codon and lineage-specific positive selection also highlighted differences in the functional domains of the RXFP genes that are under selection and potential differences between diversifying selection among teleost and mammalian genes.

Chapter 4 furthered our knowledge about the duplication history of relaxin family peptides to reveal the ancestral states of each of the four ohnologous RLN/INSL peptides. The findings about the evolution of relaxin hormones and their receptors will hopefully facilitate further research on this system in various vertebrates, including both placental and non-placental taxa. For instance, the discussed 2R-driven model of evolution should raise questions about the number of involved genes in early diverging vertebrates, such as jawless fish, whose status in relation to 2R has been debated (Kasahara 2007).

REFERENCES

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21: 2104-2105.

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. 2002. Evidence of en bloc duplication in vertebrate genomes. Nat Genet 31: 100-105.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. Genome Biol 5: R23.

Bieche I, Laurent A, Laurendeau I, Duret L, Giovangrandi Y, Frendo JL, Olivi M, Fausser JL, Evain-Brion D, Vidaud M. 2003. Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol Reprod 68: 1422-1429.

Braasch I, Brunet F, Volff J-N, Schartl M. 2009. Pigmentation pathway evolution after whole-genome duplication in fish. Mol Biol Evol 1: 479-493.

Cyranoski D. 2009. Developmental biology: Two by two. Nature 458: 826-829.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3: e314.

Donizetti A, Fiengo M, Minucci S, Aniello F. 2009. Duplicated zebrafish relaxin-3 gene shows a different expression pattern from that of the co-orthologue gene. Dev Growth Differ 51: 715-722.

Dschietzig T, Bartsch C, Greinwald M, Baumann G, Stangl K. 2006. The pregnancy hormone relaxin binds to and activates the human glucocorticoid receptor. Ann N Y Acad Sci 1041: 256-271.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

Feng S, et al. 2009. INSL3/RXFP2 signaling in testicular descent. Ann N Y Acad Sci 1160: 197-204.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. 151(4): 1531-1545.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman M. 2002. Evolutionary rate in the protein interaction network. Science 296.

Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63: 1256-1272.

Georges D, Schwabe C. 1999. Porcine relaxin, a 500 million-year-old hormone? The tunicate Ciona intestinalis has porcine relaxin. FASEB J 13: 1269-1275.

Gloriam DE, Foord SM, Blaney FE, Garland SL. 2009. Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. J Med Chem 52: 4429-4442.

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. J Mol Biol 299: 283-293.

Good-Avila SV, Yegorov S, Harron S, Bogerd J, Glen P, Ozon J, Wilson BC. 2009. Relaxin gene family in teleosts: phylogeny, syntenic mapping, selective constraint, and expression analysis. BMC Evol Biol 9: 293.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307-321.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol 8(10): R209.

Halls ML, van der Westhuizen ET, Bathgate RA, Summers RJ. 2007. Relaxin family peptide receptors--former orphans reunite with their parent ligands to activate multiple signalling pathways. Br J Pharmacol 150: 677-691.

Hisaw F. 1926. Experimental relaxation of the pubic ligamentof the guinea pig. Proc Soc Exp Biol Med 23: 661-663.

Hoffmann FG, Opazo JC. 2011. Evolution of the relaxin/insulin-like gene family in placental mammals: implications for its early evolution. J Mol Evol 72: 72-79.

Holland LZ, et al. 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res 18: 1100-1111.

Hu GB, Kusakabe M, Takei Y. 2011. Localization of diversified relaxin gene transcripts in the brain of eels. Gen Comp Endocrinol 172: 430-439.

Hughes AL, Friedman R. 2005. Loss of ancestral genes in the genomic evolution of Ciona intestinalis. Evol & Dev 7(3): 196-200.

Hughes LA, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167-170.

Huminiecki L, Heldin CH. 2010. 2R and remodeling of vertebrate signal transduction engine. BMC Biol 8: 146.

Hurst LD, Lercher MJ. 2005. Unusual linkage patterns of ligands and their cognate receptors indicate a novel reason for non-random gene order in the human genome. BMC Evol Biol 5: 62.

Ivell R, Grutzner F. 2009. Evolution and male fertility: lessons from the insulin-like factor 6 gene (Insl6). Endocrinology 150: 3986-3990.

Jaillon O, et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431: 946-957.

Kasahara M. 2007. The 2R hypothesis: an update. Curr Opin Immunol 19: 547-552.

Kasahara M, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719.

Kawamura K, et al. 2004. Paracrine regulation of mammalian oocyte maturation and male germ cell survival. Proc Natl Acad Sci U S A 101: 7323-7328.

Kemkemer C, Kohn M, Cooper DN, Froenicke L, Hogel J, Hameister H, Kehrer-Sawatzki H. 2009. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. BMC Evol Biol 9: 84.

Kinoshita M, Murata K, Naruse K, Tanaka M. 2009. Medaka: biology, management, and experimental protocols. Iowa: Wiley-Blackwell.

Kong RC, Shilling PJ, Lobb DK, Gooley PR, Bathgate RA. 2010. Membrane receptors: structure and function of the relaxin family peptide receptors. Mol Cell Endocrinol 320(1-2): 1-15.

Kuraku S, Meyer A, Kuratani S. 2009. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol Biol Evol 26: 47-59.

Liu C, et al. 2005. INSL5 is a high affinity specific agonist for GPCR142 (GPR100). J Biol Chem 280: 292-300.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494): 1151-1155.

McGowan BM, Stanley SA, Donovan J, Thompson EL, Patterson M, Semjonous NM, Gardiner JV, Murphy KG, Ghatei MA, Bloom SR. 2008. Relaxin-3 stimulates the hypothalamic-pituitary-gonadal axis. Am J Physiol Endocrinol Metab 295: E278-286.

McRory JE, Sherwood NM. 1997. Ancient divergence of insulin and insulin-like growth factor. DNA Cell Biol 16: 939-949.

Millar L, Streiner N, Webster L, Yamamoto S, Okabe R, Kawamata T, Shimoda J, Bullesbach E, Schwabe C, Bryant-Greenwood G. 2005. Early placental insulin-like protein (INSL4 or EPIL) in placental and fetal membrane growth. Biol Reprod 73: 695-702.

Mita M, Yamamoto K, Nagahama Y. 2011. Interaction of Relaxin-Like Gonad-Stimulating Substance with Ovarian Follicle Cells of the Starfish Asterina pectinifera. Zoolog Sci 28: 764-769.

Mita M, Yoshikuni M, Ohno K, Shibata Y, Paul-Prasanth B, Pitchayawasin S, Isobe M, Nagahama Y. 2009. A relaxin-like peptide purified from radial nerves induces oocyte maturation and ovulation in the starfish, Asterina pectinifera. Proc Natl Acad Sci U S A 106: 9507-9512.

Muffato M, Roest Crollius H. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays 30: 122-134.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17: 1254-1265.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York: Oxford University Press.

Nei M, Gu X, Sitnikova T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 94: 7799-7806.

Nordstrom KJ, Fredriksson R, Schioth HB. 2008. The amphioxus (Branchiostoma floridae) genome contains a highly diversified set of G protein-coupled receptors. BMC Evol Biol 8: 9.

Ohno S. 1970. Evolution by gene duplication. Berlin, New York: Springer-Verlag.

Olinski RP, Lundin LG, Hallbook F. 2006a. Conserved synteny between the Ciona genome and human paralogons identifies large duplication events in the molecular evolution of the insulin-relaxin gene family. Mol Biol Evol 23: 10-22.

Olinski RP, Dahlberg C, Thorndyke M, Hallbook F. 2006b. Three insulin-relaxin-like genes in Ciona intestinalis. Peptides 27: 2535-2546.

Park JI, Semyonov J, Chang CL, Yi W, Warren W, Hsu SY. 2008. Origin of INSL3-mediated testicular descent in therian mammals. Genome Res 18: 974-985.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25: 1795-1808.

Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198): 1064-1071.

Reinig JW, Daniel LN, Schwabe C, Gowan LK, Steinetz BG, O'Byrne EM. 1981. Isolation and characterization of relaxin from the sand tiger shark (Odontaspis taurus). Endocrinology 109: 537-543.

Samuel CS, Tian H, Zhao L, Amento EP. 2003. Relaxin is a key mediator of prostate growth and male reproductive tract development. Lab Invest 83: 1055-1067.

Sherwood OD. 2004. Relaxin's physiological roles and other diverse actions. Endocr Rev 25: 205-234.

Shiao MS, Liao BY, Long M, Yu HT. 2008. Adaptive evolution of the insulin two-gene system in mouse. Genetics 178: 1683-1691.

Sodergren E, et al. 2006. The genome of the sea urchin Strongylocentrotus purpuratus. Science 314: 941-952.

Steinetz BG, Schwabe C, Callard IP, Goldsmith LT. 1998. Dogfish shark (Squalus acanthias) testes contain a relaxin. J Androl 19: 110-115.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38: 615-643.

Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Gen 5: 366-375.

Wilkinson TN, Bathgate RA. 2007. The evolution of the relaxin peptide family and their receptors. Adv Exp Med Biol 612: 1-13.

Wilkinson TN, Speed TP, Tregear GW, Bathgate RA. 2005a. Evolution of the relaxin-like peptide family: from neuropeptide to reproduction. Ann N Y Acad Sci 1041: 530-533.

Wilkinson TN, Speed TP, Tregear GW, Bathgate RA. 2005b. Evolution of the relaxin-like peptide family. BMC Evol Biol 5: 14.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution 22(12): 2472-2479.

APPENDIX A: Using ancestral genome reconstructions to resurrect the duplication history of gene families

Multiple studies have been conducted in the last several years with the goal of understanding the evolution of genomes in the chordate lineage (Muffato and Roest Crollius 2008). I used the two most recent ancestral genome reconstruction models by Nakatani et al. (2007) and Putnam et al. (2008) (therein referred to as "N" and "P" model respectively) to clarify how the three rounds of whole genome duplications (1R, 2R and 3R) and subsequent genome rearrangements could have influenced the evolution of the *RLN/INSL* and *RXFP* families. In addition, I used the work by Kasahara et al. (Kasahara et al. 2007) to shed light on the effects of teleost-specific genome rearrangements on my genes of interest in medaka, tetraodon and zebrafish. I also referred to the reconstruction of the Eutherian ancestor genome to reconstruct the eutherian state (Kemkemer et al. 2009). Because I principally employ the Nakatani et al. (2007) model, and it includes two alternative scenarios for the genomic rearrangements that ensued between the pre-1R to the post 2R vertebrate genomes, in this appendix I also include the alternative scenarios for the gene duplication of my focal genes, which are not shown in main text.

The N-model reconstructs a later stage (compared to the P-model) in the evolution of chordate genome.

Although both the N- and P-models were constructed based on similar methodologies, the models differ in the number of ancestral chromosomes they predict and ultimately reconstruct two different ancestral genomes (Table A4). In particular, there is a significant difference in the conclusions made by each model about the pre-1R ancestor linkage groups: for example, the number of chordate linkage groups (*CLG*s, P-model) equals 17 while the number of vertebrate ancestral chromosomes (*VAC*s, N-model) is in the range of 10-13. The discrepancies between the two reconstructions can be explained by the inaccuracy of either or both models and by the evolutionary distance between the reconstructed genomes.

Putnam et al. (2008) compared vertebrate genomes to the genome of amphioxus to reconstruct the linkage groups *ancestral to both amphioxus and vertebrates*, or more accurately, olfactores (ancestor of tunicates and modern vertebrates). On the other hand, Nakatani et al. (2007) used protein-coding genes from *Ciona* and sea urchin to *outline* groups of paralogs in vertebrates without directly comparing the synteny between vertebrate and invertebrate genomes.

Overall, it is clear that the P-model reconstructs an earlier stage in the evolution of chordate karyotype (a "pre-1R protokaryotype") compared to the N-model, which shows a pre-1R genome that is structurally very close to its modern vertebrate counterpart. The evolutionary separation between the N- and P-model ("P") genomes should therefore be significant (Figure A1).

Given these assumptions, it can be hypothesized that the amphioxus-olfactores ancestral genome underwent several chromosomal fusions which led to a decrease in the number of chromosomes in the pre-1R vertebrate ancestor from 17 to 10-13 (See below and Figure

A4). Alternatively, the difference in the number of linkage groups may be attributable to the inaccuracy of one or both of the models.

How accurate are ancestral reconstructions?

Ancestral reconstructions, like any analyses indeed, are prone to errors. The accuracy of ancestral genome reconstruction is dependent on multiple factors among which the utilized methods and considered evolutionary scales are among the more prominent ones. Hence I sought for phylogenetic and small-scale synteny data confirmation for all results derived from the tracing of the history of the focal genes in this work.

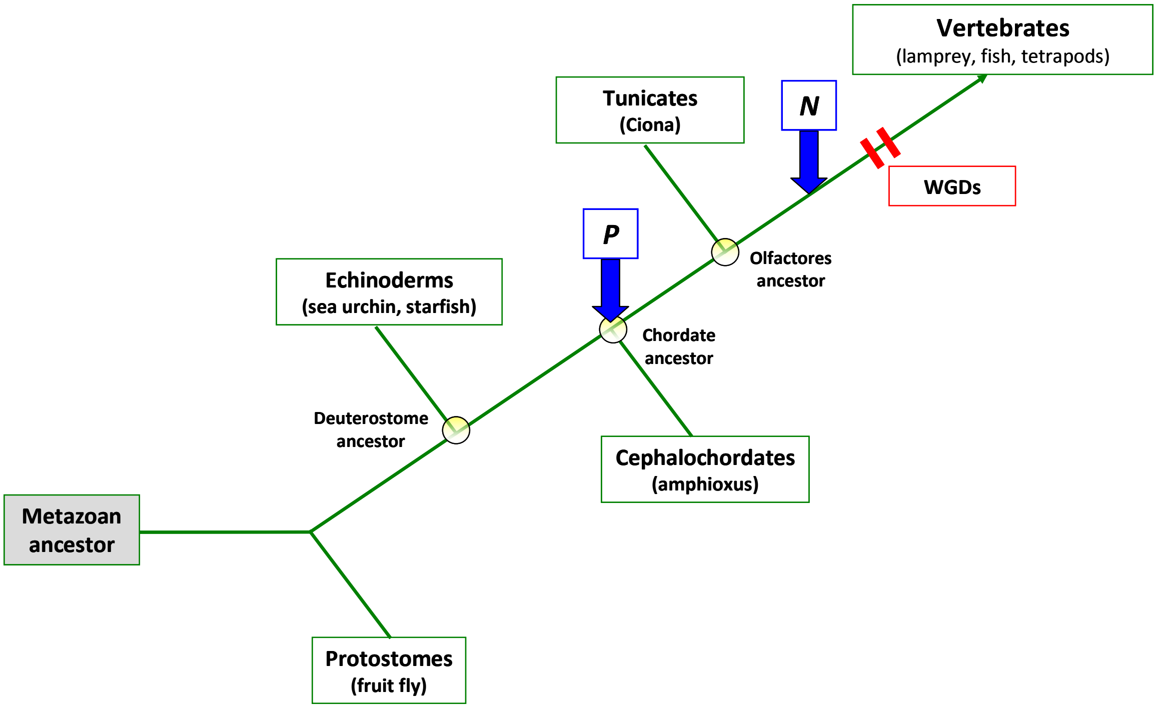


Figure A1. Simplified phylogenetic tree showing the evolutionary relationships among the groups of organisms discussed in this paper. The hypothetical ancestral genome predicted by the N-model ("N") probably belongs to an organism that existed just before 2R in early vertebrates. Tree topology adapted from Putnam et al (2008).

Tracing of the evolutionary history of genes in vertebrates using the N-model:

First, I mapped all medaka *rln/insl-rxfp* genes to ancestral pre-3R teleost chromosomes (Table A1: a-m). Each of the pre-3R teleost chromosomes as well as the human and chicken chromosomes can be inferred to be composed of *GAC*s (gnathostome ancestor chromosomes, e.g. *A0-A5, J0-J1*), which themselves arose from duplications of the ancestral vertebrate chromosomes *A-J*. This allows one to compare the sets of *GAC*s between human and medaka, and, given that the genomic location of the focal genes are known in human, chicken and the ancestor of medaka, it is then possible to trace the

chromosomal origins of the genes in the common ancestor of teleosts, human and chicken (osteichtyan ancestor).
Thus, secondly I determined which *GAC*s host each of the *RLN/INSL* and *RXFP* genes. I did this by comparing *GAC*s assigned to each of the genes in the human, medaka and chicken [Table A1: *GAC(H)*, *GAC(M)* and *GAC(C)*]and identifying the ones common to at least 2 of the analyzed genomes. For example, the comparison of the human, medaka and chicken *GAC*s for *RXFP3-1, RXFP3-3* and *RXFP3-4* led us to conclude that these genes originate from 3 post-2R *GAC*s (*A0, A4* and *A5*, respectively) (Table A1). This supported my conclusion about the ohnologous nature of *RXFP3-1*, *RXFP3-3* and *RXFP3-4,* which appear paralogous on the phylogenetic tree (Figure 1.5a).

Genes that exist in only one of the analyzed species were assigned to a *GAC* with the aid of other phylogenetic and syntenic data. For example, the *rxfp3-2* genes, which have been found in all studied teleosts, but have no traceable orthologs in human or chicken, were assigned to *GAC* "A1" using the following rationale. The medaka *rxfp3-2* gene belongs to the pre-3R chromosome "*m*", which is a mosaic of genes from 7 *GAC*s (*A1, A2, B0, B5, F0, J1* and *E1*) (Table A1). Due to absence of *GAC* data for this gene from human and chicken, it is not possible to deduce the GAC hosting *rxfp3-2* solely based on the information available for medaka. The phylogeny shows that the teleost *rxfp3-2* genes cluster together, in close proximity, to the *RXFP3-1* cluster (Figure 1.5a), suggesting that *RXFP3-1* and *3-2* are paralogs. Hence, the next step was to determine whether the teleost *rxfp3-2* gene was ohnologous to vertebrate *RXFP3/4* genes.
Although *RXFP3-2* has no tetrapod orthologs, its neighboring genes do have tetrapod orthologs, and the synteny of these neighboring genes allowed us to estimate the ancestral linkage of *RXFP3-2*. For example, medaka *rxfp3-2a* has two neighboring genes, *sirt6* (sirtuin 6, ENSORLG00000014983) and *eef2* (eukaryotic elongation factor-2, ENSORLG00000015009), and their chicken orthologs (ENSGALG00000001245 and ENSGALG00000001830) are found in chromosome 28 (see ENSEMBL genome browser). Since chicken chromosome 28 is syntenic only to *GAC "A1"* (Nakatani et al., 2007), we infer that *RXFP3-2* belongs to *GAC "A1"*. In addition, because the four *RXFP3/4* genes are mapped to 4 duplicated *GAC* chromosomes (*A0, A1, A4* and *A5*), I conclude that they are likely to be ohnologs.

An approach similar to the one described above was used to trace the ancestral origins of *INS/IGF* genes to clarify whether the relaxin and insulin/IGF genes were situated on one pre-1R *VAC* (vertebrate ancestral chromosome) and whether they arose from one ancestral pre-1R gene.

Two scenarios of the duplication and rearrangement history of VAC "A" (N-model)

In their work, Nakatani et al. (2008) proposed two scenarios for the duplication and rearrangement history of *VAC "A"*. According to one scenario (the "fission scenario", which I adopt as the framework for my analyses), a single chromosome in the pre-2R vertebrate ancestor is duplicated by 1R to produce two daughter chromosomes. One of these daughter chromosomes is further split into two linkage groups (one of them containing *AncRln-II* and the other- *AncRxfp3-II* in Figure 1.1). Hence before the onset of

2R, the post-1R vertebrate genome had a total of 3 *VAC "A"* descendants, which are duplicated by 2R to give rise to six post-2R chromosomes (*GAC "A0-A5"*).
According to the alternative scenario of *VAC "A"* evolution (the "fusion scenario", see Figure S2), the pre-2R vertebrate had two chromosomes (*VAC "A-I"* and *VAC "A-II"*), which after 1R yielded four post-1R linkage groups (*A-Ia/b* and *A-IIa/b* in Figure A2). Two of the post-1R chromosomes undergo fusion, which brings the total number of chromosomes down to 3, equaling the number of chromosomes at the onset of 2R described by the first scenario. Identical to the first scenario, 2R yields six *GAC* chromosomes (*GAC "A0-A5"*).

Essentially, the main conclusions (e.g. about the evolutionary relationships among *RLN/INSL* and *RXFP3/4* genes, their WGD-driven origination, and linkage of ancestral *RXFP3/4* and *RLN/INSL* genes) of this work are not altered by choosing either of the two scenarios. I adopt the "fission" scenario for the main text because it explains the derivation of my genes of interest using the fewest number of gene losses (compare: 4 losses in the fusion model with none in the fission model) and does not assume that an additional duplication took place in the proto-pre-2R vertebrate ancestor; i.e. the "fission" scenario is more parsimonious and is thus presented in the main text. The two major differences between the two scenarios are:

- The origination of modern *RLN/INSL* and *RXFP3/4* genes was primarily driven by the second round of WGD in the "fusion" scenario, whereas the "fission" scenario tells us that the origination of these genes was driven by both rounds of WGD.
- In the "fusion" scenario, two pairs of *RLN/INSL* ligand and *RXFP3/4* receptor genes existed in two separate linkage groups before the onset of 2R, whereas according to the "fission" scenario the two single ancestral genes were located in one linkage group in the pre-2R vertebrate ancestor.

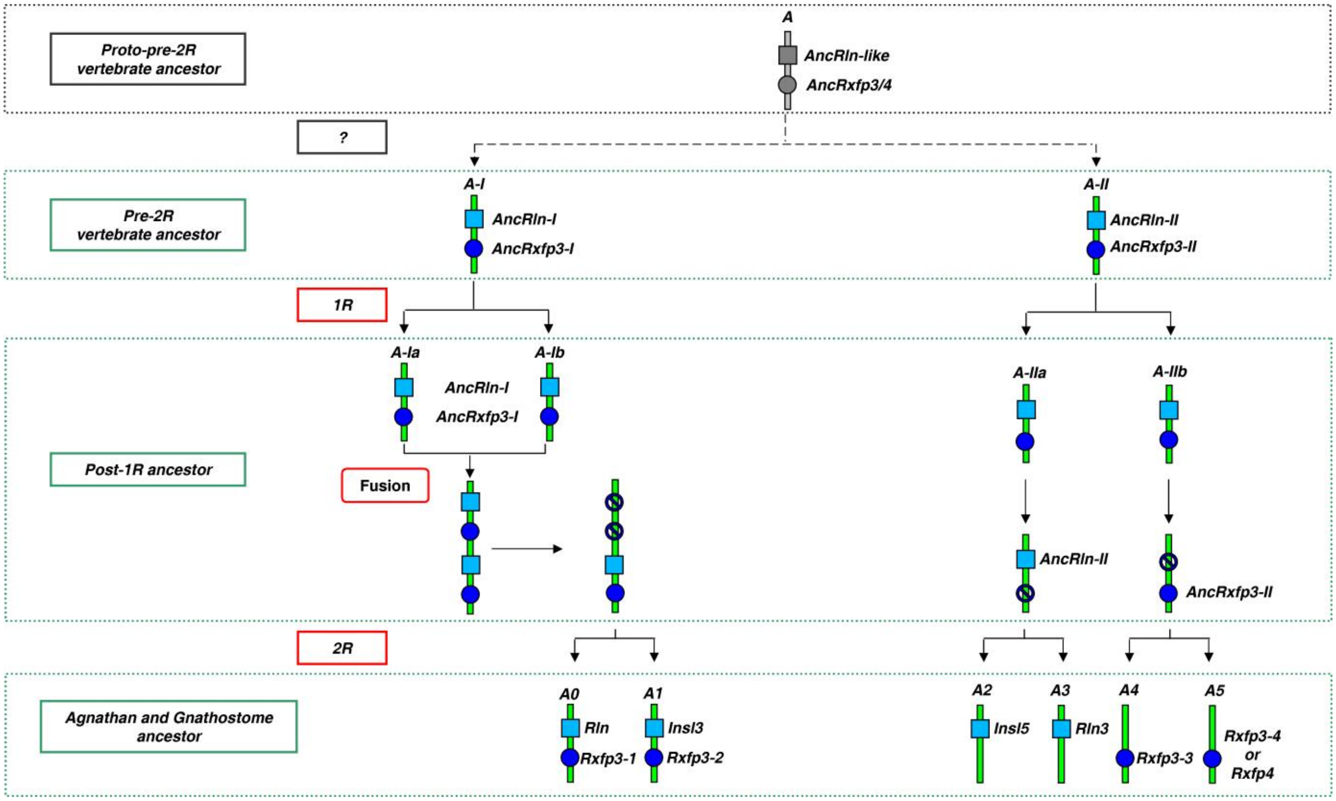


Figure A2. The alternative scenario of duplication and rearrangement history for *VAC "A"* according to the N-model. While the number and identity of post-2R daughter chromosomes in both scenarios (see Figure 1.1 for the scenario adopted here) is the same, the introduction of a *fusion* event (red box) and elimination of a *fission* event (red boxed in Figure 1.1) in this scenario results in two pre-2R chromosomes (*A-I* and *A-II*) each carrying a pair of genes (a ligand and a receptor). Hence, according to this scenario the pre-1R vertebrate already had two *Rln/Insl* ligand and two *Rxfp3/4* receptor genes (there is one of each according to the other scenario) and 1R did not play a major role in the duplication of the gene families. This scenario also implies that the twin ligand and receptor genes arose as a result of duplication in the earlier proto-pre-2R ancestor (gray boxed, question mark ("?") refers to the unknown duplication event).

The "fission" scenario is more supported by the sequence and phylogenetic data.

The phylogenetic reconstruction of RXFP3/4 genes (Figure 4a, main text) shows that the divergence of *AncRxfp3-I* from *AncRxfp3-II* occurs shortly before the divergence of Rxfp3-1/3-2 and Rxfp3-3/3-4. In the fission model, these duplication events are associated with 1R and 2R respectively, which are known to have occurred less than 50 MY apart from each other [6]. On the other hand, the fusion model would suggest that *AncRxfp3-I* and *AncRxfp3-II* were already present in the proto-pre-2R genome and had two separate, more ancient, origins.

My conclusions (N-model):

- Good-Avila et al. (2009) previously demonstrated that the *RLN/INSL* genes of teleosts and vertebrates are orthologous. Here I confirmed the synteny among the human, medaka and chicken genes (along with other vertebrate genes, see Appendix B), and by mapping them to the N-model I show that *RLN(2), RLN3, INSL3* and *INSL5* originated from one gene, which I call *AncRln-like,* in the pre-1R vertebrate ancestor and that they multiplied into four loci commensurate with the 2R events. Thus these 4 loci can be described as "ohnologs" based on their WGD-related evolutionary descent.
- According to the fission scenario (Figure 1.1), all four *RLN/INSL* genes arose as a result of 2R. After 1R, the *AncRln-like* gene duplicated giving rise, in the first instance, to the ancestor of the *RLN/INSL3* genes and, in the second instance, to the ancestor of the *RLN3/INSL5* genes. After 2R, these ancestral genes again duplicated giving rise to the 4 genes common to teleosts and tetrapods: *Rln, Insl3, Rln3* and *Insl5*.
- According to the fusion scenario (Figure A2), *RLN* and *INSL3* are 2R-ohnologs as are *RLN3* and *INSL5*, but *AncRln-I* and *AncRln-II* originate from an unknown duplication event in the proto-pre-2R ancestor.
- RXFP3 and RXFP4 receptors arose from one ancestral gene.
  - The fission scenario dictates that all *RXFP3/4* genes are 2R-ohnologs.

  [The fusion scenario implies that while *RXFP3-1* and *RXFP3-2* are ohnologs as are *RXFP3-3* and *RXFP3-4*, their parent genes, *AncRxfp3-I* and *AncRxfp3-II,* again arose in the proto-pre-2R ancestor as a result of a duplication event of an unknown nature.]
- Both *RLN/INSL* and *RXFP3/RXFP4* genes originated from one *VAC* named "A" by Nakatani et al. (2007) While *RLN(2)* and *INSL3* can be traced to the same gnathostome ancestor chromosomes (*GAC*s) as *RXFP3-1* and *RXFP3-2, RLN3, INSL5, RXFP3-3* and *RXFP4* are situated on different *GAC*s. According to the fission scenario, a logical explanation for this is that the pre-1R vertebrate ancestor had one *RLN3/INSL5*-like gene and one *RXFP3/RXFP4*-like gene which were linked on one chromosome. 2R duplicated the genes, but chromosomal rearrangements disrupted their linkage, thus the ligands and receptors were unlinked at the end of 2R.
- *RXFP1* and *RXFP2* are ohnologs.
- *RXFP1/2* and *RXFP2-like* originated from 2 *VAC*s that are different from that hosting *RXFP3/RXFP4* and *RLN/INSL* genes (*VAC "A"*). These chromosomes are known as "C" (*AncRxfp1/2*) and B or F (*AncRxfp2-like*). See main text for the discussion of the *Rxfp2-like* origins.
- Two different scenarios could explain the origin of *RXFP1/RXFP2*: these are shown in Figure A2.
- Although the tracing of the *INS/IGF* genes was problematic due to insufficient data available for medaka and other teleosts, these genes seem to have originated from an ancestral vertebrate chromosome *"D"* that is different from both *VAC* "*A"* and "*C"* that carried the ancestors of *RLN/INSL* and *RXFP* genes.

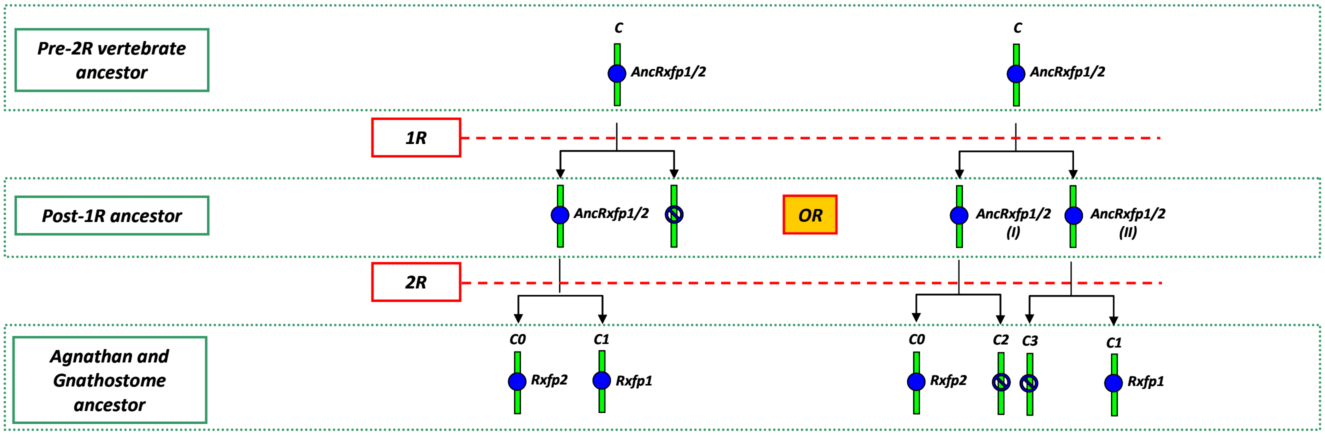


Figure A3. Two alternative scenarios for the 2R-driven duplication of the *AncRxfp1/2* gene. Note that based on the phylogenetic evidence (Figure 4b, main text), *RXFP2-like* is the paralog of *RXFP2*, in which case the two genes may have arisen as a result of 2R in *CLGs "C0"* and *"C2"*. The scenario on the right was hence adopted in the main text as more explanatory.

Search for the evidence of the presence of regions orthologous to *RLN/INSL* and *RXFP* loci in the amphioxus ancestor using the P-model:

Using their known genomic locations, each of the human *RLN/INSL* and *RXFP* genes were mapped to a chromosomal segment (Table A3: "Segment ID"). The identified chromosomal segments were then traced to *CLG*s using the oxford grid provided [3]. Additionally, the scaffold locations of amphioxus *ilp* and *rxfp1/2*-type genes were also traced, where possible, to *CLG*s using the oxford grid (Table A3). Since the oxford grid incorporates map locations from only two organisms, i.e. human and amphioxus, and because the identities of the amphioxus genes are still to be established, this method allowed us to use the genomic information pertaining only to the genes present in the human genome. In other words, the *CLG* origins of genes such as *Rxfp3-2* that have not been identified in humans (but exist in teleosts for example) could not be traced using this model.

My conclusions (P-model):

- All human *RLN/INSL* genes were traced to the same chordate linkage group (*CLG*), *CLG1*, agreeing with the N-model that all RLN family genes arose from a single ancestral gene.
- Only *RXFP3-1* was traced to *CLG1*, *RXFP3-3* was localized to *CLG2*, and the location of *RXFP4* is unclear.
- Both *RXFP1* and *RXFP2* were mapped to *CLG8*, while *RXFP2-like* was mapped to *CLG*9
- *INS* and *IGF2* were clearly mapped to a *CLG* different from those occupied by the *RLN/RXFP* genes confirming that the ancestral *INS/IGF2* and *RLN/INSL* have separate ancestral chromosome origins. Also one could conclude that the ancestral *INS/IGF2* genes were in a separate linkage group from ancestral *RLN/INSL* before the split of the amphioxus and olfactores lineages. (Following from this conclusion it is tempting to

revisit the identities of the three *INS/IGF/RLN*-like genes previously identified in *C. intestinalis* as linked on one chromosome [7]).

- Some of the amphioxus candidate *rln/insl*, *ins/igf* and *rxfp1/2* genes that were obtained from public databases (Appendix B) were assigned to the same *CLG*s as their human counterparts (the *ins/igf-like* and *rxfp1/2-like* groups). I was unable to identify any *rxfp3/4-like* genes in the amphioxus databases.

Gene gain/loss and genomic rearrangements in the pre-2R ancestor and/or inaccuracy of ancestral genome reconstruction models may account for the difference in the results obtained using the two models:

According to the results of the gene tracing method using the P-model, *RXFP3* and *RXFP4*-type genes originate from at least 2 different *CLG*s and only one of them, *RXFP3-1*, appears to have been linked to the ancestral *RLN/INSL* gene on *CLG1*. This would suggest that *RXFP4* has a different evolutionary origin from *RXFP3*. On the contrary, the N-model gene tracing method predicts that all *RXFP4* and *RXFP3*-type genes originated from one ancestral receptor gene that was linked to the ancestral *RLN/INSL* gene (*VAC "A"*, as described above).
How can this conflict be explained?
As discussed above, the ancestor linkage groups reconstructed in the P- and N-models are not equivalent. It is possible that some of the *CLG*s of the amphioxus-olfactores ancestor fused to produce "multi-*CLG*" chromosomes of the vertebrate ancestor. For instance, *CLG1*, *CLG2* and could have fused together and with other unknown *CLG*s, resulting in the so-called *VAC "A"* reconstructed by Nakatani et al. (2007). Intriguingly, amphioxus does not seem to possess *rxfp3/4*-type genes which implies that these genes appeared after the divergence of cephalochordates.
Alternatively the observed discrepancy could stem from inaccurate ancestral genome reconstruction.

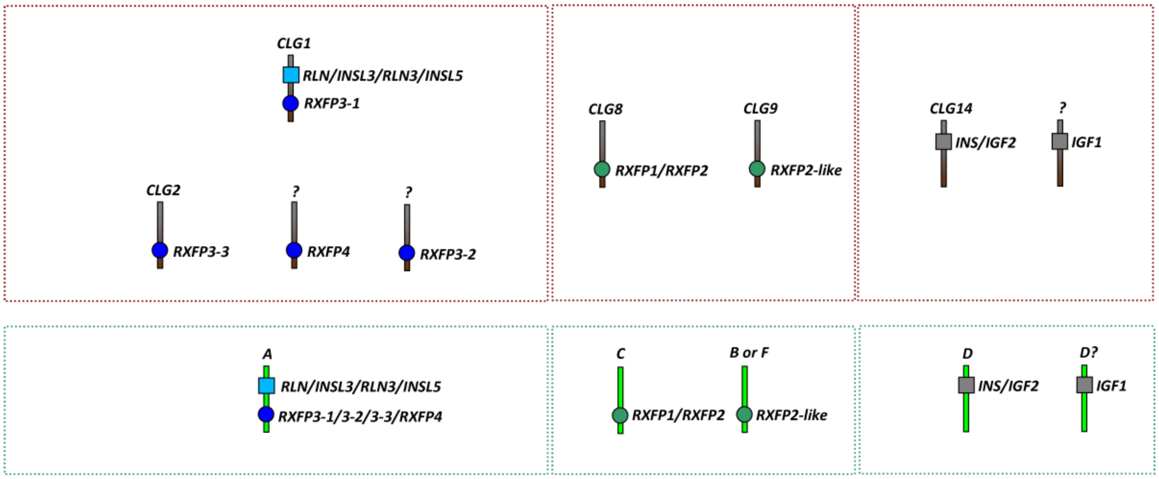


Figure A4. Comparison of the results obtained using two ancestral genome reconstructions **top:** Tracing of human *RLN/INSL* and *RXFP*-like genes in chordate linkage groups (*CLG*) using P-model; **bottom**: Tracing of human *RLN/INSL* and *RXFP* genes in pre-2R vertebrate ancestor chromosomes.

Origins of the Rxfp-Rln/Insl system in early Deuterostomes

To further explore the genomic background of insulin-related genes in invertebrate chordates I employed the synteny tool in VISTA-Point (http://pipeline.lbl.gov/cgi-bin/gateway2?bg=Brafl1&selector=vistapoint) and compared the four Amphioxus scaffolds containing Rln/Insl-like genes to the human genome.
Each scaffold shows different levels of synteny to various regions of different human chromosomes. By looking at specific regions of human chromosomes and using Nakatani et al.'s work I determined the relationship of the amphioxus scaffolds to the pre-2R vertebrate ancestor chromosomes. I find that the two scaffolds predicted by the P-model to have originated from *CLG*14 do indeed share a significant amount of synteny with *GAC*s A1, A3, D0-D3. The other two scaffolds containing a total of three Rln/Insl-like genes are distinct and exhibit synteny to different regions of the human genome (Figure A4). Probably due to its small size, the syntenic information available for scaffold 372 is not sufficient to determine its origins.

Table A1. Dataset used to reconstruct the duplication history of *RLN/INSL-RXFP* genes using model "N"

Map positions of *RLN/INSL, RXFP* and *INS/IGF* genes in human (H), medaka (M) and chicken (C) are shown. The pre-3R teleost ancestor chromosomes that gave rise to the *rln/rxfp* medaka chromosomes are shown in column 3 (pre3R)
CVLb (H): Conserved Vertebrate Linkage blocks in human, determined using the known intrachromosomal position of human genes and the protochromosome map scheme from Nakatani et al. (2007)
GAC: Gnathostome Ancestor Chromosomes, reconstructed by Nakatani et al. (2007)
GAC (H): GACs potentially hosting the genes of interest based on the identified CVLb (H)
GAC (C): GACs hosting *RLN/INSL-RXFP*genes in chicken
GAC (R): GAC Resolved to contain *RLN/INSL-RXFP* genes in the gnathostome ancestor
GAC (M): GACs hosting the genes of interest based on the comparison between medaka and human
CVLb (R): CVLb resolved to contain *Rln/Insl-Rxfp* genes in the gnathostome ancestor

	Map positions								**Ancestral linkage groups**		
Gene	**H**	**C**	**M***	**pre-3R**	**CVLb (H)****	**GAC (H)**	**GAC (M)**	**GAC (C)**	**CVLb (R)****	**GAC (R)**	**VAC**
RLN (2)	9: 5.30	Z: 28.2	12: 14.80	*i*	50	*A0*	*A0*	*A0*	50	*A0*	***A***
INSL3	19: 17.93	n/a	17: 60.74	*m*	89-90	*A3-A1*	*A1*	n/a	90	*A1*	
INSL5	1: 67.26	8: 29.36	4: 16.56; 17: 16.79	*m*	2	*A2*	*A2*	*A2*	2	*A2*	
RLN3	19: 14.14	n/a	1: 31.36; 8: 30.65	*f-e-d*	89	*A3*	*A3 (e)*	n/a	89	*A3*	
RXFP3-1	5: 33.93	Z: 9.7	12: 10.00	*i*	27	*A0*	*A0*	*A0*	27	*A0*	***A***
RXFP3-2	n/a	n/a	17: 22.92; 4: 8.24	*m; m*	n/a	n/a	*A1-A2-B0-B5-F0-J1-E1*	n/a	88-90-111	*A1*	
RXFP3-3	15: 79.12†	n/a	3: 10.20-10.95; 6:?	*j; j-k*	76	*A4*	*A4 (j)*	n/a	76	*A4*	
RXFP4	1: 155.91	n/a	11: 6.39	*b*	4-5-109	*A5-A2-Un*	*A5*	n/a	4	*A5*	
RXFP1	4: 159.24	4: 23.0	10: ?	*f-g*	25	*C1*	*C1*	*C1*	25	*C1*	***C***
RXFP2	13: 32.31	1: 179.1†	14: 9.82	*g-h*	71	*C0*	*C0*	*C0*	71	*C0*	
RXFP2-like	X: 67.94†	4: 0.57	n/a	n/a	108-117	*B0-F4*	n/a	*C1-B0-F4*	108-117	*B0-F4*	***B/F***

Table A2. Dataset used to reconstruct the duplication history of *Ins/Igf* genes using model "N"

Tetrapods: Human and Chicken						**Ancestral linkage groups**	
Gene	**H**	**C**	**CVLb (H)**	**GAC (H)**	**GAC (C)**	**CVLb (R)**	**GAC (R)**
INS1	11: 2.153	5: 14.85	60	D1	D1	60	*D1*
IGF2	11: 2.150	5: 14.88	60	D1	D1	60	*D1*
IGF1	12: 102.79	1: 57.33	68	D0	D0	68	*D0*

Fish: Medaka (M) and Tetraodon (T)

	T	**pre-3R (T)**	**M^**	**pre-3R (M)**	**pre-3R M+T*****	**GAC (R)**
ins1	1: 3.38	*f/g/m*	21: ?	*c?*	*g*	**B0/C2/F4**
ins2	7: 10.6	*g/h*	14: ?	*g-h?*	*g*	**B0/C2/F4**
igf1	19:1.39	*k*	23: 20.44	*k*	*k*	**D0**
igf2	13: 6.6	*j/k*	n/a	*?*	*k*	**D0**

* two positions are given for 3R duplicates located on two separate modern medaka chromosomes originating from one pre3R teleost ancestor chromosome

** when multiple blocks match the same location/GAC, all possible block numbers are given separated by dashes (e.g. **#-#-#**)

† pseudogene

^ not all genome regions hosting Ins/Igf genes in medaka are assembled into chromosomes, therefore mapping of these genes is imprecise

*** based on shared ancestral teleost linkage groups, since all Tetraodon genes can be mapped to distinct chromosomes, results obtained for Tetraodon overrode Medaka
Genomic region around Tetraodon Ins1 is syntenic to the region of Human Chromosome 5 under CVLb 28, which is equivalent to C2

Table A3. Dataset used to reconstruct the origins of *Rln/Insl-Rxfp* genes in the chordate ancestor using model "P"

Human genes were mapped to genome segments determined by Putnam *et al*., amphioxus candidate gene locations (scaffolds) were also assigned to CLGs based on the Oxford grid in Putnam *et al*.

Segment ID: Human Genome Segment boundaries corresponding to genomic locations as described in Putnam *et al.* [2]

CLG: Chordate Linkage Group, Putnam *et al.* identified 17 CLGs which correspond to the hypothetical chromosomes of the ancestor of amphioxus and *Olfactores* (tunicates and vertebrates)

Human genes			
Gene	**Location**	**Segment ID**	**CLG**
RLN (2)	9: 5.30	9.1	1
INSL3	19: 17.93	19.3	1
INSL5	1: 67.26	1.5	1
RLN3	19: 14.14	19.2	1
RXFP3-1	5: 33.93	5.1	1
RXFP3-3	15: 79.12†	15.2	2
RXFP4	1: 155.91	1.13	un*
RXFP1	4: 159.24	4.4	8
RXFP2	13: 32.31	13.1	8
RXFP2-lik	X: 67.94†	X.5	9
INS	11: 2.153	11.1	14
IGF2	11: 2.150	11.1	14
IGF1	12: 102.79	12.9	un*

*segment not mapped to a CLG

Amphioxus rln/insl/rxfp and ins/igf candidates			
	Gene ID	**Scaffold**	**CLG**
ilp4	7253642	bf_v2_277	un*
ilp5	7255900	bf_v2_277	un*
ilp6	7230317	bf_v2_196	4
ilp2	***n/a***	bf_V2_190	14
ilp1	7235917	bf_V2_190	14
ilp3	7251652	bf_v2_243	14
rxfp1/2-like	7252026	bf_v2_249	7-8
	7229038	bf_v2_150	8
	7221608	bf_v2_150	8
	7207790	bf_v2_21	un*
	7209355	bf_v2_21	un*

Table A4. Comparison of the two reconstruction models used in this study.

		N-model	**P-model**
Brief outline of The procedure used		Used the *C. intestinalis* and *S. purpuratus* genomes to outline ohnologs in human, mouse, dog, and pufferfishes (tetraodon and fugu); grouped ohnologs into GACs Compared human versus medaka and pufferfishes to determine the effect of WGD3 and subsequent rearrangements on the teleost chromosome evolution. Chicken genome was employed to verify the correctness of the reconstructed amniote ancestral genome	Compared the paralogons of human, chicken, stickleback, fugu and amphioxus Subdivided vertebrate chromosomes into segments based on the identified paralogons Employed the information from the amphioxus genome to directly reconstruct the chordate linkage groups
Predicted numbers of chromosomes in the ancestral genomes			
Pre-1R ancestors	Olfactores-amphioxus	n/a	17
	Vertebrate	10-13	n/a
Post-2R ancestors	Gnathostome	40	69
	Osteichtyan	31	37-49
	Amniote	26	?
Pre-3R ancestor	Teleost	13	12

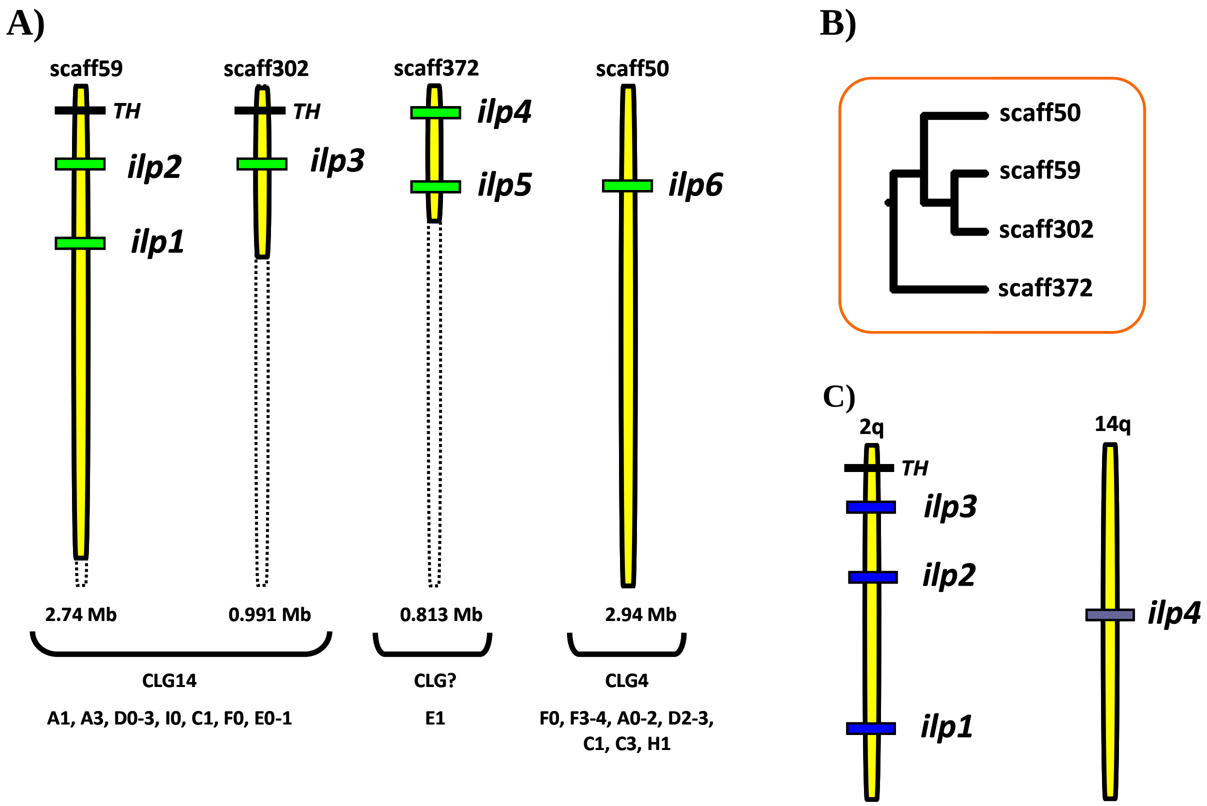


Figure A5. A) Four amphioxus scaffolds hosting *ilp* genes. Phylogenetically *ilp1* is the closest to the vertebrate *INS* gene, *ilp5* structurally resembles vertebrate *INSL5* and *ilp4* is found outside the vertebrate *RLN/INSL* clade (see *Figure 1.4*). Interestingly, both *ilp2* and *ilp3*, which seem to have originated from a duplication of *ilp1*, appear to be orthologous to the starfish relaxin-like gene (GSS). Overall, amphioxus seems to possess a set of genes which represents a continuity of evolution from the more conserved *INS/IGF*-type gene to more divergent *RLN/INSL*-type genes. Length of scaffolds shown in Mb. **B)** Phylogenetic tree is based on global alignment done in MLAGAN (VISTA-Point), it confirms the common origin of scaffolds 59 and 302 (*CLG14*); **C)** Genomic location of *ilp* genes in *C.intestinalis*. *ilp4* is a novel gene for the first time identified in this study. Olinski et al. (2006) proposed that *ilp2* and *ilp3* are orthologous to the vertebrate *INS/IGF* locus (based on the close linkage of these gene pairs), while *ilp1* is the ortholog of vertebrate *RLN/INSL*. Note that phylogenetically both *ilp1* and *ilp2* are close to the amphioxus *ilp1* gene, and *ilp3*-*ilp4* cluster with a fruifly "relaxin-like" gene possbly due to long-branch attraction (*Figure 1.4,* main text). Taking into account that another tunicate (*C. productum*) possesses 2 *ilp* genes which are phylogenetically close to amphioxus *ilp1*, it is more likely that the Ciona *ilp* genes are highly divergent duplicates of the amphioxus *ilp1*-like gene.
TH: tyrosine hydroxylase gene, typically found next to the vertebrate *INS/IGF* loci.

References
(Appendix A)

Good-Avila SV, Yegorov S, Harron S, Bogerd J, Glen P, Ozon J, Wilson BC. 2009. Relaxin gene family in teleosts: phylogeny, syntenic mapping, selective constraint, and expression analysis. BMC Evol Biol 9: 293.

Kasahara M, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719.

Kemkemer C, Kohn M, Cooper DN, Froenicke L, Hogel J, Hameister H, Kehrer-Sawatzki H. 2009. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. BMC Evol Biol 9: 84.

Muffato M, Roest Crollius H. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays 30: 122-134.

Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198): 1064-1071.

APPENDIX B: Database IDs, genomic locations and other information pertaining to the genes used in the study

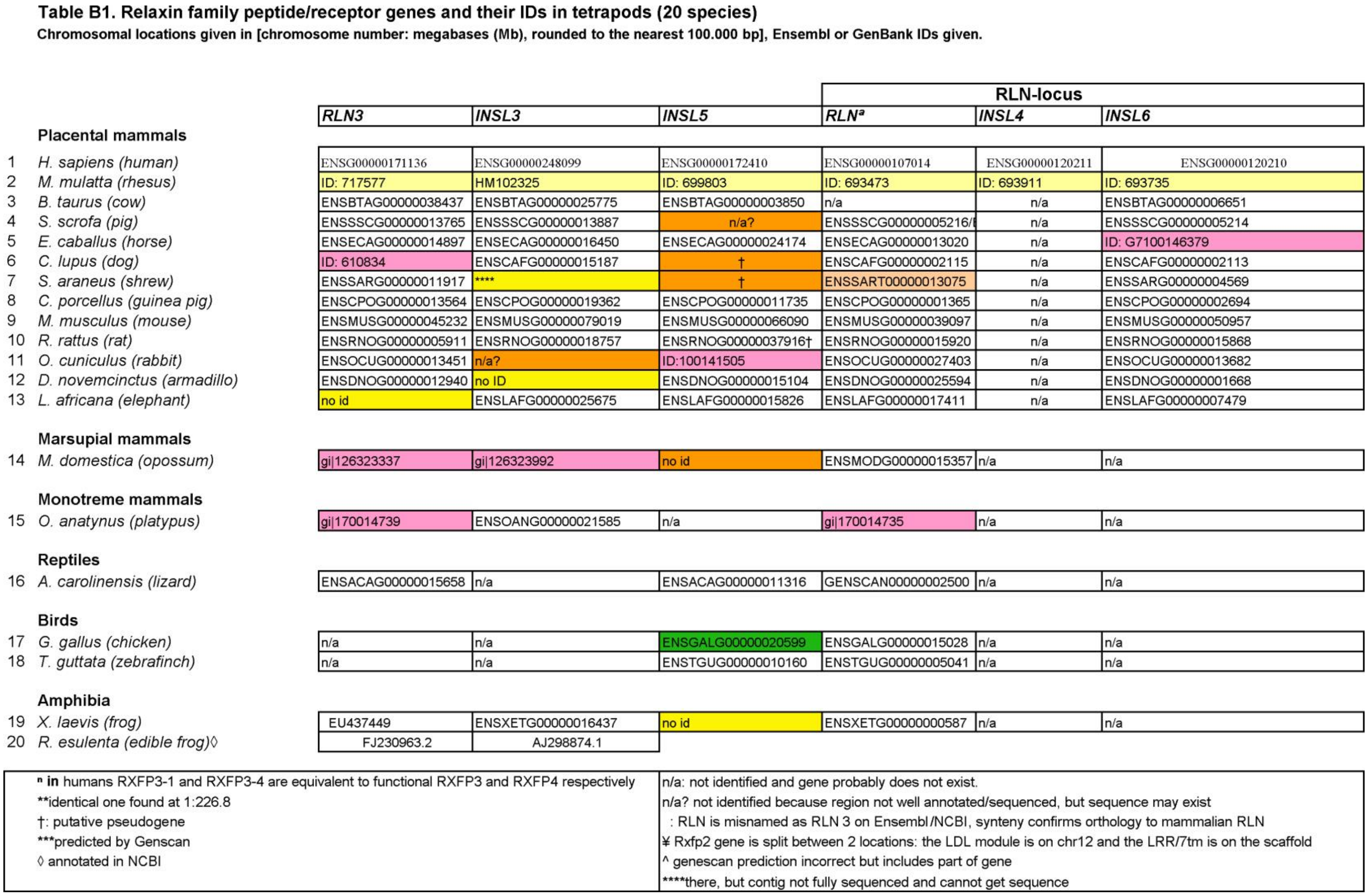


Table B1. Relaxin family peptide/receptor genes and their IDs in tetrapods (20 species)

Chromosomal locations given in [chromosome number: megabases (Mb), rounded to the nearest 100.000 bp], Ensembl or GenBank IDs given.

					RLN-locus		
		RLN3	***INSL3***	***INSL5***	***RLNa***	***INSL4***	***INSL6***
	Placental mammals						
1	*H. sapiens (human)*	ENSG00000171136	ENSG00000248099	ENSG00000172410	ENSG00000107014	ENSG00000120211	ENSG00000120210
2	*M. mulatta (rhesus)*	ID: 717577	HM102325	ID: 699803	ID: 693473	ID: 693911	ID: 693735
3	*B. taurus (cow)*	ENSBTAG00000038437	ENSBTAG00000025775	ENSBTAG00000003850	n/a	n/a	ENSBTAG00000006651
4	*S. scrofa (pig)*	ENSSSCG00000013765	ENSSSCG00000013887	n/a?	ENSSSCG00000005216/I	n/a	ENSSSCG00000005214
5	*E. caballus (horse)*	ENSECAG00000014897	ENSECAG00000016450	ENSECAG00000024174	ENSECAG00000013020	n/a	ID: G7100146379
6	*C. lupus (dog)*	ID: 610834	ENSCAFG00000015187	†	ENSCAFG00000002115	n/a	ENSCAFG00000002113
7	*S. araneus (shrew)*	ENSSARG00000011917	****	†	ENSSART00000013075	n/a	ENSSARG00000004569
8	*C. porcellus (guinea pig)*	ENSCPOG00000013564	ENSCPOG00000019362	ENSCPOG00000011735	ENSCPOG00000001365	n/a	ENSCPOG00000002694
9	*M. musculus (mouse)*	ENSMUSG00000045232	ENSMUSG00000079019	ENSMUSG00000066090	ENSMUSG00000039097	n/a	ENSMUSG00000050957
10	*R. rattus (rat)*	ENSRNOG00000005911	ENSRNOG00000018757	ENSRNOG00000037916†	ENSRNOG00000015920	n/a	ENSRNOG00000015868
11	*O. cuniculus (rabbit)*	ENSOCUG00000013451	n/a?	ID:100141505	ENSOCUG00000027403	n/a	ENSOCUG00000013682
12	*D. novemcinctus (armadillo)*	ENSDNOG00000012940	no ID	ENSDNOG00000015104	ENSDNOG00000025594	n/a	ENSDNOG00000001668
13	*L. africana (elephant)*	no id	ENSLAFG00000025675	ENSLAFG00000015826	ENSLAFG00000017411	n/a	ENSLAFG00000007479
	Marsupial mammals						
14	*M. domestica (opossum)*	gi|126323337	gi|126323992	no id	ENSMODG00000015357	n/a	n/a
	Monotreme mammals						
15	*O. anatynus (platypus)*	gi|170014739	ENSOANG00000021585	n/a	gi|170014735	n/a	n/a
	Reptiles						
16	*A. carolinensis (lizard)*	ENSACAG00000015658	n/a	ENSACAG00000011316	GENSCAN00000002500	n/a	n/a
	Birds						
17	*G. gallus (chicken)*	n/a	n/a	ENSGALG00000020599	ENSGALG00000015028	n/a	n/a
18	*T. guttata (zebrafinch)*	n/a	n/a	ENSTGUG00000010160	ENSTGUG00000005041	n/a	n/a
	Amphibia						
19	*X. laevis (frog)*	EU437449	ENSXETG00000016437	no id	ENSXETG00000000587	n/a	n/a
20	*R. esulenta (edible frog)◊*	FJ230963.2	AJ298874.1				

ⁿ in humans RXFP3-1 and RXFP3-4 are equivalent to functional RXFP3 and RXFP4 respectively	n/a: not identified and gene probably does not exist.
**identical one found at 1:226.8	n/a? not identified because region not well annotated/sequenced, but sequence may exist
†: putative pseudogene	: RLN is misnamed as RLN 3 on Ensembl /NCBI, synteny confirms orthology to mammalian RLN
***predicted by Genscan	¥ Rxfp2 gene is split between 2 locations: the LDL module is on chr12 and the LRR/7tm is on the scaffold
◊ annotated in NCBI	^ genescan prediction incorrect but includes part of gene
	****there, but contig not fully sequenced and cannot get sequence

🟩 gene name needs to be updated on ensembl
🟧 not found or a pseudogene
🟨 found but not annotated
▨ NCBI

		RXFP1	***RXFP2***	***RXFP2-like***	***RXFP3-1^{n}***	***RXFP3-3***	***RXFP4***
	Placental mammals						
1	*H. sapiens (human)*	ENSG00000171509	ENSG00000133105	†	ENSG00000182631	GENSCAN00000038299^	ENSG00000173080
2	*M. mulatta (rhesus)*	ID: 701107	ID: 721969	†?	ID: 698115	ID: 100426387†	ID: 718025
3	*B. taurus (cow)*	ENSBTAG00000010306	ENSBTAG00000015132	†?	ENSBTAG00000039929	ENSBTAG00000026976	ID: 450212
4	*S. scrofa (pig)*	ENSSSCG00000008875	ENSSSCG00000009336	†?	ENSSSCG00000016820	ENSSSCG00000001768	ENSSSCG00000006503
5	*E. caballus (horse)*	ENSECAG00000013594	ENSECAG00000014103	†?	ENSECAG00000012797	†?	ENSECAG00000007292
6	*C. lupus (dog)*	ENSCAFG00000008672	ENSCAFG00000006501	†?	ID: 489237	†?	†
7	*S. araneus (shrew)*	ENSSARG00000013665	ENSSARG00000009883	†?	ENSSARG00000000381	†?	n/a
8	*C. porcellus (guinea*	ENSCPOG00000015517	ENSCPOG00000009157	†?	ENSCPOG00000009428	†?	ENSCPOG00000003454
9	*M. musculus (mouse*	ENSMUSG00000034009	ENSMUSG00000053368	†?	ENSMUSG00000060735	†?	ENSMUSG00000049741
10	*R. rattus (rat)*	ENSRNOG00000024120	ENSRNOG00000000897	†?	ENSRNOG00000023126	†?	†
11	*O. cuniculus (rabbit)*	ENSOCUG00000001208	ENSOCUG00000000751	†?	ENSOCUG00000004946	†?	ENSOCUG00000026819
12	*D. novemcinctus (arn*	ENSDNOG00000001016	ENSDNOG00000018859	†?	ENSDNOG00000012031	†?	ENSDNOG00000017055
13	*L. africana (elephant)*	ENSLAFG00000014435	ENSLAFG00000012550	†?	ENSLAFG00000017082	†?	ENSLAFG00000011205
	Marsupial mammals						
14	*M. domestica (oposs*	ENSMODG00000001973	ENSMODG00000009382	ENSMODG00000012676	ENSMODG00000020402	ENSMODG00000025367	ENSMODG00000024291
	Monotreme mammals						
15	*O. anatynus (platypu*	ENSOANG00000005387	ENSOANG00000015476		ENSOANG00000001969	†?	n/a
	Reptiles						
16	*A. carolinensis (lizard*	ENSACAG00000016552	n/a	ENSACAG00000007727	ENSACAG00000008321	n/a	n/a
	Birds						
17	*G. gallus (chicken)*	ENSGALG00000009429	n/a	ENSGALG00000004543	ENSGALG00000017411	n/a	n/a
18	*T. guttata (zebrafinch*	ENSTGUG00000005573	n/a	n/a	ENSTGUG00000001946	n/a	n/a
	Amphibia						
19	*X. laevis (frog)*	ENSXETG00000019493	ENSXETG00000019186	ENSXETG00000011511	ENSXETG00000011511	n/a	ENSXETG00000001632
20	*R. esulenta (edible frog)*◊						

Table B2. Mammalian lineage-specific duplicates in the *RLN*-locus

Human	**ID**
RLN1	ENSG00000107018
RLN2	ENSG00000107014
Rabbit	
Relaxin-like protein SQ10 Precursor	ENSOCUG00000027934
Relaxin-like protein SQ10 Precursor	ENSOCUG00000027403
unnamed	ENSOCUG00000026316
unnamed	ENSOCUG00000008099
unnamed	ENSOCUG00000008103
unnamed	ENSOCUG00000025640
Shrew	
unnamed	ENSSARG00000013075
unnamed	ENSSARG00000004245
unnamed	ENSSARG00000010511
Armadillo	
rln-a	ENSDNOG00000025594
unnamed	ENSDNOG00000025720
Pig	
Prorelaxin Precursor	ENSSSCG00000005216
Relaxin R-II1 A chain	ENSSSCG00000005213

Table B3. Relaxin family peptide/receptor genes and their chromosomal locations in tetrapods (20 species)

<u>Chromosomal locations given in [chromosome number: megabases (Mb), rounded to the nearest 100.000 bp], all locations are from Ensembl, unless otherwise specifie</u>

					RLN-locus		
		RLN3	***INSL3***	***INSL5***	***RLN***	***INSL4***	***INSL6***
	Placental mammals						
1	*H. sapiens (human)*	19: 14.14	19: 17.93	1: 67.26	9: 5.30	9: 5.23	9: 5.16
2	*M. mulatta (rhesus)*	19: 13.71	19: 17.4	1: 69.58	15: 71.83	15: 71.92	15: 72.00
3	*B. taurus (cow)*	7: 9.87	7: 5.26	3: 84.21	n/a	n/a	8: 41.52
4	*S. scrofa (pig)*	2:57.0	2:62.1	6*	1:226.8**	n/a	1:226.74
5	*E. caballus (horse)*	7:44.8	21:2.6	5:94.1	23:26.7	n/a	23:26.6
6	*C. lupus (dog)*	20:51.5***	20:48.1	5:46.7†	1: 96.6	n/a	1: 96.5
7	*S. araneus (shrew)*	sc79318: 0.0006:	sc_1461:1032	†?	scaffold_70875: 4	n/a	sc2354: 0.012
8	*C. porcellus (guinea pig)*	sc42: 12.8	sc72: 7.8	sc2: 66.3	sc21: 14.7	n/a	sc21: 14.7
9	*M. musculus (mouse)*	8: 86.6	8: 74.2	4: 102.7	19: 29.4	n/a	19:29.4
10	*R. rattus (rat)*	19: 25.8	16: 18.9	5: 123.9†	1: 233.0	n/a	1: 233.0
11	*O. cuniculus (rabbit)*	sc1049: 0.028	n/a?	13:99.4	chr1	n/a	chr1
12	*D. novemcinctus (armadillo)*	sc236409: 0.0001	sc31772: 0.017	sc2247: 0.3	sc1575: 0.061	n/a	sc119039: 0.003
13	*L. africana (elephant)*	sc26: 29.1	sc26: 24.2	sc17: 27.1	sc6: 93.9	n/a	sc88: 0.58
	Marsupial mammals						
14	*M. domestica (opossum)*	3: 446.3	3: 476.4	2: 24.49	6: 166.9	n/a	n/a
	Monotreme mammals						
15	*O. anatynus (platypus)*	ultc605:0.02	c19353: 0.007	n/a?	X5: 11.5	n/a	n/a
	Reptiles						
16	*A. carolinensis (lizard)*	sc132: 2.7	n/a?	sc619: 0.48	sc13: 3.2	n/a	n/a
	Birds						
17	*G. gallus (chicken)*	n/a?	n/a?	8: 29.36	Z: 28.2	n/a	n/a
18	*T. guttata (zebrafinch)*	n/a?	n/a?	8: 26.7	Z: 63.9	n/a	n/a
	Amphibia						
19	*X. laevis (frog)*	sc649: 0.087	sc969: 0.20	sc431: 0.98	sc86: 2.1	n/a	n/a
20	*R. esulenta (edible frog)*◊	FJ230963.2	AJ298874.1				

ⁿ **in** humans RXFP3-1 and RXFP3-4 are equivalent to functional RXFP3 and RXFP4 respectively

* not annotated along with syntenic genes, but probably exists

**identical one found at 1:226.8

†: putative pseudogene: rat insl5 is a processed pseudogene

***predicted by Genscan

◊ annotated on NCBI

n/a?: not identified, but likely exists

: RLN is misnamed as RLN3 on Ensembl/NCBI, synteny confirms orthology to mammalian RLN

¥ Rxfp2 gene is split between 2 locations: the LDL module is on chr12 and the LRR/7tm is on the scaffold

	Placental mammals	***RXFP1***	***RXFP2***	***RXFP2-like***	***RXFP3-1***^{n}	***RXFP3-3***	***RXFP3-4***^{n}
1	*H. sapiens (human)*						
2	*M. mulatta (rhesus)*	4: 159.24	13: 32.31	X: 67.94†	5: 33.93	15: 79.12†	1: 155.91
3	*B. taurus (cow)*	5: 15.07	17: 11.10	†?	6: 34.04	7: 57.95†	1: 13.45
4	*S. scrofa (pig)*	17: 42.34	12: 29.01	†?	20: 42.31	21: 30.35	3: 16.15
5	*E. caballus (horse)*	8:41.1	11:7.7	†?	16:17.5	7: 54.0	4: 98.0
6	*C. lupus (dog)*	2:75.8	17:11.0	†?	21:30.7	1: 115.6†	5: 42.2
7	*S. araneus (shrew)*	15: 58.7	25: 11.3	†?	4: 77.1◊	†?	7: 44.8†
8	*C. porcellus (guinea pig)*	sc5672: 0.004	sc3003: 0.1	†?	sc4216: 0.008	†?	n/a?
9	*M. musculus (mouse)*	sc7: 31.8	sc6: 33.2	†?	sc29:21.7	†?	10:16.6
10	*R. rattus (rat)*	3: 79.5	5: 150.8	†?	15: 10.96	†?	3: 88.5
11	*O. cuniculus (rabbit)*	2: 171.1	12: 5.3	†?	2: 60.4	†?	2: 180.8†
12	*D. novemcinctus (armadi*	sc58: 1.2	sc59: 2.4	†?	11: 55.7	†?	13:36.9
13	*L. africana (elephant)*	sc1259: 0.003	sc3001: 0.024	†?	sc6302:0.04†	†?	sc4957:0.006
		sc61: 9.6	sc11: 5.3	†?	sc7: 37.9	†?	sc33: 4.4
	Marsupial mammals						
14	*M. domestica (opossum)*						
		5: 116.5	4: 303.2	Un: 50.4, X?	3: 242.6	1: 9.8	2: 190.8
	Monotreme mammals						
15	*O. anatynus (platypus)*						
		12: 13.5/c5907: (	ultc336: 0.46¥	?	c1755: 0.011	†?	n/a?
	Reptiles						
16	*A. carolinensis (lizard)*						
		sc284: 0.6	n/a	sc1398: 0.05	sc3: 1.7	n/a	n/a?
	Birds						
17	*G. gallus (chicken)*						
18	*T. guttata (zebrafinch)*	4: 23.0	1: 179,14 †	4: 0.57	Z: 9.7	n/a	n/a
		4: 29.7	n/a	n/a?	Z: 41.0	n/a	n/a
	Amphibia						
19	*X. laevis (frog)*						
20	*R. esulenta (edible frog)*◊	sc110: 0.6	sc80: 1.6	sc422: 0.30	sc803: 0.128	n/a	sc9769: 0.009

Table B4. Relaxin family peptide/receptor genes and their IDs in teleost fish (5 species)

Chromosomal locations given in [chromosome number: megabases (Mb), rounded to the nearest 100.000 bp], all locations are from Ensembl, unless otherwise specified

rln/insl genes

Species	*rln3a*	*rln3b*	*insl3*	*insl5a*	*insl5b*	*rln*[a]
D. rerio (zebrafish)	ENSDARG00000070780	ENSDARG00000039854	ENSDARG00000035862	ENSDARG00000070966	ENSDARG00000069294	100329416
O. latipes (medaka)	ENSORLG00000011777	ENSORLG00000010278	GENSCAN00000085130	GENSCAN00000098652	NO ID	ENSORLG00000009974
G. aculeatus (stickleback)	ENSGACG00000018985	ENSGACG00000012435	NO ID	NO ID	ENSGACG00000016154	ENSGACG00000017364
T. nigroviridis (tetraodon)	GSTENG00026277001	†?	GSTENG00020897001	EU437461	EU437463	EU437459
T. rubripes (takifugu)	ENSTRUG00000010031	ENSTRUG00000012677	SINFRUG00000162280	GENSCAN00000013221	GENSCAN00000015952	ENSTRUG00000005640

rxfp1/2 and rxfp3/4 genes

Species	*rxfp1*	*rxfp2a*	*rxfp2b*	*rxfp2-like*	*rxfp3-1*	*rxfp3-2a*
D. rerio (zebrafish)	ENSDARG00000090071	ENSDARG00000032820	ENSDARG00000019660	ENSDARG00000068731	ENSDARG00000057410	**ENSDARG00000022739**
O. latipes (medaka)			n/a	n/a	ENSORLG00000006539	ENSORLG00000014985
G. aculeatus (stickleback)	ENSGACG00000016581	ENSGACG00000020550	n/a	n/a	ENSGACG00000016296	ENSGACG00000015315
T. nigroviridis (tetraodon)	ENSTNIG00000013038	ENSTNIG00000009913	n/a	n/a	ENSTNIT00000003086	ENSTNIG00000015329
T. rubripes (takifugu)	ENSTRUG00000016132	ENSTRUT00000005652	n/a	n/a	ENSTRUG00000014489	ENSTRUG00000007126

Species	*rxfp3-2b*	*rxfp3-3a2*	*rxfp3-3a1*	*rxfp3-3b*	*rxfp3-3a3*	*rxfp4*
D. rerio (zebrafish)	ENSDARG00000061846	ENSDARG00000062111	ENSDARG00000069028	ENSDARG00000059348	ENSDARG00000069246	
O. latipes (medaka)	no id	ENSORLG00000001754	ENSORLG00000002054	ENSORLG00000019204	n/a	ENSORLG00000003213
G. aculeatus (stickleback)	ENSGACG00000012856	ENSGACG00000016895	ENSGACG00000016952	ENSGACG00000008049	n/a	ENSGACG00000003931
T. nigroviridis (tetraodon)	ENSTNIG00000010632	ENSTNIG00000009550	ENSTNIG00000009561			ENSTNIG00000009161
T. rubripes (takifugu)	ENSTRUG00000017932	ENSTRUG00000016840	ENSTRUG00000016739	ENSTRUG00000014434	n/a	gi|74096006

Table B5. Relaxin family peptide/receptor genes and their chromosomal locations in teleost fish

Chromosomal locations given in [chromosome number: megabases (Mb), rounded to the nearest 100.000 bp], all locations are from Ensembl, unless otherwise specified

	rln3a	*rln3b*	*insl3*	*insl5a*	*insl5b*	*rln*[a]
Species						
D. rerio (zebrafish)	3: 19.03	1: 50.17	2: 20.05	6: 30.40	2: 9.92	21: 0.20
O. latipes (medaka)	1: 31.36	8: 30.65	17: 60.74	4: 16.56	17: 16.79	12: 14.80
G. aculeatus (stickleback)	9: 15.05	11: 11.63	3: 8.50	8: 8.10	3: 9.58	14: 7.44
T. nigroviridis (tetraodon)	18: 2.04	n/a	15: 3.88	1: 15.66	15: 3.44	4: 1.08
T. rubripes (takifugu)	sc189: 0.355	sc141: 0.678	sc212: 0.364	sc55: 0.131	sc166: 0.545	sc243: 0.093

	rxfp1	*rxfp2a*	*rxfp2b*	*rxfp2-like*	*rxfp3-1*	*rxfp3-2a*
Species						
D. rerio (zebrafish)	14: 51.06	10: 34.81	15: 30.98	5: 37.50	21: 18.31	2: 53.78 (ENSEM), 4:? (NCBI)
O. latipes (medaka)	10: ?*	14: 9.82	n/a	n/a	12: 10.00	17: 22.92
G. aculeatus (stickleback)	4: 2.39	7: 19.90	n/a	n/a	14: 3.51	3: 7.51
T. nigroviridis (tetraodon)	20: 2.24	7: 3.13	n/a	n/a	4: ?†	15: 4.70
T. rubripes (takifugu)	sc5: 0.003	sc2436: 0.007	n/a	n/a	sc44: 1.50	sc200: 0.130

	rxfp3-2b	*rxfp3-3a2*	*rxfp3-3a1*	*rxfp3-3b*	*rxfp3-3a3*	*rxfp4*
Species						
D. rerio (zebrafish)	22: 20.42	7: 58.58	7: 35.21	25: 1.41	18: 19.81	n/a
O. latipes (medaka)	4: 8.24	3: 10.95	3: 10.20	6: ?	n/a	11: 6.39
G. aculeatus (stickleback)	8: 16.28	2: 18.72	2: 19.11	19: 9.18	n/a	10: 5.35
T. nigroviridis (tetraodon)	1: 13.78	5: 10.38	5: 10.62	13: 12.75	n/a	21: 1.73
T. rubripes (takifugu)	sc25: 0.761	sc1: 3.39	sc1: 3.12	sc4528: 0.004	n/a	sc138: 0.096◊

[a] rln is the orthologue of the human RLN2 (H2) gene

†: putative pseudogene

◊ annotated on NCBI as GPR100

* ultracontig115: 1,259,584-1,305,610 (Ensembl) corresponds to chr10 on UTGB

Table B6. Relaxin family peptide/receptor + ilp (insulin-like peptide) genes and their chromosomal locations in Ascidians

(Tunicates- Ciona: genomes sequenced & assembled, 2 species; Chelyosoma productum (C. productum)- not sequenced)

	Species	Name on phylogeny	Gene ID (Ensembl)	Gene ID (NCBI)	Map location	Olinski *et al.* naming
novel ilp ◊	*C. intestinalis*	*ilp4*	ENSCING00000005306	100177666	14q: 3.31	n/a
	C. savignyi	*ilp4*	n/a	n/a	reftig_76: 1,317,679-	n/a
rxfp1/2-like*	*C. intestinalis*	*rxfp1/2-L*	n/a	100178659	3p: 3.28	n/a
rxfp3-4-like		*rxfp3-L1*	ENSCING00000012511	?	1q: 4.67	n/a
	C. intestinalis	*rxfp3-L2*	ENSCING00000014875	100185531	6q: 0.975	n/a
ilp		*ilp1*	ENSCING00000003666	100170003	2q: 5,456,963-5,458,855	*ins-L1*
	C. intestinalis	*ilp2*	n/a	100170005	2q: 1,054,089-1,058,637**	*ins-L2*
		ilp3	n/a	100170004	2q: 1,061,012-1,065,562**	*ins-L3*
"ins"	*C. productum****	*"ins"*	n/a	n/a	n/a	*n/a*
"igf"		*"igf"*	n/a	n/a	n/a	*n/a*

◊ found via PHI-BLAST (see explanation below)

* rxfp1/2 gene obtained from Kamesh et al. and verified using the Ensembl C. intestinalis browser

** map locations of ins-l2 and ins-l3 obtained from Ensembl (which does not have these genes identified) using NCBI sequences as BLAT queries

*** sequences obtained from: McRory JE and Sherwood NM (1997) Ancient divergence of insulin and insulin-like growth factor. DNA and Cell Biology 16(8):939-49.

Table B7. ***Rxfp1/2*****-type + ilp (insulin-like peptide) genes and their chromosomal locations in Lancelets (amphioxus)**
3 species

Branchiostoma floridae (genome sequenced and assembled into contigs)

	Name on phylogeny	Gene ID (NCBI)	Protein ID (JGI)	Scaffold: version2	Scaffold: version1*	Exact map location
ilps	*ilp2*	n/a**		bf_V2_190	59	
	ilp1	7235917	121099	bf_V2_190	59	scaffold_59:2166507-2182058
	ilp3	7251652	97394	bf_v2_243	302 (243)	scaffold_302:862279-904829
	ilp4	7253642	100967	bf_v2_277	41 (372)	
	ilp5	7255900	100968	bf_v2_277	41 (372)	scaffold_372:110792-121518
	ilp6	7230317	74371	bf_v2_196	73 (50)	scaffold_50:2135877-2137051
rxfp1/2-type***	*rxfp1/2-L1*	7252026	63628	bf_v2_249		
	rxfp1/2-L2	7229038	134702	bf_v2_150		
	rxfp1/2-L3	7221608	63627	bf_v2_150		
	rxfp1/2-L4	7207790	204729	bf_v2_21		
	rxfp1/2-L5	7209355	98820	bf_v2_21		

* and as appears in Holland *et al.*

***rxfp1/2 sequences were obtained from NCBI referring to Nordström *et al.*, with the exception of BRAFLDRAFT_231009, which could not be found in the current database

bf_v2_21 shares 2 genes with rxfp2 region

** gene prediction no longer valid

Ilp genes from different amphioxus species

	Name on phylogeny	Genbank ID	Species	NCBI naming
IGF-like	*ilp_B. belcheri*	GU388424.1	*Branchiostoma belcheri*	*igf-L_B_belcheri*
INS-like	*ilp_B. californiensis*	M55302.1	*Branchiostoma californiensis*	*ins-L_B_californiensis*

A search strategy similar to the one described for *C. intestinalis* (see Methods, Chapter 1) was used to look for the Insulin Superfamily genes in the *B. floridae* proteome at NCBI. A total of 6 genes were detected; two of them, "igf-L1" and "igf-L2" are flanked by the tyrosine hydroxylase (TH)-like genes whose orthologs in vertebrates flank the Igf genes; "igf-L2" was also determined to be more related to INS/IGF than to RLN/INS based on our gene history tracing , and both of these genes were found using Ins/Igf genes as queries; "igf-L3" is located in close proximity to igf-L1. The remaining 3 genes could not be assigned to a class of genes based on synteny: rln-L1 and rln-L2 were found using starfish GSS, insl5 and rln3 as queries; rln-L3 was found using starfish GSS and vertebrate Insl3 as queries. The six amphioxus candidate genes were mapped to scaffolds using the BLASTn (Altschul et al., 1997) tool as implemented in the Amphioxus genome database (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html).

Table B8. GSS (Gonad-stimulating substance) & *Rxfp1/2* -type genes in Echinoderms (2 species)

Starfish (Asterina pectinifera) GSS IDs are from GenBank,

Sea Urchin (Strongylocentrotus purpuratus) rxfp1/2 IDs are from SpBase (http://www.spbase.org/SpBase/)

	Species	Name on phylogeny	Gene ID	mRNA ID	Protein ID
GSS	*A. pectinifera (starfish)*	gss_Starfish	n/a	AB496611.1	BAI44654.1
rxfp1/2s	*S. purpuratus (sea urchin)*	rxfp1/2-L1	SPU_000792		
		rxfp1/2-L2	SPU_000840		
		rxfp1/2-L3	SPU_001502		
		rxfp1/2-L4	SPU_003492		
		rxfp1/2-L5	SPU_003527		
		rxfp1/2-L6	SPU_004308		
		rxfp1/2-L7	SPU_005497		
		rxfp1/2-L8	SPU_009179		
		rxfp1/2-L9	SPU_011953		
		rxfp1/2-L10	SPU_013866		
		rxfp1/2-L11	SPU_015134		
		rxfp1/2-L12	SPU_016206		
		rxfp1/2-L13	SPU_019187		
		rxfp1/2-L14	SPU_019188		
		rxfp1/2-L15	SPU_019240		
		rxfp1/2-L16	SPU_019676		
		rxfp1/2-L17	SPU_020381		
		rxfp1/2-L18	SPU_020408		
		rxfp1/2-L19	SPU_022337		
		rxfp1/2-L20	SPU_023168		
		rxfp1/2-L21	SPU_023629		
		rxfp1/2-L22	SPU_024157		
		rxfp1/2-L23	SPU_025906		
		rxfp1/2-L24	SPU_026409		
		rxfp1/2-L25	SPU_027093		
		rxfp1/2-L26	SPU_027094		
		rxfp1/2-L27	SPU_027097		
		rxfp1/2-L28	SPU_028324		

Table B9. Vertebrate INS/IGF genes in phylogenetic analyses (11 species)

	Medaka (*O. latipes*)						
Name	ins1	ins2	igf1	igf2			
Map location	scaffold498: 0.025	scaffold223: 0.409	Chr23: 20.44	scaffold1060: 0.013			
Ensembl ID	ENSORLG00000018432	ENSORLG00000018994	ENSORLG00000016443	ENSORLG00000018930			
Chromosomes*	21	14	23	14?			
				missing b-chain, not included in final phylogeny			
	Tetraodon (*T. nigroviridis*)						
Name	ins1	ins2	igf1	igf2			
Map location	1: 3.4	7: 10.6	19: 1.4	13: 6.6			
Ensembl ID	ENSTNIG00000004978	ENSTNIG00000004978	ENSTNIG00000012663	ENSTNIG00000010464			
	Stickleback (*G. aculeatus*)						
Name	ins	igf1	igf2				
Map location	scaffold_132: 0.057	groupIV: 32.1	groupXIX: 13.3				
Ensembl ID	ENSGACG00000001771	ENSGACG00000020042	ENSGACG00000011125				
	Human (*H. sapiens*)						
Name	INS	IGF1	IGF2				
Map location	Chr11: 2.15	Chr12: 102.79	Chr11: 2.15				
Ensembl ID	ENSG00000129965	ENSG00000017427	ENSG00000167244				
	Zebrafish (*D. rerio*)						
Name	insb	insa	igf1a	igf1b	igf2a	igf2b	ins-like
Map location	Chr14: 29.89	Chr5: 36.46	Chr 4: 17.12	Chr8: 14.14			Chr3:32.6
Ensembl/NCBI ID	ENSDARG00000034610	ENSDARG00000035350	ENSDARG00000014109	ENSDARG00000058058	NM_131433.1	NM_001001815	XM_003198132.1
	Clawed frog (*X. tropicalis*)						
Name	ins	igf1	igf2a	igf2b			
Map location	scaffold_419: 0.868	scaffold_243: 0.047	scaffold_587: 0.296	scaffold_419: 0.670			
Ensembl ID	ENSXETG00000014029	ENSXETG00000002532	ENSXETG00000002876	ENSXETG00000014020			
	Chicken (*G. gallus*)						
Name	ins	igf1	igf2				
Map location	Chr5: 14.85	Chr1: 57.33	Chr5: 14.88				
Ensembl ID	ENSGALG00000006552	ENSGALG00000012755	ENSGALG00000006555				
	Mouse (*M. musculus*)						
Name	ins1	ins2	igf1	igf2			
Map location	19: 52.34	7: 149.86	10: 87.32	7: 149.84			
Ensembl ID	ENSMUSG00000035804	ENSMUSG00000000215	ENSMUSG00000020053	ENSMUSG00000048583			
	Platypus (*O. anatynus*)						
Name	igf1	igf2					
Map location	n/a	n/a					
NCBI ID	100075905	791102					
	Hagfish (*M. glutinosa*)						
Name	ins-like	igf-like					
Genbank ID	V00649.1	M57735.1					
	Lamprey (*P. marinus*)						
Name	igf	ins					
Genbank ID	AB081462.1	*Genscan prediction on the PreEnsembl Lamprey database*					

* putative chromosomes from UTGB database: the medaka genome is not fully assembled

Table B10. Fruit fly *(Drosophila melanogaster)* insulin-like peptides (*ilp*) and *rxfp1/2* -type genes

all sequences obtained from Ensembl Metazoa (http://www.metazoa.ensembl.org/)

Name	*ilp1*	*ilp2*	*ilp3*	*ilp4*	*ilp5*	*ilp6*	*ilp7*
Map location	3L: 9,79	3L: 9,79	3L: 9,79	3L: 9,80	3L: 9,82	X: 2,23	X: 3,56
ID	FBgn0044051	FBtr0076329	FBtr0076373	FBgn0044049	FBtr0076371	FBtr0070406	FBtr0070577

rxfp1/2

Name	*lgr3*	*lgr4*
Map location	3R: 21	X: 13
ID	FBgn0039354	FBgn0085440

Table B11. Relaxin family peptide-like genes in cartilaginous fish (4 species)

Sequences for elephant shark were found via megablast/tblastn from Trace Archives (NCBI, Elephant shark WGS)

Cartilaginous fish	Name on phylogeny*	Sequences (Genbank)	Possibly:
S.acanthias (Dogfish)	*rfpl_Dogfish*	gi|27734670|sp|P11953.2|RELX_SQUAC	*rln*
C. taurus (Sand tiger shark)	*rfpl_SandTigerShark*	gi|27734666|sp|P01349.2|RELX_ODOTA	*rln*
R. erinacea (Little skate)	*rfpl_Skate*	gi|32172395|sp|P11952.1|RELX_RAJER	*insl5*
C. milii (Elephant shark)	*rfpl1_ElephantShark*	*n/a*	*rln*
	rfpl2_ElephantShark	*n/a*	*rln3*

Because of the absence of an assembled version of the shark genome, the locations of the obtained sequences could not be mapped. Although numerous hits were produced when the shark Trace Archive was blasted with the reference RXFP sequences, these traces could not be assembled with enough confidence to yield shark RXFP-like sequences that could be used in phylogenetic reconstruction.

Table B12. Relaxin family peptide/receptor genes in jawless fish (lamprey: genome sequenced, pre-assembled)

Sequences for lamprey were found via megablast/tblastn from Trace Archives (NCBI)

	Name on phylogeny	Map location	Pre-Ensembl Genscan ID
rln(3)-like			
	rfpl1	Contig61593:5820:6505	n/a
	rfpl2	Contig7836:4268:4340	n/a
rxfp3/4-like	*rxfp3-L1*	Contig3136: 36,508-40,602	GENSCAN00000044236
	rxfp3-L2	Contig57509.1: 238-3,235	GENSCAN00000070350
	rxfp3-L3	Contig581: 1-54,828	GENSCAN00000048927 (first half of the gene), GENSCAN00000142421 (second half)
	rxfp3-L4	Contig14581: 1-3,029	GENSCAN00000072822
rxfp1/2-like			
	rxfp1/2-L	Contig19837.2: 883-6,619	GENSCAN00000077907

* rfpl: relaxin family peptide-like

Genscan entries in most cases had to be edited to get rid of contaminating adjacent cds

† possible pseudogene

APPENDIX C: Supplementary Figures (Chapter 1)

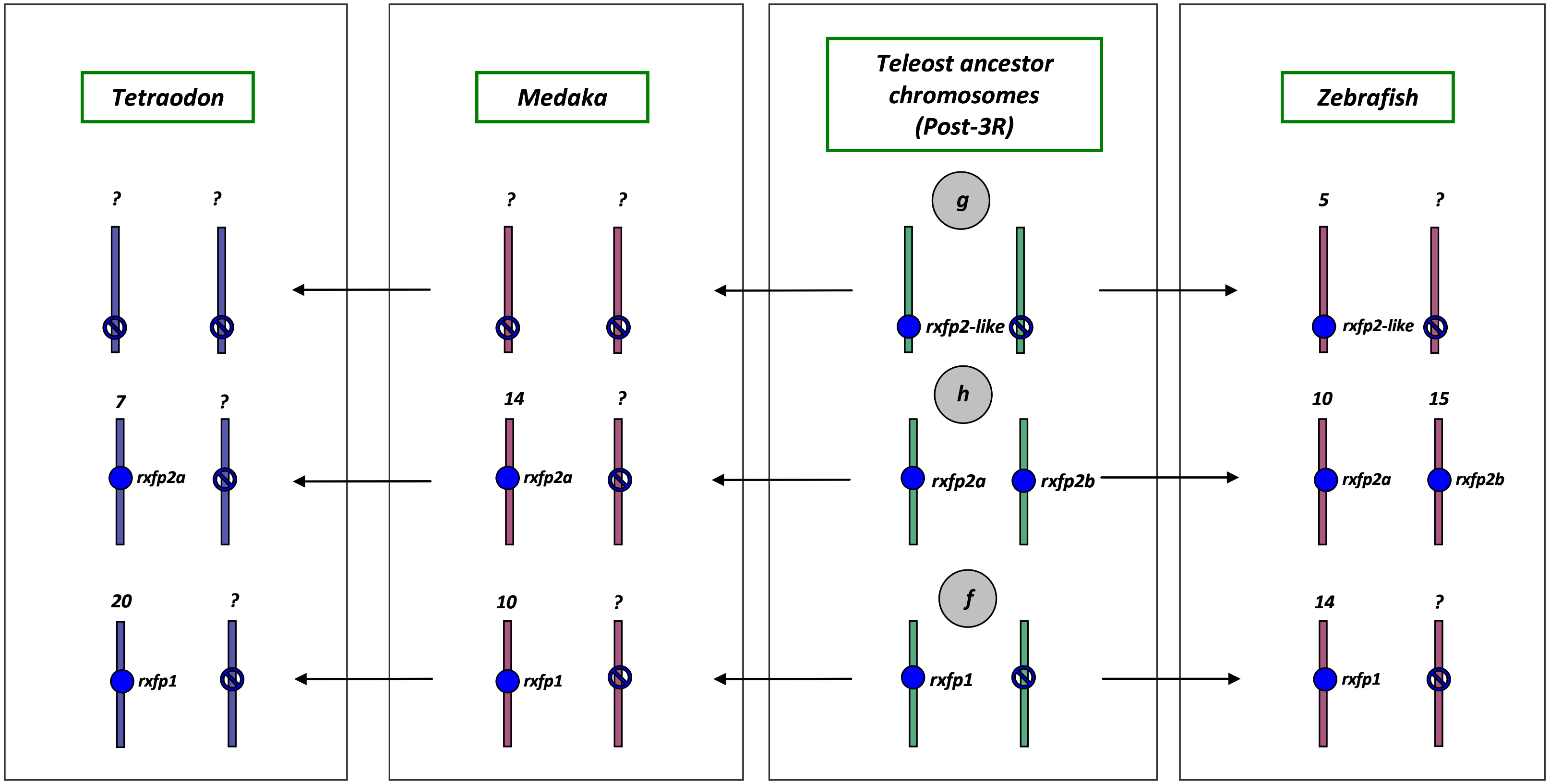


Figure C1. Linkage relationships among *rxfp1/2*-type genes in teleosts

Figure C2 *(next two pages)*. The expanded versions of *Figure 5* (in main text). **a)** Tree showing the "*INS-IGF* Vertebrates & Protochordates" clade in detail. **b)** Tree showing the "*RLN/INSL* Vertebrates & *ilp5* Amphioxus" clade in detail.

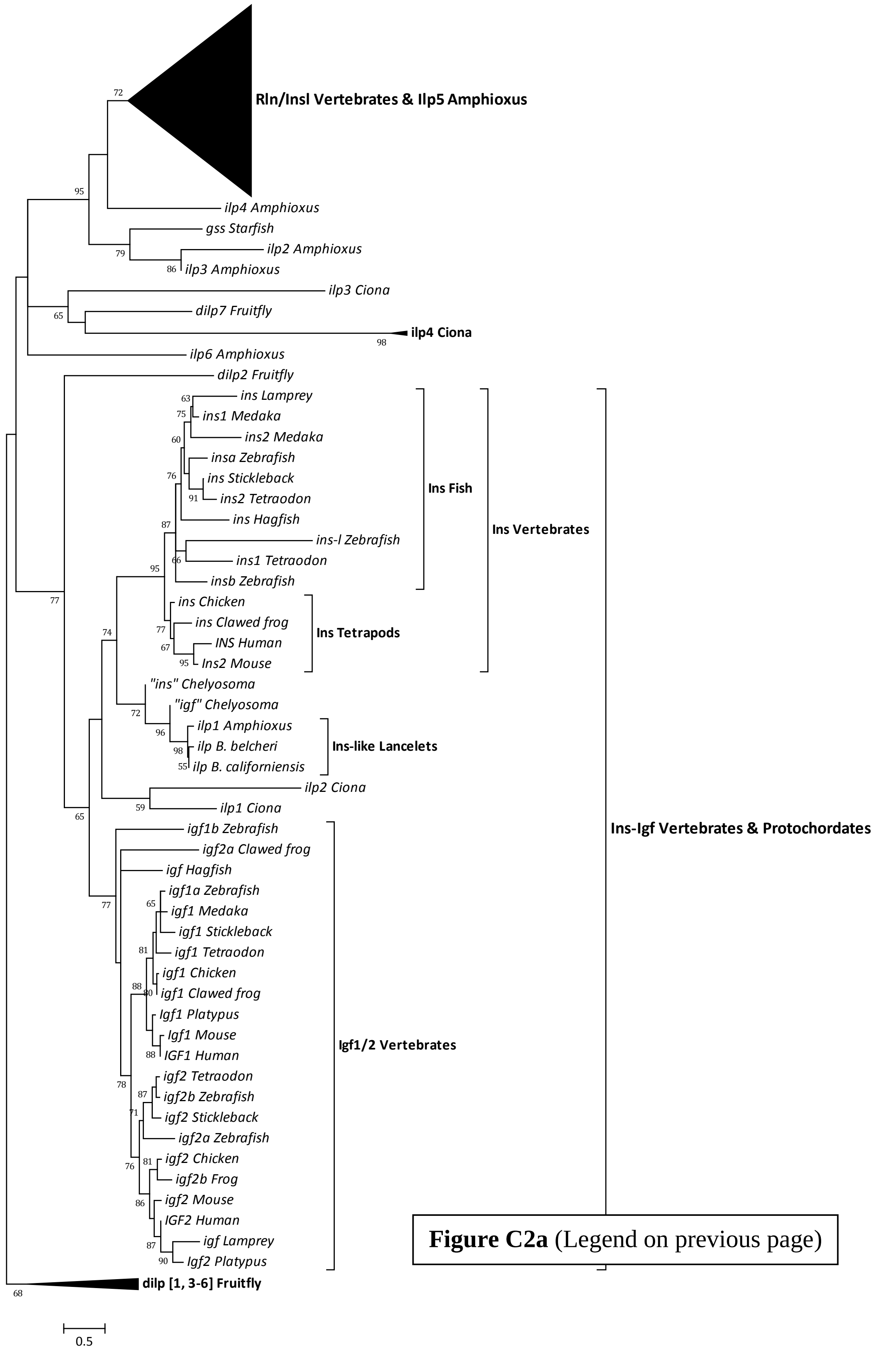


Figure C2a (Legend on previous page)

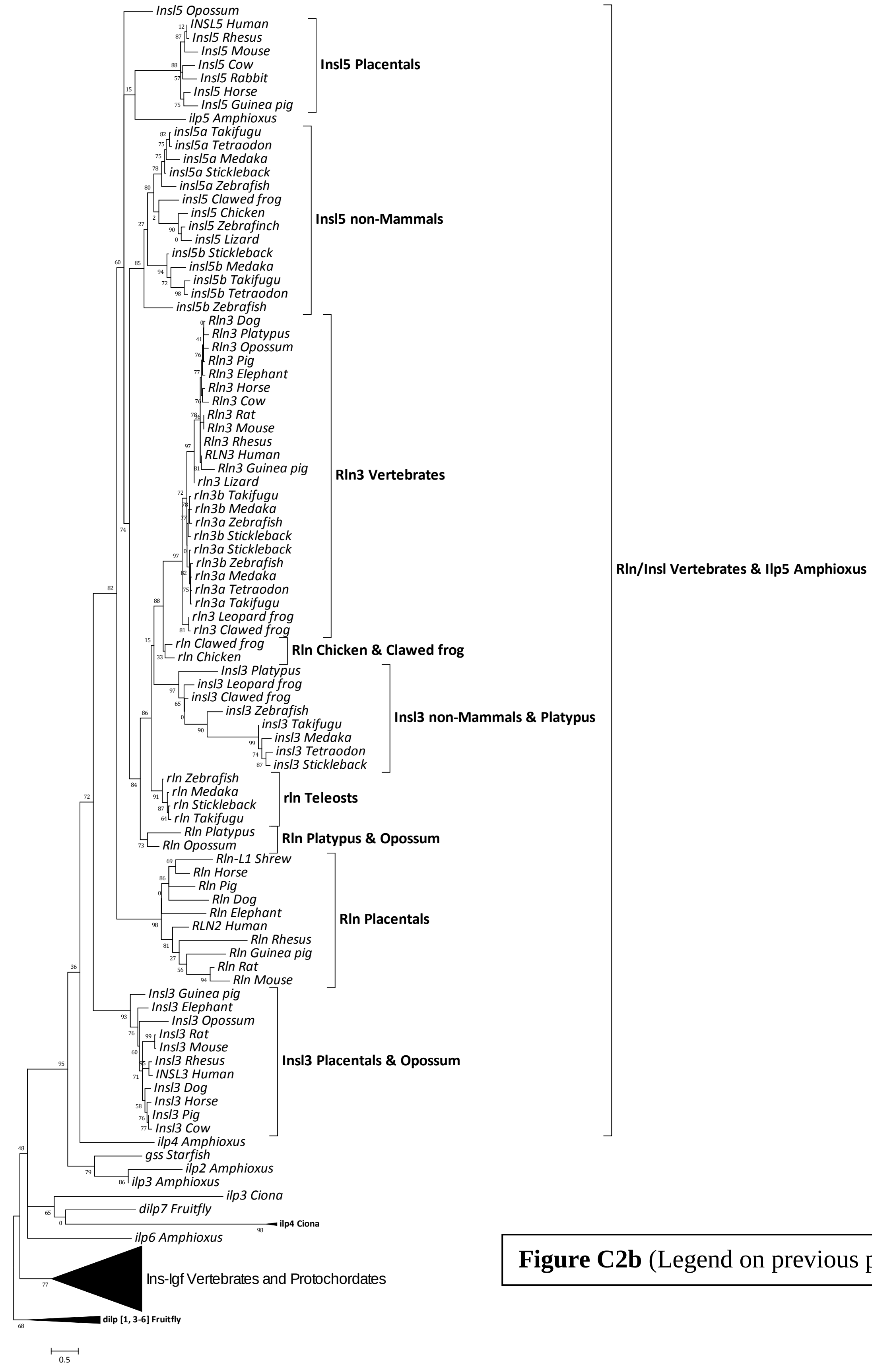


Figure C2b (Legend on previous page)

APPENDIX D: Supplementary Figures (Chapter 2)

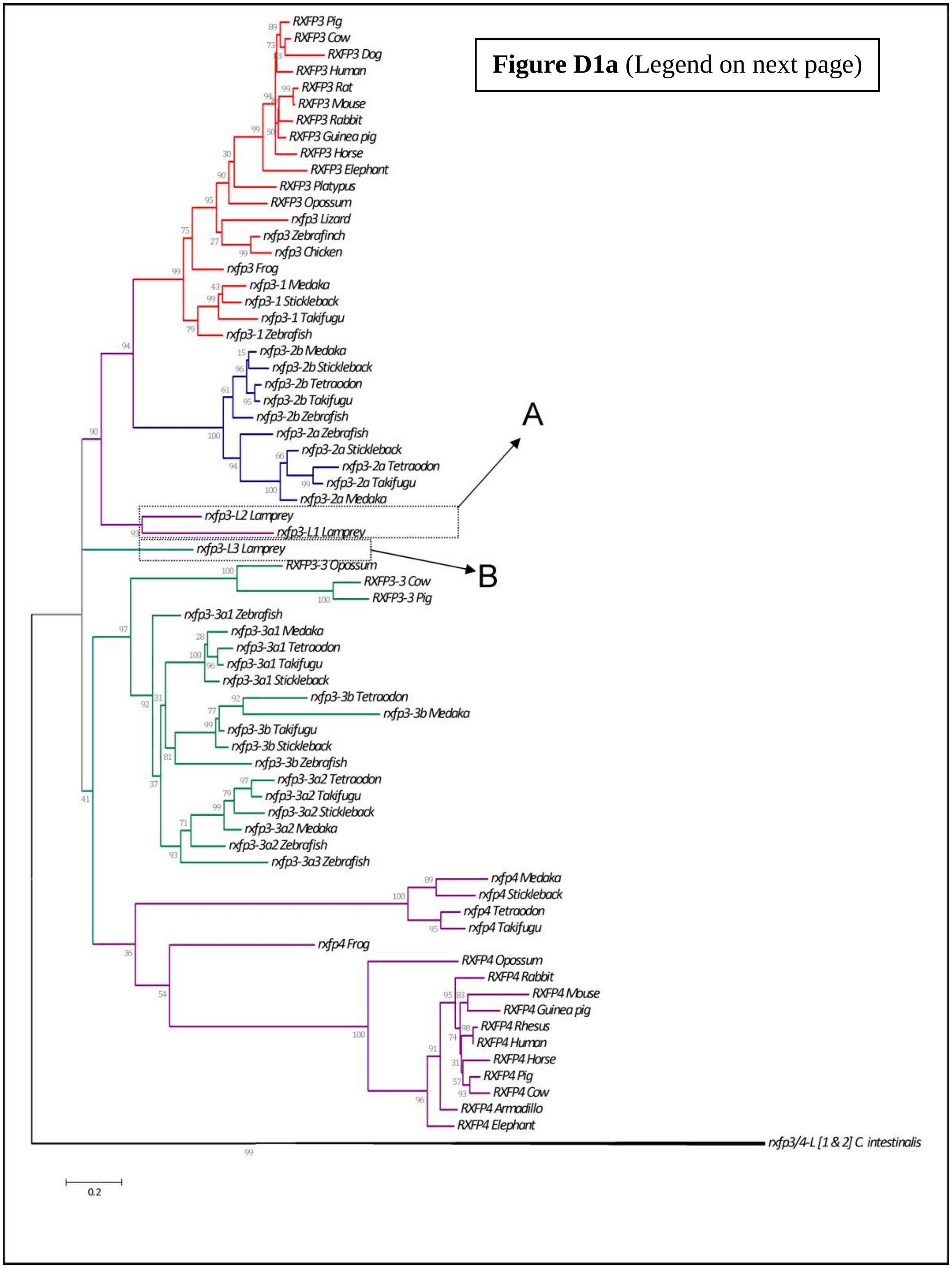


Figure D1a (Legend on next page)

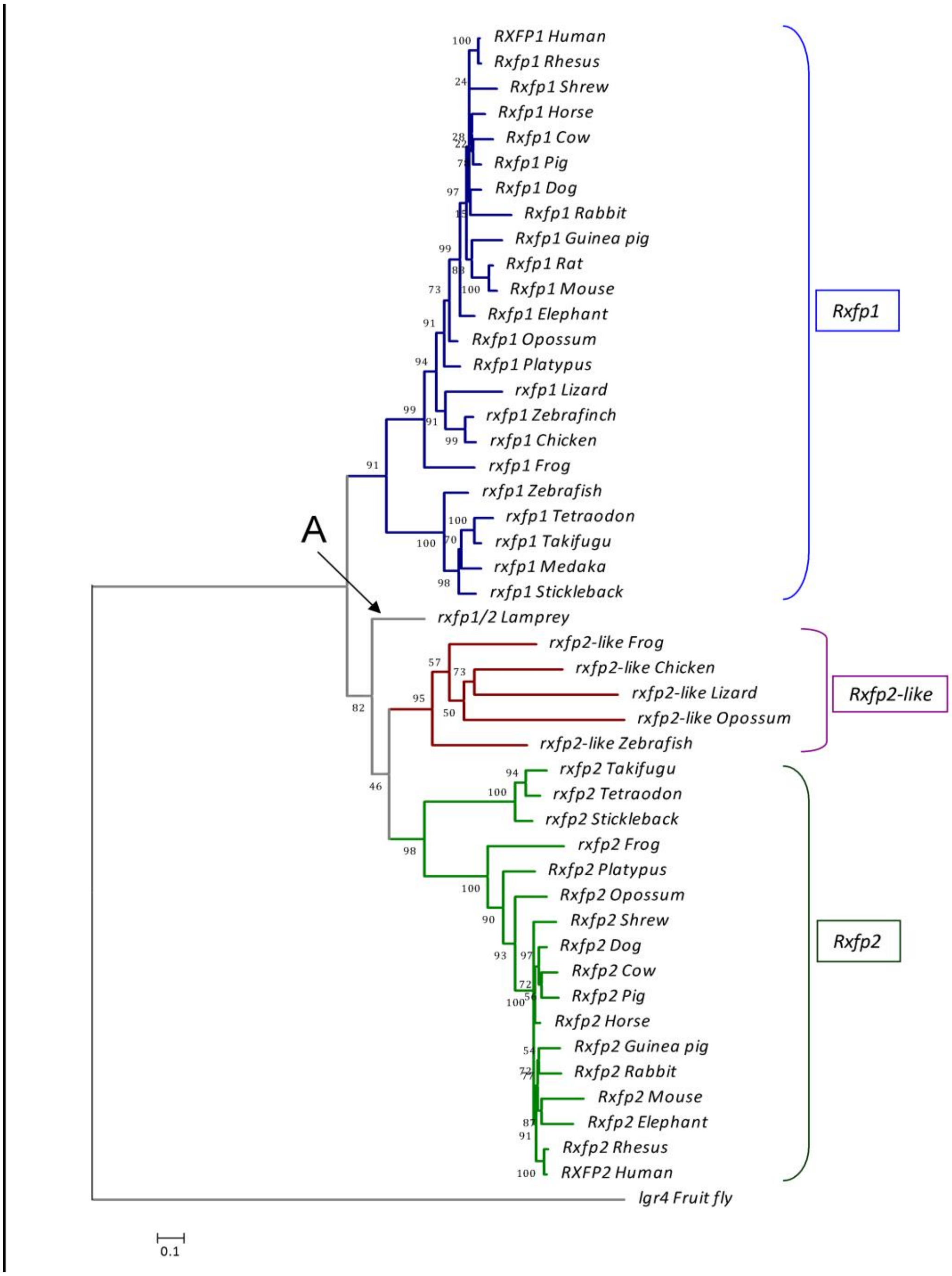


Figure D1. Phylogenetic reconstruction of evolutionary relationships among the RXFP receptors of tetrapods, teleosts and novel lamprey rxfp sequences. **a)** ***(previous page)*** lamprey rxfp3-L1 and rxfp3-L2 (“A”) are at the base of the RXFP3-1/3-2 clade, the rxfp3-L3 sequence (“B”); **b) (above)** rxfp1/2 (“A”) is at the base of the RXFP2/RXFP2-like clade.

APPENDIX E: Selection analyses (Chapter 3)

Table E1. Raw data used to build histograms in Chapter 3.

	Type of Selection				**Type of Selection**		
	purifying	***neutral***	***positive***		***purifying***	***neutral***	***positive***
rln	0.39257	0.08531	0.52212	rln	0.43781	0.56219	0
rxfp1	0.79686	0.11583	0.08732	rxfp1	0.81697	0.12334	0.05969
insl3	0.73771	0.10518	0.1571	insl3	0.71866	0.16757	0.11376
rxfp2	0.74526	0.15082	0.08643	rxfp2	0.72576	0.19448	0.0629
rxfp2-like	0.6448	0.1495	0.2057	rxfp2-like*	0.6448	0.1495	0.2057
rln3	0.92972	0.03525	0.03503	rln3a	0.9693	0.0241	0.0066
rxfp3-1	0.90509	0.03581	0.05685	rln3b	0.96403	0.01761	0.01836
insl5	0.60042	0.33076	0.06883	rxfp3-1	0.92523	0.05126	0.02228
rxfp4	0.79991	0.07244	0.12766	insl5a	0.60428	0.39572	0
				insl5b	0.57975	0.38124	0.03902
				rxfp4	0.7616	0.12558	0.11282
				rxfp3-2a	0.85722	0.02068	0.1121
				rxfp3-2b	0.93831	0.06169	0
				rxfp3-3a1	0.89025	0.04174	0.06802
				rxfp3-3a2	0.93659	0.0444	0.01906
				rxfp3-3b	0.92527	0.04924	0.02549