J Mol Evol (2012) 75:73–78
DOI 10.1007/s00239-012-9517-0

INSL4 Pseudogenes Help Define the Relaxin Family Repertoire in the Common Ancestor of Placental Mammals

José Ignacio Arroyo · Federico G. Hoffmann · Sara Good · Juan C. Opazo

Received: 16 April 2012 / Accepted: 3 August 2012 / Published online: 9 September 2012

Abstract The relaxin/insulin-like (*RLN*/*INSL*) gene family comprises a group of signaling molecules that perform physiological roles related mostly to reproduction and neuroendocrine regulation. They are found on three different locations in the mammalian genome, which have been called relaxin family locus (RFL) A, B, and C. Early in placental mammalian evolution, the ancestral proto-*RLN* gene at the RFLB locus underwent successive rounds of small-scale duplications resulting in variable number of paralogous genes in different placental lineages. Most placental mammals harbor copies of the RLN2 and INSL6 paralogs in the RFLB. However, the origin of an additional paralog, INSL4 (also known as placentin), has been controversial as its phyletic distribution does not converge with its phylogenetic position. In principle, by searching for INSL4 genes in representative species of all major groups of mammals we can gain insights into when the gene originated and better reconstruct its evolutionary history. Here we identified INSL4 pseudogenes in two laurasiatherian, (alpaca and dolphin) and one xenarthran (armadillo) species. Phylogenetic and synteny analyses confirmed that the identified pseudogenes are orthologs of INSL4. According to these results, the proto-*RLN* gene in the RFLB underwent two successive tandem duplications which gave rise the INSL6 and INSL4 paralogs in the last common ancestor of placental mammals. The INSL4 gene was subsequently inactivated or lost from the genome in all placentals other than catarrhine primates, where its product became functionally relevant. Our results highlight the contribution of relatively old gene duplicates to the gene complement of extant species.

Keywords INSL4 · Placentin · Differential retention · Gene family evolution · Mammals · Gene duplication

Electronic supplementary material The online version of this article (doi:10.1007/s00239-012-9517-0) contains supplementary material, which is available to authorized users.

J. I. Arroyo · J. C. Opazo (✉)
Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
e-mail: jopazo@gmail.com

F. G. Hoffmann
Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Starkville, MS, USA

F. G. Hoffmann
Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Starkville, MS, USA

S. Good
Department of Biology, University of Winnipeg, Winnipeg, MB, Canada

Introduction

The relaxin (RLN)/insulin-like (INSL) family of peptides is related to insulin and INSL growth factors, and includes signaling molecules that perform a variety of physiological roles mostly related to reproduction and neuroendocrine regulation (Bathgate et al. 2003; Sherwood 2004; Park et al. 2005; McGowan et al. 2008). Recent analyses revealed that the two whole genome duplications that occurred early in vertebrate evolution are linked to the initial expansion of this group of genes (Hoffmann and Opazo 2011; Yegorov and Good 2012). Members of this gene family are found on three different genomic locations in mammals, which have been called relaxin family locus (RFL) A, B, and C (Park et al. 2008). The number and

nature of genes at these three loci are well conserved in most mammalian groups, with the exception of the RFLB locus (Fig. 1; Park et al. 2008; Hoffmann and Opazo 2011; Arroyo et al. 2012a, b). This locus underwent successive small-scale duplications such that its gene repertoire is variable among mammalian lineages. A duplication event of the proto-*RLN* gene in the eutherian ancestor gave rise to the INSL6 gene, located upstream from RLN2 (Fig. 1); and two additional genes have also been identified in primates. The INSL4 gene, also known as placentin, is found in tarsier and anthropoids (haplorhines), and an additional paralog of RLN2, RLN1 is present in apes (Fig. 1; Arroyo et al. 2012a, b). Previously, and because apes are the only group with copies of both RLN1 and RLN2, these genes were thought to have derived from a duplication event in the last common ancestor of the group. More comprehensive phylogenies, however, suggested that the duplication event that gave rise to these genes predates the radiation of anthropoids (apes, New World monkeys, and Old World monkeys; Arroyo et al. 2012a). The evolutionary history of the INSL4 gene is also complex, it was first identified in catarrhines, the group that includes apes and Old World monkeys, and so its origin was traced back to the common ancestor of catarrhines (Bieche et al. 2003; Park et al. 2008). Similar to the case of the RLN1 and RLN2 paralogs in apes, phylogenetic reconstructions of INSL4 indicate an older evolutionary origin as all INSL4 genes were placed in a monophyletic group sister to laurasiatherian RLN2 sequences (Hoffmann and Opazo 2011). The identification of INSL4 pseudogenes in a New World monkey and tarsier species, which are not catarrhine primates, provided independent confirmation, and together these results push back the origin of INSL4 to at least the last common ancestor of haplorhines (Arroyo et al. 2012a). Although a strict reconciliation of the INSL4 gene tree with the organismal phylogeny would suggest that the origin of INSL4 actually

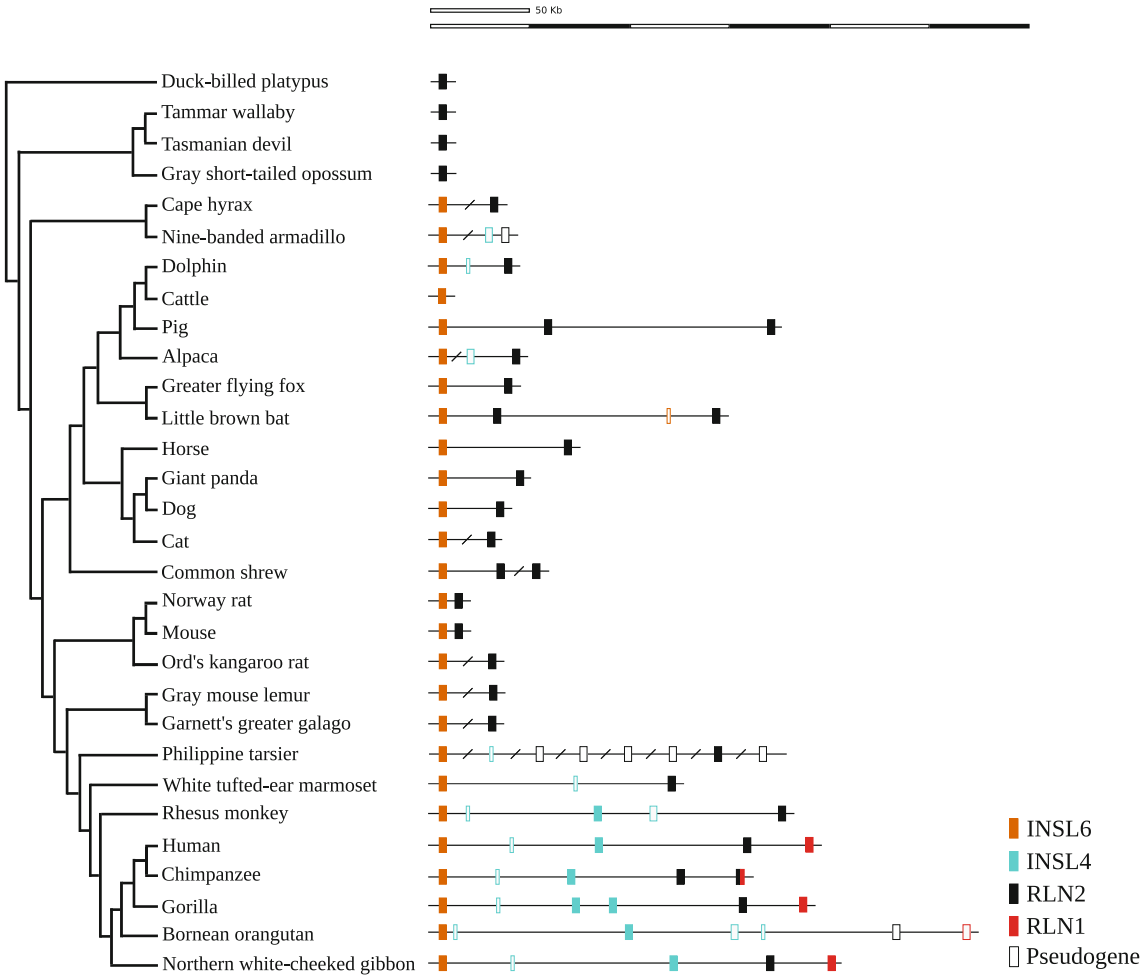

Fig. 1 Genomic structure of the RFLB in mammals. Diagonal slashes indicate that the gene was identified in different genomic pieces. The orientation of the clusters is from 5′ (on the left) to 3′ (on the right). Phylogenetic relationships are based on a loose consensus of recent studies (Van Rheede et al. 2003; Hallstrom and Janke 2010; Nery et al. 2012)

predates the origin of primates, it was not possible to reach a conclusive answer (Arroyo et al. 2012a). Accordingly, the main aim of this study was to exhaustively search for traces of INSL4 genes in representative mammals in order to reconstruct the duplication history of the *INSL4* gene and test competing evolutionary scenarios regarding its origin.

Through bioinformatic searches we were able to identify INSL4 pseudogenes in three non-primate mammals: two laurasiatherian species, alpaca and dolphin, and one xenarthran, nine-banded armadillo (Fig. 1). In the alpaca (*Vicugna pacos*), we identified both exons of the INSL4 pseudogene, which were characterized by the presence of nucleotide insertions and multiple stop codons. Synteny is also well conserved as the RLFB locus in alpaca is flanked by janus kinase 2 (JAK2), RNA terminal phosphate cyclase-like 1 (RCL1) and adenylate kinase 3 (AK3) on the 5′ side, and by the chromosome 9 open reading frame 46 (C9Orf46), CD274 molecule (CD274), and programmed cell death 1 ligand 2 (PDCD1LG2) on the 3′ end (Supplementary Fig. 1). In dolphin (*Tursiops truncatus*), we were only able to identify traces of the second exon which was found between the *INSL6* and *RLN2* genes as expected for an INSL4 sequence (Fig. 1), and is also characterized by the presence of multiple stop codons. In this case, synteny is also well conserved, as genes found up- and downstream are conserved relative to other mammals (Supplementary Fig. 1). Finally, in the nine-banded armadillo (*Dasypus novemcinctus*), we identified traces of the first exon and a partial match to the second exon at the 5′ side of a RLN2 pseudogene, as expected for an INSL4 sequence (Fig. 1), and, as in the other two cases, it possesses multiple stop codons and indels. However, given the fragmentary nature of the genomic data, we were not able to assess synteny conservation in this species.

We then conducted phylogenetic analysis to resolve orthologous relationships between the INSL4 pseudogenes found in the two laurasiatherian and one xenarthran species and the different genes found at the RFLB locus in other mammalian species (Fig. 2). The tree topologies obtained are generally concordant with the evolutionary history proposed for the genes located on the RFLB locus (Fig. 2). As previously, the monophyletic clade containing all INSL6 orthologs was recovered as sister to the clade that includes the INSL4, RLN2, and RLN1 genes of placental mammals (Fig. 2; Hoffmann and Opazo 2011; Arroyo et al. 2012a, b). With regard to the INSL4 genes, maximum likelihood and Bayesian phylogenies placed all INSL4 sequences in a monophyletic clade that includes the laurasiatherian and xenarthran pseudogenes (Fig. 2). The pseudogenes identified in alpaca and dolphin were placed sister to the clade containing all primate INSL4 sequences with strong support, whereas the nine-banded armadillo INSL4 pseudogene was placed as sister to all the other INSL4 sequences with moderate support (Fig. 2). In agreement with its phyletic distribution, which now includes all main groups of placental mammals (Euarchontoglires, Laurasiatheria, and Atlantogenata), the INSL4 clade was recovered as sister to the clade that includes all RLN2 and RLN1 sequences from placental mammals other than the *RLN2* gene from cape hyrax (Fig. 2).

Our revised model for the evolution of the RFLB in placental mammals is graphically depicted in Fig. 3. According to this model, two successive duplications of a proto-*RLN* gene gave rise to the *INSL6* and *INSL4* genes in the last common ancestor of placental mammals, estimated at 104.7–176.1 mya (Kumar and Hedges 2011), after they diverged from marsupials. Our reconstruction of the RFLB in the last common ancestor of placental mammals indicates the presence of three different paralogs, INSL6–INSL4–RLN2, a scenario slightly more complex than the one proposed by Yegorov and Good (2012), in which only two genes (*INSL6* and *RLN2*) were present. After that, we find that the *INSL4* gene was either secondarily inactivated or altogether lost from the genomes of all placentals other than catarrhine primates. In the last common ancestor of anthropoid primates a proto-*RLN* gene gave rise to the RLN1 and RLN2 genes. Although the two-gene arrangement was present in the last common ancestor of anthropoid primates, only apes appear to have retained both copies, whereas New and Old World monkeys independently retained the RLN2 paralog (Arroyo et al. 2012a). The *INSL6* gene has remained as a single-copy gene in most placental mammals.

It is interesting to note that although in catarrhine primates the INSL4 product performs essential physiological roles, mostly related to fetal and placental growth and development, and bone formation (Laurent et al. 1998; Millar et al. 2005), in all other placental species this gene appears not to be indispensable. It is possible that during its early evolution the product of INSL4 had a redundant physiological role, and that its retention in catarrhines was a matter of chance. However, the possession of multiples gene copies may provide opportunities for physiological innovation because it allows duplicated genes to acquire new functions. In support of this scenario, Wilkinson et al. (2005) estimated an event of positive Darwinian selection in the branch leading to the INSL4 clade of catarrhine primates, which is probably related to the novel role this gene acquired in this group. According to the primate tree of life, this evolutionary event occurred after the divergence of catarrhines and New World monkeys (ca. 45 mya) but before the catarrhine radiation (ca. 30 mya).

More in general our results illustrate the evolutionary potential of the differential retention of relatively old duplicates in shaping genome complement, as both the

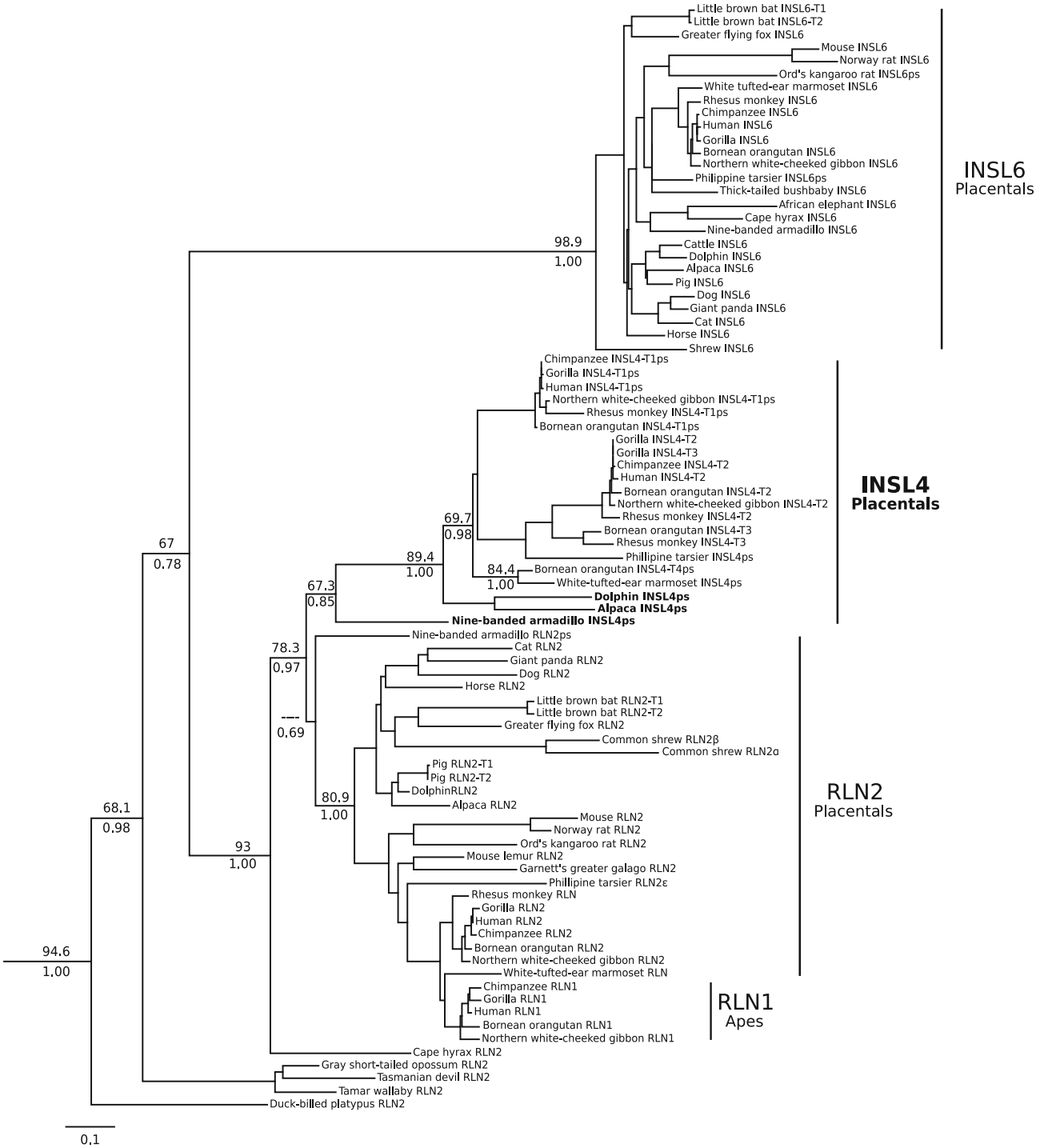

Fig. 2 Maximum likelihood phylogram depicting phylogenetic relationships among the RFLB genes in mammals. The INSL4 pseudogenes identified in laurasiatherians and xenarthrans are in bold. Number above the nodes correspond to maximum likelihood bootstrap support values, and numbers below the nodes correspond to Bayesian posterior probabilities. Sequences from the chicken (*Gallus gallus*), western clawed frog (*Xenopus tropicalis*), fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), pufferfish (*Tetraodon nigroviridis*), and stickleback (*Gasterosteus aculeatus*) were used as outgroups

duplicate RLN1 and RLN2 paralogs of apes, and the *INSL4* genes of primates derive from duplications that are much older than what their phyletic distributions suggest. In both cases, the resulting genes have been lost in most species, but may have acquired functionally important roles in those lineages that have retained them.

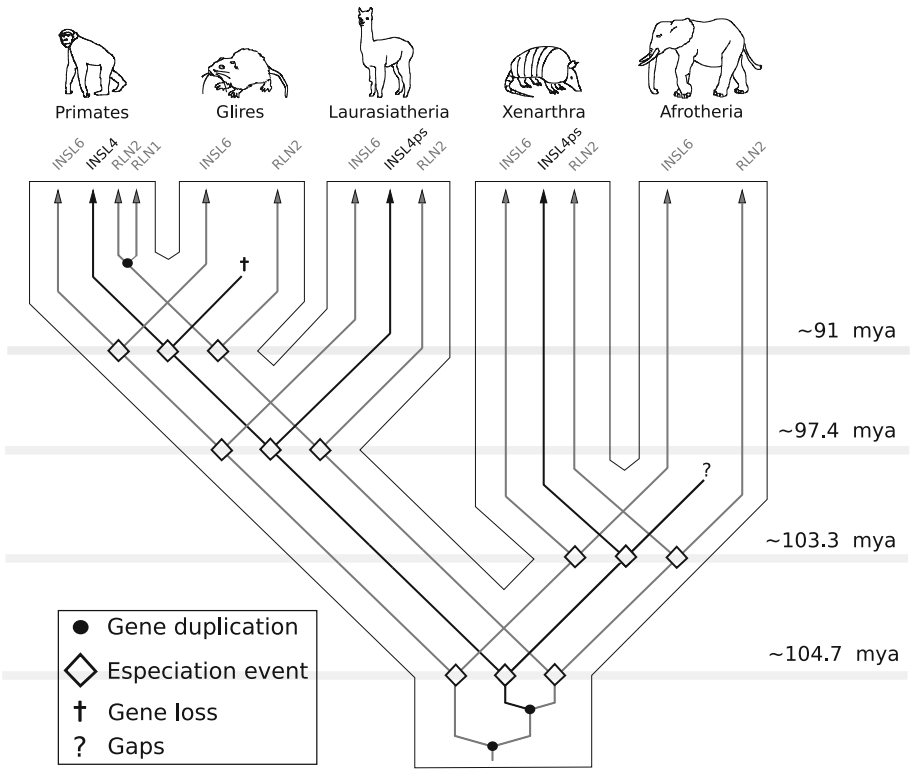

Fig. 3 An evolutionary model for the evolution of the RFLB in placental mammals. According to this model, two successive duplications of a proto-*RLN* gene gave rise to the INSL6 and INSL4 genes in the last common ancestor of placental mammals, estimated at 104.7–176.1 mya (Kumar and Hedges 2011), after they diverged from marsupials. Our reconstruction of the RFLB in the last common ancestor of placental mammals indicates the presence of three different paralogs, INSL6–INSL4–RLN2, a scenario slightly more complex than the one proposed by Yegorov and Good (2012), in which only two genes (INSL6 and RLN2) were present. After that, we find that the INSL4 gene was either secondarily inactivated or altogether lost from the genomes of all eutherians other than catarrhine primates. In the last common ancestor of anthropoid primates a proto-*RLN* gene gave rise to the RLN1 and RLN2 genes, although the two-gene arrangement was present in the last common ancestor of anthropoid primates, only apes appear to have retained both copies, whereas New and Old World monkeys independently retained the RLN2 paralog (Arroyo et al. 2012a). The INSL6 gene has remained as a single-copy gene in most placental mammals

Materials and Methods

We searched for traces of RLN/INSL genes in the RFLB locus in representative species of all main groups of placental mammals (Euarchontoglires, Laurasiatheria, and Atlantogenata) for which good quality data was available (Supplementary Table 1). This strategy was designed based on previous research which indicates that INSL4 most probably derives from the differential retention of a gene that was originated during the radiation of placental mammals, rather than an evolutionary innovation of haplorhine primates (Hoffmann and Opazo 2011; Arroyo et al. 2012a, b). In support of this claim, other groups have failed to find traces of INSL4 outside placental mammals (Park et al. 2008). However, to err on the safe side, we also included sequences from other vertebrate groups (marsupials, monotremes, birds, amphibians, and fish), to improve our phylogenetic design to accurately infer homology relationships (Supplementary Table 1). We manually identified RLN/INSL-like genes by comparing annotated exon sequences to unannotated genomic sequences, using the program Blast2seq (Tatusova and Madden 1999) available from NCBI. Putatively functional genes were characterized by an intact open reading frame with the canonical two exon/one intron structure typical of vertebrate RLN/INSL-like genes, whereas pseudogenes were identifiable because of their high sequence similarity to functional orthologs and the presence of inactivating mutations, and/or the lack of exons. To distinguish among tandemly arrayed genes copies, we index each gene-copy with the symbol T followed by a number that corresponds to the linkage order in the 5′ to 3′ orientation, thus, the first gene in the cluster is labeled T1, the second T2, and so forth.

Sequences were aligned using the L-INS-i strategy from Mafft v.6 (Katoh et al. 2009). We estimated phylogenetic relationships among the different RFLB locus genes in all major groups of mammals (monotremes, marsupials, and

placentals), using sequences from chicken (*Gallus gallus*), western clawed frog (*Xenopus tropicalis*), fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), pufferfish (*Tetraodon nigroviridis*), and stickleback (*Gasterosteus aculeatus*) as outgroup. We used a maximum likelihood and a Bayesian approach, as implemented in the program Treefinder version March 2011 (Jobb et al. 2004) and Mr.Bayes v3.1.2 (Ronquist and Huelsenbeck 2003), respectively. The different domains of the *RLN*/*INSL*-like genes probably evolve under somewhat different evolutionary regimes. To accommodate this, each of the domains (signal peptide, and peptides B, C, and A), were set as independent partitions and were allowed to have and independent model of nucleotide substitution. For each of these partitions, the best fitting model of nucleotide substitution was estimated separately using the "propose model" routine from Treefinder version March 2011. In the case of maximum likelihood, we estimated the best tree under the selected models, and assessed support for the nodes with 1,000 bootstrap pseudoreplicates. In Bayesian analysis, two simultaneous independent runs were performed for 30×10^6 iterations of a Markov Chain Monte Carlo algorithm, with six simultaneous chains sampling trees every 1,000 generations. Support for the nodes and parameter estimates were derived from a majority rule consensus of the last 15,000 trees sampled after convergence. The average standard deviation of split frequencies remained 0.01 after the burn-in threshold.

Acknowledgments We thank Sergey Yegorov for comments on the manuscript. This study was funded by Grants to JCO from the Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT 11080181). FGH acknowledges grant support from the National Science Foundation (EPS-0903787).

References

Arroyo JI, Hoffmann FG, Opazo JC (2012a) Gene turnover and differential retention in the relaxin/insulin-like gene family in primates. Mol Phylogenet Evol 63:768

Arroyo JI, Hoffmann FG, Opazo JC (2012b) Gene duplication and positive selection explains unusual physiological roles of the relaxin gene in the European rabbit. J Mol Evol 74:52

Bathgate RA, Samuel CS, Burazin TC, Gundlach AL, Tregear GW (2003) Relaxin: new peptides, receptors and novel actions. Trends Endocrinol Metab 14:207

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

Hallstrom BM, Janke A (2010) Mammalian evolution may not be strictly bifurcating. Mol Biol Evol 27:2804

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

Jobb G, Av Haeseler, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4:18

Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537:39

Kumar S, Hedges SB (2011) TimeTree2: species divergence times on the iPhone. Bioinformatics 27:2023

Laurent A, Rouillac C, Delezoide AL, Giovangrandi Y, Vekemans M, Bellet D, Abitbol M, Vidaud M (1998) Insulin-like 4 (INSL4) gene expression in human embryonic and trophoblastic tissues. Mol Reprod Dev 51:123

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

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

Nery MF, González DJ, Hoffmann FG, Opazo JC (2012) Resolution of the laurasatherian phylogeny: evidence from genomic data. Mol Phylogenet Evol 64:685

Park JI, Chang CL, Hsu SY (2005) New insights into biological roles of relaxin and relaxin-related peptides. Rev Endocr Metab Disord 6:291

Park J-I, Semyonov J, Chang CL, Yi W, Warren W, Hsu SYT (2008) Origin of INSL3-mediated testicular descent in therian mammals. Genome Res 18:974

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572

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

Tatusova TA, Madden TL (1999) BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247

Van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O (2003) The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and therians. Mol Biol Evol 23:587

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

Yegorov S, Good S (2012) Using paleogenomics to study the evolution of gene families: origin and duplication history of the relaxin family hormones and their receptors. PLoS ONE 7: e32923