UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

MARCOS TADEU GERALDO

EVOLUÇÃO E ORGANIZAÇÃO GENÔMICA DO GENE *FOXL2* EM VERTEBRADOS

Dissertação apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, como parte dos requisitos para obtenção do título de Mestre, pelo Programa de Pós-Graduação em Ciências Biológicas (Genética).

Botucatu - SP

2012

Universidade Estadual Paulista
"Júlio de Mesquita Filho"

Instituto de Biociências de Botucatu

MARCOS TADEU GERALDO

EVOLUÇÃO E ORGANIZAÇÃO GENÔMICA DO GENE *FOXL2* EM VERTEBRADOS

Dissertação apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, como parte dos requisitos para obtenção do título de Mestre, pelo Programa de Pós-Graduação em Ciências Biológicas (Genética).

Orientador: Cesar Martins

Botucatu - SP

2012

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO DE AQUIS. E TRAT. DA INFORMAÇÃO
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ***SULAMITA SELMA CLEMENTE COLNAGO***

Geraldo, Marcos Tadeu.
Evolução e organização genômica do gene FOXL2 em vertebrados / Marcos Tadeu Geraldo. – Botucatu: [s.n.], 2012

Dissertação (mestrado) – Universidade Estadual Paulista, Instituto de Biociências de Botucatu
Orientador: Cesar Martins
Capes: 20204000

1. Evolução molecular. 2. Vertebrados.

Palavras-chave: Cópias parálogas; Duplicação gênica; Forkhead box L2; Genoma de vertebrados.

A tragédia não é quando um homem morre.
A tragédia é aquilo que morre dentro de um homem, enquanto ele ainda está vivo.

Alberto Schweitzer

AGRADECIMENTOS

Primeiramente ao meu orientador Cesar, pela grande oportunidade oferecida, pelo apoio e pela confiança a mim depositada.

Ao Programa de Pós-Graduação em Ciências Biológicas (Genética) do Instituto de Biociências de Botucatu.

Ao Departamento de Morfologia pelo apoio institucional e pelos espaços físicos.

A todos da Seção de Pós-Graduação por toda ajuda referente ao controle da documentação e solicitude para atender a todas as minhas dúvidas.

Às agências de fomento à pesquisa CNPq e FAPESP pelo apoio financeiro.

Um especial agradecimento ao Kbelo (Guilherme) por ter sido a pessoa que acompanhou de perto o desenvolvimento do meu trabalho, além dos seus ensinamentos e conselhos, que com certeza, foram fundamentais para os resultados desta dissertação.

Ao professor Antônio Sérgio (UFABC) pela colaboração, com ensinamentos imprescindíveis, conseguindo sanar muitas das minhas dúvidas e direcionando a análise de meus resultados.

A todos do Laboratório Genômica Integrativa e integrantes de outros laboratórios com quem mantenho contato, em especial ao Bruno, Eder e Taka (Rafael) pela amizade e risadas, principalmente, durante as horas do café (e foram muitas!).

Ao técnico Zé (José Eduardo Bozano) do Departamento de Morfologia pelo seu bom-humor e sua sagacidade para fazer brincadeiras que sempre animaram meus dias.

Um agradecimento imenso à minha família, em especial aos meus pais José Eduardo e Maria Helena por todo o carinho, apoio e dedicação em todos os passos de minha vida. A eles sou eternamente grato! Agradeço também aos meus irmãos Cássia e Eduardo pelo carinho, risadas e apoio durante minha vida.

Um agradecimento especial à minha namorada Natália, que entrou em minha vida durante o meu mestrado e tem sido parte fundamental de meus dias, como excelente companheira e por acreditar sempre em meu potencial.

A tudo que me aconteceu nesses dois anos de mestrado sou grato!

SUMÁRIO

Resumo **1**

Abstract **2**

1. Introdução Geral **3**

1.1. Evolução molecular e reconstrução filogenética **3**

1.2. Duplicação gênica **6**

1.3. Gene *FOXL2*: aspectos funcionais e evolutivos **10**

2. Objetivo **15**

3. Materiais e Métodos **16**

4. Resultados e Discussão: Manuscrito **17**

Abstract **20**

Introduction **21**

Materials and Methods **23**

Results and Discussion **29**

Conclusion **36**

Tables **40**

Figures **41**

Supplementary Material Captions **49**

5. Referências Bibliográficas **51**

RESUMO

FOXL2 representa um importante gene, membro da família do domínio *forkhead*, responsável principalmente pelo desenvolvimento de ovários durante a determinação e diferenciação sexual de fêmeas. Os estudos evolutivos realizados anteriormente foram limitados considerando o número de cópias gênicas e unidades taxonômicas analisadas. Para analisar um grande número de sequências de *FOXL2* e contribuir para um cenário mais claro da história evolutiva do gene, uma ampla análise evolutiva do domínio *forkhead* e do gene *FOXL2* foi realizada. Um total de 2.464 sequências para o domínio *forkhead* foram recuperadas e, subsequentemente, 62 sequências representativas para o *FOXL2* foram utilizadas na análise evolutiva do gene. A contribuição mais importante deste estudo foi a descoberta de um novo subgrupo de ortólogos de *FOXL2* (parálogos para outros *FOXL2*) presente no celacanto *Latimeria chalumnae* (celacanto do Oeste do Oceano Índico), na ave *Taeniopygia guttata* (tentilhão) e no marsupial *Monodelphis domestica* (gambá). Este novo cenário indica um evento de duplicação gênica em um ancestral de vertebrados ósseos (Euteleostomi). Com base nas análises das regiões sintênicas de ambas as cópias de *FOXL2*, o evento de duplicação não ocorreu em um único gene, mas em um bloco de genes. Além disso, a distribuição da cópia duplicada se mostrou complexa entre os vertebrados, especialmente em tetrápodas, e os resultados apoiam fortemente uma perda desta cópia nas espécies de euterianos. Finalmente, o cenário observado no presente estudo sugere uma atualização para a nomenclatura do gene *FOXL2*, estendendo a nomenclatura de *FOXL2A* e *FOXL2B*, sugerida para teleósteos, para todas as sequências de vertebrados, contribuindo, portanto, para o estabelecimento de um novo contexto evolutivo para o gene *FOXL2*.

Palavras-chave: *Forkhead box L2*, genoma de vertebrados, cópias parálogas, duplicação gênica

ABSTRACT

The *FOXL2* gene is an important member of the forkhead domain family, primarily responsible for the development of ovaries during female sex determination and differentiation. The evolutionary studies conducted previously were limited considering the number of gene copies and taxonomic units analyzed. To analyze a large number of sequences for *FOXL2* and contribute to a clearer picture of the evolutionary history of the gene, a broad evolutionary analysis of the forkhead domain and *FOXL2* was performed. A total of 2,464 sequences for the forkhead domain were recovered, and subsequently, 62 representative sequences for *FOXL2* were used in the evolutionary analysis of this gene. The most important contribution of this study was the discovery of a new subgroup of *FOXL2* orthologs (paralogs to other *FOXL2* genes) observed in the coelacanth *Latimeria chalumnae* (West Indian Ocean coelacanth), in the avian *Taeniopygia guttata* (zebra finch) and in the marsupial *Monodelphis domestica* (opossum). This new scenario indicates a gene duplication event in an ancestor of bony vertebrates (Euteleostomi). Furthermore, based on the analysis of the syntenic regions of both *FOXL2* copies, the duplication event did not occur in a single gene but in a block of genes. Moreover, the duplicated copy distribution was shown to be complex across vertebrates, especially in tetrapods, and the results strongly support a loss of this copy in eutherian species. Finally, the scenario observed in this study suggests an update for *FOXL2* gene nomenclature, extending the actual suggested teleost naming of *FOXL2A* and *FOXL2B* to all vertebrate sequences and contributing to the establishment of a new evolutionary context for the *FOXL2* gene.

Key words: Forkhead box L2, vertebrate genome, paralog copies, gene duplication

1. INTRODUÇÃO GERAL

1.1. Evolução molecular e reconstrução filogenética

A evolução molecular é uma área de estudos que surgiu e avançou muito a partir do acúmulo e disponibilidade de dados genéticos de sequências de nucleotídeos e proteínas, dos avanços computacionais e ainda do desenvolvimento de métodos estatísticos sofisticados (Yang 2006). Os estudos de evolução molecular têm como principal objetivo a reconstrução das relações evolutivas entre genes e espécies, além da investigação das forças e mecanismos do processo evolutivo (Yang 2006).

A filogenia é a história evolutiva de um grupo de genes ou organismos e o principal objetivo de sua inferência é descrever as relações evolutivas em termos de uma relação de ancestralidade comum, sendo normalmente representada por um diagrama de ramos (árvore filogenética) (Felsenstein 2004). Dentro de uma árvore filogenética podem ser observados três tipos de relações, definidas por Hennig (1966): monofilia, parafilia e polifilia. Os grupos considerados como monofiléticos e parafiléticos são aqueles que possuem uma única origem evolutiva, e neste contexto, os grupos monofiléticos incluem todos os descendentes de um único ancestral. No entanto, se um desses descendentes é retirado, os grupos restantes são denominados como parafiléticos. Já os grupos polifiléticos representam aqueles considerados como resultado de evolução convergente, e os caracteres que apoiam o grupo estão ausentes no ancestral comum mais recente.

Durante a reconstrução filogenética com base em dados moleculares, um pressuposto básico é de que as sequências utilizadas sejam homólogas, podendo ser sequências tanto de nucleotídeos como de aminoácidos, porém, a escolha do tipo de grupo de dados não é consensual. Em alguns casos, a utilização de dados de aminoácidos é preferida para inferir filogenia principalmente porque existem mais estados possíveis para os caracteres (20

estados), em oposição aos nucleotídeos (4 estados) (revisado por Simmons, Ochoterena, Freudenstein 2002). Contudo, isto é contraposto com o fato de que o aumento no número de caracteres em sequências de nucleotídeos pode conduzir a uma melhor resolução da árvore. Um exemplo disso é quando a família de genes de interesse tem apenas um domínio altamente conservado e pouco divergente nas sequências de aminoácidos (revisado por Harrison e Langdale 2006).

Para a reconstrução filogenética, independentemente do tipo de sequência utilizada, é preferível buscar o maior número possível de sequências similares, a fim de reduzir ou evitar associações imprecisas entre grupos terminais (Hillis 1996). Esta busca pode ser realizada acessando bancos de dados genéticos privados ou públicos. No caso dos bancos públicos, um dos mais conhecidos é o GenBank, alocado no NCBI – *National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov). Dentre as ferramentas de busca por sequências similares, uma das mais (senão a mais) utilizada é o programa BLAST (*Basic Local Alignment Search Tool*), desenvolvido por Altschul et al. (1990). O BLAST compõe um conjunto de programas que foram derivados do algoritmo original, entre eles: BLASTN, BLASTP, BLASTX, TBLASTN e TBLASTX. Desses programas, a amostragem de sequências pode ser otimizada usando as opções TBLASTN e TBLASTX, já que em ambos os casos, as sequências de nucleotídeos recuperadas são traduzidas nas seis matrizes de leitura, maximizando, portanto, o potencial para a recuperação de sequências similares ao gene ou proteína de interesse. Um parâmetro importante para ser analisado ao se recuperar sequências por meio de BLAST, é o valor E (Valor Esperado), que é uma indicação do grau de similaridade entre a sequência inicial usada nas pesquisas e a sequência recuperada, sendo que, quanto mais próximo o valor E é de zero, maior o grau de similaridade entre as duas sequências (revisado por Lemey, Salemi, Vandamme 2009).

Depois que os dados de sequências foram recuperados e cuidadosamente alinhados, existem tipos diferentes de análises filogenéticas que podem ser implementadas. De forma simplificada, os principais métodos de inferência filogenética, com base na revisão de Harrison e Langdale (2006), são:

a) Métodos de distância (e.g. agrupamento de vizinhos e evolução mínima): calculam as distâncias entre os pares de sequências e agrupam aquelas que são mais similares. Computacionalmente, esta abordagem requer baixo nível de processamento, devido a uma maior simplicidade para realizar os cálculos e, portanto, a análise é conduzida com alta rapidez. Entretanto, os métodos de distância não permitem analisar quais caracteres contribuíram para grupos particulares. Tal como acontece com outros métodos, o resultado pode depender da ordem em que as entidades são adicionadas à árvore gerada inicialmente.

b) Máxima parcimônia: assume que a melhor árvore para explicar a evolução do conjunto de dados é aquela em que há menor número de alterações possíveis (maior parcimônia). Um problema neste método surge quando árvores com agrupamentos diferentes são igualmente consideradas como plausíveis, dificultando, ou até mesmo, impedindo a geração de uma árvore-consenso final.

c) Máxima verossimilhança: calcula a probabilidade de uma determinada árvore se encaixar no conjunto de dados, levando em consideração, um modelo específico de evolução. A análise começa com uma árvore inicial, que pode ser gerada por algum método de distância, como agrupamento de vizinhos. Posteriormente, ocorrem trocas nos ramos até que a árvore com o maior valor de verossimilhança e, portanto, com maior probabilidade de ajuste aos dados, é adquirida. Este valor é uma função tanto da topologia da árvore como dos comprimentos dos ramos (número de mudanças de estado dos caracteres). Métodos de máxima

verossimilhança são as técnicas mais exigentes computacionalmente para análise filogenética, porém, permitem uma análise mais explícita das suposições feitas sobre a evolução das sequências.

d) Inferência Bayesiana: é outro método de verossimilhança, no entanto, a diferença é que um conjunto distinto de pressupostos (*priors*) é introduzido no modelo original. A reconstrução filogenética por inferência Bayesiana é baseada na distribuição da verossimilhança e da probabilidade posterior das árvores, utilizando algoritmos MCMC (*Markov Chain Monte Carlo*) (revisado por Lemey, Salemi, Vandamme 2009). Neste método, o tempo computacional requerido é bem maior que dos outros métodos em virtude da geração de um conjunto extenso de árvores.

1.2. Duplicação gênica

A grande diversidade morfológica e de espécies presente nos eucariotos tem sido atribuída, frequentemente, aos eventos de duplicação de genes, cromossomos ou genomas, pois são considerados como as principais fontes para o surgimento de novos genes (Ohno 1970; Holland et al. 1994; Long et al. 2003; Semon e Wolfe 2007; Conant e Wolfe 2008). De forma simplificada, a duplicação gênica pode ocorrer a partir de três processos principais: (a) *crossing-over* desigual, (b) retrotransposição, ou (c) não disjunção dos cromossomos ou cromátides-irmãs durante a meiose (revisado por Freeman e Herron 2009).

A duplicação gênica por *crossing-over* desigual ocorre quando os cromossomos não se pareiam adequadamente e, consequentemente, o *crossing-over* ocorre em porções não-homólogas dos cromossomos pareados. Agora, durante a retrotransposição, um RNA-mensageiro (mRNA) é retrotranscrito, gerando um DNA complementar (cDNA) que

posteriormente é inserido em algum ponto do genoma (revisado por Zhang 2003). Dessa forma, se a região retrotranscrita continha algum gene, esse será duplicado, total ou parcialmente, dependendo da cobertura da região retrotranscrita.

Por último, as duplicações cromossômicas e genômicas ocorrem possivelmente pela não-disjunção cromossômica na meiose I ou pela não-disjunção das cromátides-irmãs na meiose II (revisado por Freeman e Herron 2009).

A importância evolutiva da duplicação em larga-escala está atrelada ao aumento significativo do material genético, aumentando a chance de que novos genes surjam e adquiram novas funções (revisado por Freeman e Herron 2009). Dessas duplicações em larga-escala, a duplicação de um genoma como um todo tem sido reportada em animais e fungos, e é particularmente frequente nas plantas (Wolfe e Shields 1997; Bowers et al. 2003; Kellis, Birren, Lander 2004; Dehal e Boore 2005; Jaillon, Aury, Wincker 2009). Mesmo considerada menos frequente em animais, a duplicação do genoma tem sido normalmente associada com importantes transições, especialmente em cordados, em que estudos reportam a ocorrência de duplicação de todo o genoma coincidindo com o surgimento dos grupos de vertebrados, gnatostomados e teleósteos (Holland et al. 1994; Meyer e Schartl 1999; Postlethwait et al. 2000; Hoegg et al. 2004; Crow et al. 2006).

Apesar da importância associada às duplicações gênicas, ainda existe um grande debate sobre os detalhes evolutivos de como novos genes surgem a partir de uma duplicação (Zhang 2003; He e Zhang 2005; Li, Yang, Gu 2005).

Evidência de duplicação gênica foi reportada já no início do século passado, em 1911, em um trabalho de Kuwada com milho (*Zea mays*), em que o pesquisador sugeriu a existência de cromossomos originados de um evento de duplicação e propôs que a geração de diferentes linhagens pode ser atribuída ao referido evento (revisado por Taylor e Raes 2004).

Posteriormente, Bridges (1935) atribuiu às duplicações como um processo capaz de gerar cromossomos maiores compostos por genes iguais, o que possibilitaria padrões de mutação diferenciada entre esses genes e, consequentemente, podendo gerar efeitos funcionais distintos. Além disso, Serebrowsky (1938) propôs que a seleção negativa pode estar reduzida em genes que se apresentem duplicados. O autor também concluiu, ao caracterizar *achaete* e *scute* (genes ligados ao cromossomo X de *Drosophila melanogaster*), que ambas as cópias podem ser alteradas, pois um gene pode influenciar diferentes aspectos do fenótipo e, por conseguinte, após a duplicação poderia ocorrer uma distribuição de funções entre as cópias duplicadas.

Claramente, estes trabalhos iniciais forneceram os primeiros indícios do que atualmente é conhecido como os processos de neofuncionalização e subfuncionalização para as cópias gênicas duplicadas. No entanto, o tema duplicação gênica somente começou a ser intensamente debatido a partir do trabalho de Ohno (1970) publicado no livro *Evolution by gene duplication*. Neste trabalho, Ohno definiu o modelo de neofuncionalização, apesar do termo propriamente dito ter sido cunhado posteriormente por Force et al. (1999). Segundo Ohno, após a duplicação de um gene, uma das cópias manteria a função original do gene ancestral, enquanto a cópia adicional seria redundante e, se fixada por deriva genética, estaria livre de seleção negativa. Posteriormente, a cópia adicional poderia se tornar um pseudogene, ou então, ser perdida pela acumulação de mutações neutras com perda de função gênica. Em casos menos frequentes, durante a acumulação de mutações, a cópia gênica adicional poderia adquirir uma nova função, que seria mantida por seleção. Um exemplo muito referido na literatura é o caso das glicoproteínas *antifreeze* dos peixes antárticos originadas por duplicação do gene do tripsinogênio, em que a cópia duplicada se neofuncionalizou, e a proteína resultante se tornou capaz de se ligar a cristais de gelo, prevenindo o peixe do congelamento (Deng et al. 2010).

A subfuncionalização, por sua vez, mesmo depois do trabalho de Serebrowsky (1938), somente teve sua definição melhor compreendida com o trabalho de Force et al. (1999), em que o modelo proposto de subfuncionalização foi denominado de Duplicação-Degeneração-Complementação (DDC). Este modelo, da mesma forma que a neofuncionalização proposta por Ohno, é baseado na redundância das cópias duplicadas. Desta forma, no modelo DDC, haveria o acúmulo de mutações em ambas as cópias (mutações neutras com degeneração da função gênica) conduzindo a uma redução na eficácia funcional das cópias duplicadas. Assim, caso ocorra a manutenção de ambas as cópias por deriva genética, nenhuma delas seria capaz de desempenhar a função gênica original de forma independente, portanto, ambas as cópias necessariamente seriam preservadas por seleção a fim da manutenção do *fitness*.

É importante mencionar que os processos de neofuncionalização e subfuncionalização não são mutuamente excludentes. Como exemplo, as formas duplicadas podem evoluir por subfuncionalização em termos de estrutura de proteína, mas por neofuncionalização em termos de perfis de expressão (Tirosh e Barkai 2007). Dessa forma, é possível perceber que as estratégias evolutivas desempenhadas pelas cópias duplicadas são diversificadas, em que diferentes modelos de evolução podem ser sugeridos e agrupados, como pode ser observado na revisão de Innan e Kondrashov (2010). Por exemplo, há modelos em que a duplicação em si é vantajosa, como no caso em que o *fitness* é incrementado pelo aumento na dosagem gênica. Neste caso, se esse aumento for vantajoso, então o gene duplicado pode ser fixado por seleção positiva e a manutenção desse gene será condicionada com a conservação das condições que favoreceram a sua seleção.

Dentro do contexto da evolução gênica, diferentes relações de homologia podem ser definidas de acordo com o evento de geração das cópias a partir de um ancestral comum (Figura 1). Quanto à origem, a homologia entre genes pode ser diferenciada quando cópias são originadas por um evento de especiação (ortologia, *orto* = igual, preciso) ou por um

evento de duplicação (paralogia, *para* = em paralelo) (Fitch 1970). Posteriormente, outras relações de homologia surgiram para referenciar casos específicos: onologia (em homenagem a Susumu Ohno) no caso de genes parálogos que se originaram de um evento de duplicação do genoma (Wolfe 2000) e xenologia para genes homólogos resultantes de transferência horizontal (Gray e Fitch 1983).

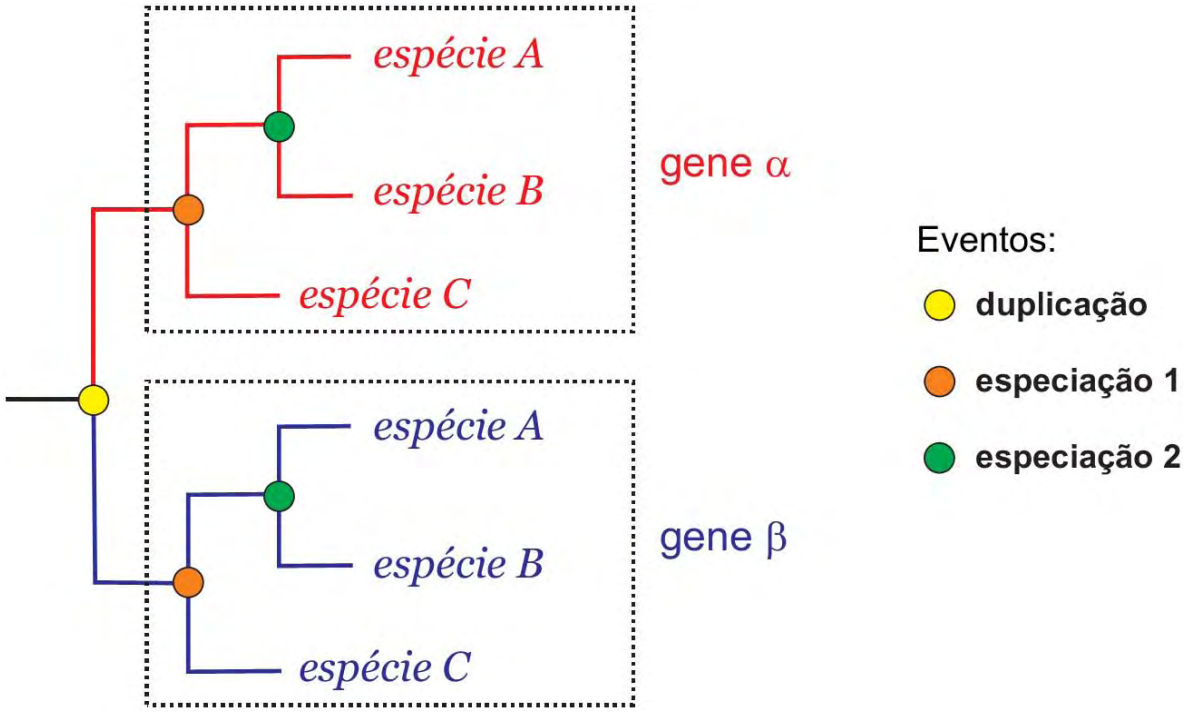


Figura 1: Esquema da caracterização da paralogia e ortologia, com base nos respectivos eventos de geração das sequências (duplicação e especiação). Sequências originadas de um evento de duplicação (e.g. gene α da espécie A e gene β da espécie A = genes parálogos) e sequências originadas de um evento de especiação (e.g. gene α da espécie A e gene α da espécie B = genes ortólogos).

1.3. Gene *FOXL2*: aspectos funcionais e evolutivos

O gene *FOXL2* foi identificado pela primeira vez em 1998 ao se buscar genes envolvidos com o desenvolvimento da glândula pituitária em camundongos (Treier et al. 1998), sendo considerado membro da família de fatores de transcrição de ligação ao DNA

compostos pelo domínio denominado *forkhead* ou *winged helix* (Weigel et al. 1989; Mahlapuu et al. 2001; Lehmann et al. 2003). O domínio, que segundo a determinação no banco de dados Pfam apresenta tamanho de 96 aminoácidos (Finn et al. 2010), foi identificado a partir da determinação da estrutura cristalográfica da família HNF-3γ/FOXA, envolvida no desenvolvimento dos tecidos de pâncreas e fígado (Clark et al. 1993). Muitas proteínas da família *forkhead* são consideradas como reguladores do desenvolvimento (Bravieri et al. 1997), como desenvolvimento crânio-faringeal – FOXE1 (Lehmann et al. 2003), crescimento celular e resposta à insulina – FOXO1 (Gross, van den Heuvel, Birnbaum 2008), formação de pêlo e diferenciação dos queratinócitos – FOXN1 (Nehls et al. 1994), e formação e funcionamento do ovário – FOXL2 (Cocquet et al. 2003; Baron et al. 2004; Uhlenhaut e Treier 2006; Veitia 2010; Jaubert et al. 2011). A descoberta dos papéis de cada membro da família de proteínas foi realizada por meio de estudos envolvendo defeitos congênitos associados com mutações, normalmente observadas na região do domínio *forkhead* (Carlsson e Mahlapuu 2002).

O gene *FOXL2* está presente em diferentes grupos de metazoários, porém, sua funcionalidade e expressão se encontram melhor compreendidos em vertebrados. O papel do gene *FOXL2* em vertebrados é fundamental tanto para o desenvolvimento como para a maturação e posterior funcionamento ovariano (Uhlenhaut et al. 2009). A expressão de *FOXL2* é fêmea-específica e pode ser detectada em estágios iniciais do desenvolvimento gonadal (Cocquet et al. 2002). Este padrão de expressão dimórfico entre os sexos é conservado entre os diferentes grupos de vertebrados, atuando na formação do ovário, particularmente, com a diferenciação em células granulosas (Guerrero-Estévez e Moreno-Mendoza 2009). Dessa forma, o gene *FOXL2* (*Forkhead box L2*) se encontra como um dos importantes candidatos à determinação e diferenciação sexual de vertebrados (Baron et al. 2004).

Estudos mais recentes em mamíferos associam determinadas mutações em *FOXL2* com a Síndrome Blefarofimose-Ptose-Epicanto Inverso (BPES), caracterizada pela má-formação das pálpebras (BPES tipo II) ou, em alguns casos, associada à falência prematura do ovário (BPES-tipo I) (Crisponi et al. 2001; De Baere et al. 2001). A ocorrência de algumas dessas mutações foram descritas numa região exclusiva do gene de mamíferos caracterizada por ser rica em alaninas (14 alaninas) (Figura 2). Moumne et al. (2008) observou que expansões de 10 alaninas (total de 24 alaninas) foram identificadas em aproximadamente 30% dos pacientes com BPES e são a principal causa da BPES tipo II.

Figura 2: Representação da sequência tipo selvagem da proteína FOXL2 de humanos (número de acesso NP_075555). Em destaque o domínio *forkhead* (vermelho), determinado segundo o banco de dados Pfam (Finn et al. 2010) e a região rica em poli-alaninas (azul) exclusiva de mamíferos. Figura gerada pelo *software* Geneious 4.8.5 (Drummond et al. 2009).

Uhlenhaut et al. (2009) mostraram, em camundongos, que *FOXL2* é requerido para prevenir a transdiferenciação das células foliculares de um ovário adulto em células semelhantes às testiculares. Segundo os autores, a perda somente de *FOXL2* parece ser suficiente para iniciar a conversão completa do ovário em um testículo. Além disso, estudos da cinética da transdiferenciação sugerem que *FOXL2* e outro gene chamado *SOX9* (*SRY-box 9*), envolvido na via de formação de testículos, são mutuamente excludentes em mamíferos. A possível exclusão mútua entre esses genes pode representar um modelo ativo de alternância, na qual dois genes centrais inibem um ao outro, direta ou indiretamente, conduzindo a dois estados ou destinos alternativos. Consequentemente, *FOXL2* representaria um repressor na via de masculinização durante o desenvolvimento gonadal feminino (Uhlenhaut et al. 2009).

Sobre o aspecto evolutivo do gene *FOXL2*, os trabalhos se limitam a poucas espécies de vertebrados, portanto, num contexto filogenético bem mais restrito, e os resultados mais interessantes até o momento se referem à descrição de cópias parálogas em teleósteos. O primeiro trabalho a relatar tais cópias foi o de Baron et al. (2004) ao identificar a presença de uma cópia extra com similaridade ao *FOXL2* nas espécies de peixes *Oncorhynchus mykiss* (truta-arco-íris), *Takifugu rubripes* (fugu) e *Tetraodon nigroviridis* (baiacu-verde-pintado). A filogenia dessas sequências demonstrou uma topologia com dois grupos monofiléticos bem estabelecidos: um clado com a cópia duplicada de teleósteos (denominada de *FOXL2B*) e outro com a cópia original (*FOXL2* de tetrápodas e *FOXL2A* de teleósteos), sendo esta topologia típica de cópias duplicadas. Outro trabalho também identificou cópias duplicadas nos peixes *Danio rerio* (paulistinha), *Gasterosteus aculeatus* (esgana-gata) e *Oryzias latipes* (medaka) (Jiang et al. 2011). A filogenia destas cópias também resultou em dois clados distintos para as cópias *FOXL2A* e *FOXL2B*. Para os autores, a identificação de cópias parálogas nessas espécies estaria em concordância com o evento de duplicação de genoma ocorrido exclusivamente em teleósteos. No entanto, nenhum destes trabalhos realizou uma

busca minuciosa e extensa de possíveis cópias duplicadas de *FOXL2* em outros grupos de vertebrados.

Até o momento, muito pouco se sabe sobre a função da cópia *FOXL2B* em teleósteos. Em um estudo de expressão envolvendo a espécie de teleósteo *Oncorhynchus mykiss*, ambas as cópias parálogas se expressaram em ovário, no entanto, *FOXL2B* se expressou mais tardiamente, e essa diferença combinada com divergências nas sequências entre *FOXL2A* e *FOXL2B* indica, segundo os autores, uma possível neofuncionalização, que possibilitou a manutenção da cópia duplicada em teleósteos (Baron et al. 2004). Em outro trabalho, neste caso envolvendo a espécie de teleósteo *Salmo salar* (salmão) (von Schalburg et al. 2011) relatou-se que ambas as cópias são expressas bem no início do desenvolvimento, sendo que em machos a expressão de *FOXL2A* se mostrou restrita em comparação a *FOXL2B*, que ao contrário, se mostrou presente na maioria dos tecidos extra-gonadais analisados. Em contraste, a expressão de *FOXL2B* foi restrita aos tecidos de fêmeas adultas, essencialmente confinada em pele, brânquias e ovário.

Devido aos poucos estudos evolutivos envolvendo *FOXL2*, associado à importância desse gene nos processos de determinação e diferenciação sexual de vertebrados e, em face do recente contexto com a presença de cópias duplicadas de *FOXL2* em teleósteos, o presente estudo buscou proporcionar um melhor entendimento em relação ao cenário evolutivo do gene *FOXL2* em diferentes grupos de vertebrados.

2. OBJETIVO

O objetivo do presente estudo foi compreender a evolução e organização genômica do gene *FOXL2*, com enfoque na distribuição e origem das cópias duplicadas no contexto evolutivo de vertebrados.

3. MATERIAIS E MÉTODOS

Os materiais e métodos adotados nesta dissertação encontram-se reunidos no manuscrito apresentado no tópico *Resultados e Discussão*, página 23.

4. RESULTADOS E DISCUSSÃO

Os resultados e as discussões desta dissertação são apresentados na forma de manuscrito submetido à revista *Molecular Biology and Evolution*.

The discovery of *FOXL2* paralogs in coelacanth and tetrapod genomes reveals an ancient duplication in vertebrates

Research Article

Marcos Tadeu Geraldo[1], Guilherme Targino Valente[1], Antônio Sérgio Kimus Braz[2] and Cesar Martins[1*]

[1] Integrative Genomics Laboratory, Department of Morphology, Institute of Biosciences, Sao Paulo State University – UNESP, Botucatu, SP, 18618-000, Brazil.

[2]Laboratory of Computational Biology and Bioinformatics, Center of Natural and Human Sciences, Federal University of ABC – UFABC, Santo Andre, SP, 09210-170, Brazil

***Corresponding Author:** C Martins, Institute of Biosciences, Department of Morphology, Sao Paulo State University – UNESP, Botucatu, SP, 18618-000, Brazil. Phone/Fax: ++55(14)38800462; E-mail: cmartins@ibb.unesp.br

Key words: Forkhead box L2, vertebrate genome, paralog copies, gene duplication

Running Head: Ancient duplication of *FOXL2* in vertebrates

List of nonstandard abbreviations

aLRT: approximate likelihood ratio test, **BPES**: blepharophimosis, ptosis, epicanthus inversus syndrome, **CAB39**: calcium-binding protein 39, **CDS**: coding sequence, **CLSTN1**: calsyntenin-1, **CLSTN2**: calsyntenin-2, **COPB2:** coatomer protein complex subunit beta 2, **CTNNBIP1:** beta-catenin-interacting protein 1, **FOXA**: forkhead box A, **FOXE1**: forkhead box E1, **FOXL2A:** isoform A of the forkhead box L2, **FOXL2B**: isoform B of the forkhead box L2, **FOXN1**: forkhead box N1, **FOXO1**: forkhead box O1, **HES6**: hairy and enhancer of split 6, **HMM**: hidden markov model, **HNF-3γ**: hepatocyte nuclear factor 3-gamma, **ITM2C**: integral membrane protein 2C, **iTOL**: interactive Tree of Life, **KLHL17**: Kelch-like protein 17, **LZIC**: leucine zipper and CTNNBIP1 domain-containing protein, **MRPS22**: mitochondrial 28S ribosomal protein S22, **NMNAT1**: nicotinamide nucleotide adenylyltransferase 1, **NMNAT3**: nicotinamide nucleotide adenylyltransferase 3, **NNI**: nearest neighbor interchange, **NOC2L**: nucleolar complex protein 2 homolog, **ORF**: open reading frame, **PER2**: period circadian protein homolog 2, **PIK3CB**: beta isoform of the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit, **PIK3CD**: delta isoform of the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit, **PLEKHN1**: pleckstrin homology domain-containing family N member 1, **RBP2**: retinol-binding protein 2, **RBP7**: retinol-binding protein 7, **SAMD11**: sterile alpha motif domain-containing protein 11, **SLC25A33**: solute carrier family 25 member 33, **SPR**: subtree pruning and regrafting, **TMEM201**: transmembrane protein 201, **UBE4B**: ubiquitin conjugation factor E4 B.

Abstract

The *FOXL2* gene is an important member of the forkhead domain family, primarily responsible for the development of ovaries during female sex determination and differentiation. The evolutionary studies conducted previously were limited considering the number of gene copies and taxonomic units analyzed. To analyze a large number of sequences for *FOXL2* and contribute to a clearer picture of the evolutionary history of the gene, a broad evolutionary analysis of the forkhead domain and *FOXL2* was performed. A total of 2,464 sequences for the forkhead domain were recovered, and subsequently, 62 representative sequences for *FOXL2* were used in the evolutionary analysis of this gene. The most important contribution of this study was the discovery of a new subgroup of *FOXL2* orthologs (paralogs to other *FOXL2* genes) observed in the coelacanth *Latimeria chalumnae* (West Indian Ocean coelacanth), in the avian *Taeniopygia guttata* (zebra finch) and in the marsupial *Monodelphis domestica* (opossum). This new scenario indicates a gene duplication event in an ancestor of bony vertebrates (Euteleostomi). Furthermore, based on the analysis of the syntenic regions of both *FOXL2* copies, the duplication event did not occur in a single gene but in a block of genes. Moreover, the duplicated copy distribution was shown to be complex across vertebrates, especially in tetrapods, and the results strongly support a loss of this copy in eutherian species. Finally, the scenario observed in this study suggests an update for *FOXL2* gene nomenclature, extending the actual suggested teleost naming of *FOXL2A* and *FOXL2B* to all vertebrate sequences and contributing to the establishment of a new evolutionary context for the *FOXL2* gene.

Introduction

The forkhead domain, also called the winged helix domain, is present in a large number of proteins that constitute a family of transcription factors that activate important pathways of embryogenesis (Weigel et al. 1989; Mahlapuu et al. 2001; Lehmann et al. 2003) and cell differentiation (Brissette et al. 1996; Dottori et al. 2001; Nakae et al. 2003). This domain, determined by the Pfam database (Finn et al. 2010) as including 96 amino acids, was named after the description of the crystal structure of the HNF-3γ/FOXA family, which is involved in pancreas and liver tissue development (Clark et al. 1993). Many proteins of the forkhead family are considered to be tissue-specific regulators of development (Bravieri et al. 1997). Examples include craniopharyngeal development – FOXE1 (Lehmann et al. 2003), cell growth and insulin responsiveness – FOXO1 (Gross, van den Heuvel, Birnbaum 2008), hair formation and keratinocyte differentiation – FOXN1 (Nehls et al. 1994), and ovarian formation and function – FOXL2 (Cocquet et al. 2002; Baron et al. 2004; Uhlenhaut and Treier 2006; Veitia 2010; Jaubert et al. 2011). The role of each protein member of the domain family was discovered primarily by studies of congenital defects associated with mutations, often observed within the forkhead domain (Carlsson and Mahlapuu 2002).

The *FOXL2* (Forkhead box L2) gene is an important member of this extensive family and is primarily responsible for sex determination and differentiation, specifically ovarian development and maturation (Uhlenhaut et al. 2009). The majority of studies involving *FOXL2* describe functional differences of mutations associated with Blepharophimosis Ptosis Epicanthus Inversus Syndrome (BPES), which is characterized by eyelid malformations (BPES type II) or, in some cases, premature ovarian failure (BPES type I) (Crisponi et al. 2001; De Baere et al. 2001). An exclusive feature of the protein, observed only in mammals, is a tract of 14 alanines, which has been suggested to be a region with strong functional

constraints (Cocquet et al. 2003). For instance, expansions of 10 alanines represent approximately 30% of the causes of BPES type II (Moumne et al. 2008).

Even though there are vast amounts of data available from clinical and expression analysis of *FOXL2*, only a few studies, restricted to a reduced set of vertebrate species from teleosts to mammals, are focused on the evolution of the gene. To date, the most interesting observation in the evolutionary analysis is the presence of paralogs in the fish group, the origin of which was suggested to be in accordance with the additional fish-specific whole genome duplication (Jiang et al. 2011). An extra copy of the *FOXL2* gene has only been reported in *Onchorrynchus mykiss* (rainbow trout), *Takifugu rubripes* (fugu), *Tetraodon nigroviridis* (pufferfish) (Baron et al. 2004), *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (stickleback), *Oryzias latipes* (medaka) (Jiang et al. 2011) and *Salmo salar* (Atlantic salmon) (von Schalburg et al. 2011).

Motivated by the previous evolutionary context of *FOXL2*, the present work was conducted to improve the robustness of *FOXL2* evolutionary history and reconstruct a global domain phylogeny in an attempt to orient the posterior gene phylogeny, which, for the first time, includes a large number of species. Evolutionary descriptions of new *FOXL2* single and duplicated copies were obtained for different groups of vertebrates, including the coelacanth (*L. chalumnae*), the bird (*T. guttata*) and the marsupial (*M. domestica*). Moreover, single copies of *FOXL2* from the neotropical cichlid species *Cichla monoculus* (peacock bass), and the chondrichthyan species *Rhizoprionodon lalandei* (Brazilian sharpnose shark) and *Callorhinchus callorynchus* (Plownose chimaera) were sequenced. The inclusion of sequences from the basal vertebrates, shark and chimaera, was important for developing a broader evolutionary analysis for *FOXL2*. Furthermore, the syntenic region analyses of both *FOXL2* copies in different vertebrate species were discussed together with the phylogenetic results to support the present conclusions about the duplication event of *FOXL2*.

Contrary to the previous report on the origin of paralog copies of *FOXL2*, this work shows an ancient and nonspecific teleostei origin of the paralog copies in ancestors of bony vertebrates, leading to a suggestion for a new nomenclature for those genes, namely, *FOXL2A* (the most studied gene form, generally known as *FOXL2*) and *FOXL2B* (for the diverged duplicated form) in vertebrates. Finally, it is suggested that *FOXL2B* probably could have undergone a neofunctionalization process.

Materials and Methods

The methodology adopted in the present work is summarized in Figure 1.

Sequence acquisition and alignment procedures

Forkhead domain sequences

To build an overview of the evolutionary history of the *FOXL2* gene, protein sequences of the forkhead family were collected from a large number of eukaryotes including fungi, plants and animals. A complete list of GenBank accession numbers, available in the Pfam database (Finn et al. 2010) for the forkhead family, was used to retrieve the corresponding nucleotide and amino acid sequences by means of gbreader software (Braz ASK et al. unpublished data). This procedure allowed for the acquisition of 2,464 sequences of all available forkhead family proteins in GenBank (Supplementary Material 1).

To avoid redundancy related to highly similar sequences, and thus the use of only a set of representative data, a clustering methodology was performed using the CD-HIT software (Li W and Godzik 2006). This program clustered proteins with identity ≥ 90% (the threshold used in the present work) and retrieved only one representative protein from each cluster. This threshold value was applied because the forkhead family exhibits a high conservation level in the domain region, and thus, lower values would indicate the loss of important representative

data. The previous analysis resulted in a set of sequences (1,022 sequences relative to 370 eukaryotic species) available in the Pfam database (Finn et al. 2010) that were aligned with the HMMER v. 3 software (Finn, Clements, Eddy 2011) using a Hidden Markov Model (HMM) profile of the forkhead domain. The final alignment was edited in GeneDoc (Nicholas, Nicholas Jr, Deerfield 1997) and Seaview (Gouy, Guindon, Gascuel 2010) software to select only the conserved regions of the domain and avoid poorly aligned portions (Supplementary Material 2).

FOXL2 sequences

Because our focus is on FOXL2, in addition to the sequences presented in the aforementioned dataset, other FOXL2 sequences were obtained from the Ensembl (Flicek et al. 2011), BouillaBase (http://BouillaBase.org) and JGI (http://genome.jgi.doe.gov/) databases (access from December/2011 to March/2012). Only full sequences annotated as FOXL2 were retrieved in the first search; however, deep searches were also conducted in all metazoan genomes available in these databases to detect additional, unreported duplicated copies. A tblastn search was performed using FOXL2 sequences from species closely related to the target species as queries. Because the resulting hits were centered only in the forkhead domain, the retrieval was expanded across the hit to obtain the complete *FOXL2* coding sequence (CDS). These sequences were submitted to an ORF (open reading frame) search and protein translation using Geneious 4.8.5 software (Drummond et al. 2009). Moreover, other FOXL2 sequences were retrieved by a blastp search directly from the GenBank database, available at NCBI (http://www.ncbi.nlm.nih.gov/). The sequences with the description annotation of FOXL2 that were not included in the list of accession numbers of the Pfam database were selected to be incorporated into the evolutionary analysis. A complete list of

FOXL2 retrieved sequences and details of the database searches are shown in Supplementary Material 3.

Experimental procedures were also employed to retrieve *FOXL2* sequences for three organisms, the chondrichthyan species *R. lalandei* and *C. callorynchus*, and the neotropical cichlid *C. monoculus*. The tissue samples were available at the Laboratory of Integrative Genomics of Sao Paulo State University, and the genomic DNA was extracted from muscle and liver tissues using the phenol-chloroform method (Sambrook and Russel 2001). The *FOXL2* sequences were amplified by PCR (polymerase chain reaction) using degenerate primers (F - 5' GTN GCN YTN ATH GCN ATG GC 3' and R - 5' CCA RTA NSW RCA RTG CAT CAT 3') constructed with the Primer3Plus software (Untergasser et al. 2007). For the construction of these primers, several FOXL2 protein sequences of vertebrates (data not shown) were retrieved based on a blastp search in the Expasy Proteomic Server (http://ca.expasy.org/) using a FOXL2 sequence from *Oreochromis niloticus* (accession number Q6JA05) as the query. These sequences were aligned using ClustalW (Thompson, Higgins, Gibson 1994), and the primers were constructed for the most conserved regions. The amplicons were ~ 250 bp, and the full gene sequences were obtained by the genome walking technique using Genome Walker™ kits (Clontech) according to the manufacturer's protocol. All amplicons were cloned in p-GEM-T plasmid vectors (Promega), and the corresponding clones were sequenced with an ABI Prism 3100 DNA sequencer (Perkin-Elmer) using ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction kits (Perkin-Elmer). Both strands from different clones were sequenced at least three times to ensure the sequencing quality. Finally, each of these sequences had their amino acid sequences predicted by an ORF search performed using the Geneious 4.8.5 program (Drummond et al. 2009). These sequences were also included in the dataset for forkhead phylogeny.

The full FOXL2 amino acid sequences present in the forkhead dataset used in the domain phylogeny (Supplementary Material 4), and their clustered full sequences, were used for the subsequent FOXL2 phylogeny reconstruction. The amino acid sequences were aligned using the Muscle algorithm (Edgar 2004), and obviously misplaced aligned regions were corrected manually. Poor-quality regions observed in the N-terminal portion were excluded from the analysis using the Seaview software (Supplementary Material 5). At the end, the reference dataset for FOXL2 phylogeny was composed of 62 amino acid sequences relevant to 50 metazoan species.

Forkhead and FOXL2 phylogenetic analysis

The choice of the best-fit model of evolution was performed with ProtTest3 (Darriba et al. 2011) for the forkhead and FOXL2 sequences using the Akaike Information Criterion (AIC) (Akaike 1974) for the best model selection.

The phylogenetic reconstruction was determined by maximum likelihood and Bayesian methods implemented in the Phyml v3.0.1 (Guindon and Gascuel 2003) and Beast v1.7.0 (Drummond and Rambaut 2007) software programs, respectively. For maximum likelihood analysis, the aLRT (SH-like) reliability test (Anisimova and Gascuel 2006) was adopted, and the values were supported by posterior probabilities obtained by Bayesian analysis. For Bayesian method generations, the burn-in was determined in Tracer (Rambaut and Drummond 2007) through log likelihood scores, and data were summarized in TreeAnnotator (Drummond and Rambaut 2007) after trees that were out of the convergence area had been discarded. The visualization and the final tree edition were performed using FigTree v1.3.1 (Drummond and Rambaut 2007) and the software package Mesquite (Maddison and Maddison 2001). In all phylogenetic analyses, the proportion of invariable sites and gamma-distributed rate variation across sites were estimated, and the substitution

rate categories were set in four categories. The random local clock model for Bayesian analysis was chosen because it allows for rate variation among lineages, and the model can run a series of local molecular clocks as described in Drummond and Suchard (2010) (see parameters in Table 1 and 2).

Evaluation of the *FOXL2* syntenic region

In attempt to provide an additional statement of the orthology for most of the newly described *FOXL2* sequences, especially the coelacanth *L. chalumnae*, the avian *T. guttata* and the marsupial *M. domestica* copies, and to understand the composition and organization of the genes in the proximity of *FOXL2*, the syntenic regions were analyzed in *D. rerio*, *G. aculeatus*, *G. morhua*, *O. latipes*, *O. niloticus*, *T. nigroviridis*, *T. rubripes*, *L. chalumnae*, *T. guttata* and *M. domestica*. Furthermore, for comparison, some ortholog sequences from representative species such as *Xenopus tropicalis* (frog) and *Homo sapiens* were included in the analysis of the syntenic regions.

Most of the synteny analysis was conducted using Ensembl with the genes identified using its genome browser. Because there were only partial transcript predictions for *O. niloticus* and *L. chalumnae*, without gene identification, each partial transcript sequence in the genomic region of *FOXL2A* and *FOXL2B* was retrieved and then queried by a blastp search against the NCBI database, with the identities and the E-values annotated from the highest similar hits from blast results (Supplementary Material 6). For each partial transcript, the retrieval was expanded to 50 kb of the flanking regions. The sequences acquired were inserted into the on-line program Softberry FGENESH (http://www.softberry.com/) using *T. rubripes* as the reference genome to recover the whole gene transcriptional region and confirm the identification of genes in these syntenic regions (Supplementary Material 6). However, for the gene *PIK3CD* (Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta

isoform), there was no transcript annotation in the *L. chalumnae FOXL2B* syntenic region in Ensembl. Because this gene appeared in our analysis as an important marker for *FOXL2B*, a tblastn search was conducted in Ensembl to identify the localization of *PIK3CD* in *L. chalumnae* using a sequence of the PIK3CD protein from *D. rerio* (accession number AAH54896) as the query. Because the resulting hit from the blast search was positioned to be syntenic with *FOXL2B*, approximately 50 kb of the flanking region was retrieved, and gene prediction was performed using Softberry FGENESH to exclude a possible pseudogene and to confirm the correct identification of PIK3CD in *L. chalumnae*. For the final predicted gene, its corresponding protein sequence (Supplementary Material 6) was used as the query for a blastp search in NCBI to check the protein prediction reliability and integrity.

Exhaustive search for the *FOXL2* duplicated copy

To assure that *FOXL2* duplicated copies were not missed during the blast search, we proceeded to perform an exhaustive search in the Ensembl genomes, isolating the flanking regions in which *FOXL2B* was located and conducting a gene name search of important *FOXL2B* vicinity markers identified in our work, including *PIK3CD*, *TMEM201* (Transmembrane protein 201), *CLSTN1* (Calsyntenin-1), *CTNNBIP1* (Beta-catenin-interacting protein 1) and *LZIC* (Leucine zipper and CTNNBIP1 domain-containing protein). We restricted our searches to annotated genomes that exhibited at least one of those markers. Once a syntenic region was identified, the nucleotide sequences of that region were retrieved. These sequences ranged from 15 to 1,000 kb, depending upon the size of the available region. A gene prediction was performed in Softberry FGENESH (http://www.softberry.com/) using *T. rubripes*, *X. tropicalis* and *H. sapiens* as reference genomes. For all predicted genes, their corresponding amino acid sequences were used as queries for a blastp search in NCBI to identify results with similarity to the forkhead protein family. When any similarity was

observed, the corresponding sequence was inserted in the domain phylogeny to verify the homology with FOXL2.

Nomenclature reference to systematic relationships

All taxonomic groups and systematic relationships mentioned in this work were based on the Interactive Tree of Life (iTOL) (Letunic and Bork 2011).

Results and Discussion

FOXL2 gene search and sequencing

During the genome searches of annotated sequences, blast and exhaustive searches for *FOXL2*, diverged duplicated copies were found in the teleosts *A. burtoni*, *G. morhua*, *M. zebra* and *O. niloticus*, in the coelacanth *L. chalumnae*, in the bird *T. guttata* and in the mammal *M. domestica* (one copy ortholog to *FOXL2A* and the other to *FOXL2B*, as will be discussed). In contrast, the *FOXL2* sequences obtained by genome walking exhibited just one copy of the gene in the chondrichthyans *R. lalandei* (accession number JX012129) and *C. callorynchus* (accession number JX012130), and in the teleost *C. monoculus* (accession number JX012131).

Interestingly, searches in the genome of another basal vertebrate, the agnathan *Petromyzon marinus* (sea lamprey), did not reveal a duplicate copy of *FOXL2*. Unfortunately, the retrieved sequence showed a partial protein (120 amino acids), and that sequence could only be utilized in the dataset for the domain phylogeny reconstruction.

All the sequences obtained for *FOXL2* from genome database searches and experimental procedures, which were posteriorly used for phylogenetic analysis, exhibited a single exon (as expected for the *FOXL2* gene) and a translated protein with variable sizes

(Supplementary Material 3). Furthermore, the present gene searches revealed the presence and absence of *FOXL2* copies.

Forkhead phylogeny

The domain phylogeny combined sequences from all forkhead family groups (A-R), and each one was represented as a monophyletic group. The FOXL2 branch was clearly identified from the rest of the forkhead proteins, and all the FOXL2 sequences obtained were placed in a monophyletic group, thus characterizing them as members of the FOXL2 subfamily (Figure 2A, Supplementary Material 4). The goal of this approach was to develop a more oriented FOXL2 phylogeny, supporting the use of sequences from the FOXL2 subfamily and also providing a precise way to establish the outgroup for the analysis. Thus, the domain phylogeny showed the following: 1- the sponge *Suberites domuncula* seems to be the proper outgroup for FOXL2 phylogeny and that 2- there is evidence of orthology among the extra FOXL2 copies. For instance, all the new duplicated copies obtained in this study (see the topic "*FOXL2* gene search and sequencing"), as well as the paralogs reported in other papers, were placed within the FOXL2 branch, including the diverged duplicated copies in the tetrapods *T. guttata* and *M. domestica* within a clade with the duplicated copies in *T. nigroviridis* and *T. rubripes*. Moreover, the sequence of the agnathan *P. marinus* that could not be inserted in the posterior FOXL2 specific phylogeny exhibited FOXL2A orthology in the domain approach (Supplementary Material 4).

A new view of *FOXL2* evolution

The FOXL2 phylogeny based only on the amino acid sequences of the forkhead domain did not exhibit a total resolute tree, probably due to the high conservation observed in this region, and was thus characterized by a low phylogenetic signal. In this case, based only

on an almost complete amino acid sequence alignment, a tree with well-supported branches was generated (Figure 2B).

The first thing that is evident in the phylogeny of FOXL2 is that it is possible to split the tree into three main clades (evidenced by a black bar in the right side of Figure 2B). Based on this result, together with other findings of the present work, a new nomenclature for the *FOXL2* gene in vertebrates is proposed. In summary, one clade includes the *FOXL2* gene of non-vertebrate taxa; another includes the two distinct copies of *FOXL2* from vertebrates, which we propose have evolved from the same duplication event. One copy is known in tetrapods as *FOXL2* and is considered the most thoroughly studied form of the gene (here named *FOXL2A*), and the other copy includes the duplicated diverged form (here named *FOXL2B*).

The representative species of non-vertebrates included in the FOXL2 evolution context were the following: Mollusca (*Crassostrea gigas*), Hemichordata (*Saccoglossus kowalevskii*), Urochordata (*Ciona intestinalis*) and Cephalochordata (*Branchiostoma floridae*). However, the phylogenetic placement of those groups could not be precisely established, possibly due to an accelerated substitution pattern represented by the long branches (Figure 2B).

The FOXL2A sequences obtained from genomes and experimental procedures, together with other sequences already known, were grouped within the FOXL2A clade of vertebrates, thus confirming their orthology. Moreover, the monophyly of major groups, such as chondrichthyans, teleosts and tetrapods, was clearly identified in the FOXL2A clade. However, the relationships among these groups were not accurately established, including the coelacanth sequence (Figure 2B). The unsolved placement of those taxa probably is due to the absence in the analysis of important intermediate vertebrate groups such as Dipnoi (lungfish), which would help to create a broader context to improve the evolutionary analysis. The same

observation (inaccurate topology) can be applied to some taxa within the teleost and tetrapod groups. Interestingly, FOXL2B copies were placed in a monophyletic group, including the newly described ones in this study (*A. burtoni, G. morhua, M. zebra, O. niloticus, L. chalumnae, T. guttata* and *M. domestica*) (Figure 2B). The FOXL2B sequences were represented as a separate clade from the FOXL2A sequences. The main point of this finding is the fact that the *L. chalumnae, T. guttata* and *M. domestica* duplicated copies were positioned within the FOXL2B clade, which refutes the hypothesis, proposed by Jiang et al. (2011), that the paralog FOXL2 copies are a product of the duplication event exclusive of teleosts. Thus, we hypothesized that the FOXL2B copies arose from an old duplication event; however, it was necessary to perform an analysis in the syntenic region of the predicted paralog copies and compare them to the teleost copies, to reinforce the orthology of the duplicated sequences.

In some cases, orthology among sequences can be inferred from synteny analysis because small syntenic blocks of the genome are frequently maintained and inherited throughout distant lineages, such as the *H. sapiens* and *D. rerio* genomes (Postlethwait et al. 2000). The present synteny analysis clearly showed that the flanking gene composition for *FOXL2* copies was similar in the species analyzed. For a clearer presentation, only the common syntenic genes among the analyzed species are shown in Figure 3.

In the *FOXL2A* synteny analysis (Figure 3A), a common general composition as well as gene orientation was observed in all species analyzed with emphasis on *MRPS22* (mitochondrial 28S ribosomal protein S22) and *PIK3CB* (beta isoform of the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit) that were observed near the *FOXL2A* copies (except *MRPS22* for *O. niloticus* and *T. nigroviridis*; however, this result can be related to the limited scaffold size in the case of *O. niloticus*). The *FOXL2A* syntenic

region for the fish O*ryzias latipes* was not analyzed here because syntenic information for this region is currently unavailable in Ensembl.

A conservation pattern was also displayed in the syntenic region of the *FOXL2B* copy (Figure 3B). In this case, *PIK3CD* and *TMEM201* were both close to *FOXL2B* in all analyzed species, except for *G. morhua*, which does not show the copy of *TMEM201*; however, this result can also be related to the limited scaffold size. It was possible to observe some rearrangements in the orientation and repositioning of some genes. For instance, the *TMEM201* gene was located upstream to *FOXL2B* in *D. rerio, L. chalumnae, T. guttata* and *M. domestica*, while it was downstream in all other analyzed species, suggesting the occurrence of rearrangements in this genomic region. Another feature observed in the *FOXL2B* cluster of *L. chalumnae, T. guttata* and *M. domestica* was the inversion rearrangement involving the *CLSTN1* and *CTNNBIP1* genes.

The syntenic regions of all described *FOXL2B* copies, including the new ones in this study, exhibited high composition conservation, which supports the hypothesis of the orthology of these sequences, also shown by the phylogeny of FOXL2. Based on this scenario of orthology among all *FOXL2B* genes, the hypotheses of the duplicated copies being generated in an exclusive duplication event of teleosts (Jiang et al. (2011) can be discarded. The duplication event must have occurred in an ancestor of bony vertebrates, for the *FOXL2B* copy to appear in teleosts, coelacanth and tetrapods (Figure 4). Furthermore, the *FOXL2A* and *FOXL2B* copies would have suffered, posteriorly, an additional duplication event exclusive of teleosts. In this context, the generated copies were possibly lost, as the presence of new duplicated copies, besides *FOXL2A* and *FOXL2B,* was not reported in teleosts. Indeed the presence of a possible *FOXL2A* pseudogene described in *S. salar*, with short identity with the forkhead domain, and exhibiting a divergent N and C terminal regions when compared to

FOXL2A and *FOXL2B* (von Schalburg et al. 2011), may represents a remnant of the *FOXL2* copies generated in the exclusive whole genome duplication occurred in teleosts.

More information involving genome sequencing and/or better assembly for agnathan and chondrichthyan species (which exhibit just one copy of *FOXL2*) is required for the exact placement of the *FOXL2* duplication event. For one of the chondrichthyan species analyzed, *Scyliorhinus canicula*, the sequence was obtained from an expression study involving FOXL2 (Wotton, French, Shimeld 2007), and only a copy similar to FOXL2A was observed. Furthermore, the nucleotide sequencing analysis reported here for *R. lalandei* and *C. callorynchus* also revealed only *FOXL2A* copies. However, in this case, there is a possibility that the degenerate primers used could only anneal to *FOXL2A* sequences, precluding the detection of paralog copies in chondrichthyans, which are the same as those observed in *C. monoculus* and possibly in the other teleost species in which the *FOXL2* sequences were obtained by PCR procedures in previous studies.

Moreover, a significant observation was that the *PIK3CB* and *PIK3CD* genes in the *FOXL2A* and *FOXL2B* syntenic regions, respectively, have been reported to have evolved from a duplication that gave rise to the two forms (Brown and Auger 2011). Futhermore, the same study verified that both *PI3K* copies remained in teleosts and tetrapods; therefore, the duplication event also occurred at least in an ancestor of bony vertebrates. Taken together with our present analysis, this scenario indicates that *FOXL2A* and *FOXL2B* were not products of a single gene duplication. However, searches of other genes observed in the syntenic region of *FOXL2*, such as *CLSTN1* and *CLSTN2* (Calsyntenin-2); *RBP2* (Retinol-binding protein 2) and *RBP7* (Retinol-binding protein 7); and *NMNAT1* (Nicotinamide nucleotide adenylyltransferase 1) and *NMNAT3* (Nicotinamide nucleotide adenylyltransferase 3), did not reveal any related evidence associating them with a duplication event. In this context, it would be very informative to perform a phylogenetic analysis of those genes to

understand their evolutionary history and their relationship to a *FOXL2* and *PI3K* duplication event.

An additional important observation was that no duplicated copy of *FOXL2* was detected in any of the 23 eutherian genomes searched, even in the exhaustive search in 18 of those species, thus suggesting that the *FOXL2B* copy was lost in this group (Figure 4, Supplementary Material 7). *FOXL2B* was also not detected in the amphibian species *X. tropicalis*, in the sauropsidian *Anolis carolinensis* (anole lizard), *Gallus gallus* (chicken) and *Meleagris gallopavo* (turkey), in the monotreme *Ornithorhynchus anatinus* (platypus) or in the methaterian species *Sarcophilus harrisii*. These results indicate an evolutionary process characterized by events of independent losses of *FOXL2* copies during the evolutionary history of bony vertebrates. However, except for eutherians, few species were analyzed, and because most of these genomes are still in an assembly process, this scenario of independent losses will only be clearly understood and confirmed with future studies.

FOXL2B activity

The scenario described in the current work for paralog copies also raises an important issue about the role of FOXL2B, especially with regard to sexual determination and differentiation. In an expression study involving the fish *O. mykiss* (Baron et al. 2004), *FOXL2B* was shown to be expressed later than *FOXL2A* in ovaries. The authors suggested that this difference indicated a neofunctionalization process for FOXL2B. Furthermore, in another study with the fish, *S. salar* (von Schalburg et al. 2011), *FOXL2B* exhibited a similar but also very distinctive pattern of expression compared to *FOXL2A*. This difference was observed especially in males, in which *FOXL2B* was expressed at higher levels in testes, while in females, it was restricted to the gills, skin and ovary. Although the authors did not

draw any conclusions about the roles of FOXL2B, the different patterns of expression in different types of tissues could indicate a neofunctionalization process for *FOXL2B*.

Another interesting observation was based on the analysis of the polyalanine tract that is exclusive to mammals. In *M. domestica*, the FOXL2A protein showed 16 alanines (Supplementary Material 5), unlike the value of 14 alanines suggested by Cocquet et al. (2003). In addition, the FOXL2B protein in *M. domestica* did not exhibit the polyalanine tract, which may indicate a change in the protein function compared to the FOXL2A protein. However, how FOXL2B adapted its function during the evolution of teleosts and tetrapods remains unknown pending additional functional studies.

Conclusion

The broad study with the domain approach and the incorporation of new FOXL2 duplicated sequences in the inferred phylogeny provided the basis for a novel discussion of evolutionary studies of FOXL2. The current study identified duplicated forms of *FOXL2* in different vertebrate groups, which suggests the need for a revision of the nomenclature for *FOXL2A* and *FOXL2B*. Notably, the established phylogeny, combined with synteny analysis, strongly supported the conclusion that the predicted FOXL2B copies in coelacanth and tetrapods are indeed orthologous to all FOXL2B copies of teleosts. Furthermore, the duplication event that gave rise to the paralog copies was not unique to teleosts and was a common event that occurred in an ancestor of bony vertebrates. Our results also suggest that this duplication was not exclusive to *FOXL2*. From the description of new *FOXL2B* copies in more fish species, obtained by genome searches, it is reasonable to believe that there are more teleost species, not yet reported, that carry a *FOXL2B* copy. In this context, it remains unknown whether most species in this group carry this additional copy, while others have lost it, or if all teleost species have the duplicated version of *FOXL2*. The distribution of *FOXL2B*

was shown to be diversified in tetrapods, with a strong indication of loss of this copy in eutherians. In addition to this newly reported scenario, more information is required for understanding the role of the *FOXL2B* copy in vertebrates, especially in relation to sexual determination and differentiation.

Acknowledgements

This study was supported by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) from Brazil. The authors are grateful to Dr. OBF Gadig for providing tissue samples of *C. callorynchus*.

Literature cited

As referências deste artigo encontram-se reunidas ao final desta dissertação no item 5, *Referências Bibliográficas*, página 51.

Tables

Table 1. Maximum likelihood reconstruction tree parameters.

Data	Substitution model	Tree searching operations	Starting tree	Branch support
Forkhead	LG	Best of NNI	Neighbor- joining	aLRT (SH-like)
FOXL2	Dayhoff	Best of SPR	Neighbor- joining	aLRT (SH-like)

NNI: nearest neighbor interchange, SPR: subtree pruning and regrafting, aLRT: approximate likelihood ratio test

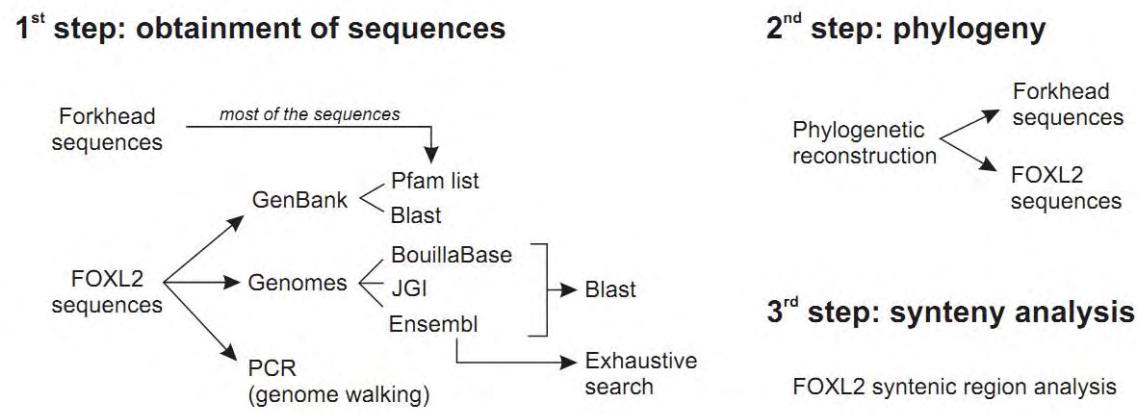
Table 2. Bayesian method reconstruction tree parameters

Data	Substitution model	Molecular clock model	Base frequencies	Starting tree	Generations/ Burn-in	Sample frequency	Branch support
Forkhead	WAG	Random local clock	Estimated	Randomly generated	50,000,000/ 17,000	1,000	Posterior probability
FOXL2	Dayhoff	Random local clock	Estimated	Randomly generated	30,000,000/ 5,000	1,000	Posterior probability

Figures

Figure 1. Fluxogram of the methodology adopted in the present work. In (A), general steps performed for the evolutionary analysis. In (B), details from the 2nd step of (A): procedures for forkhead domain and FOXL2 phylogenetic reconstruction.

(A)


1st step: obtainment of sequences
Forkhead sequences
most of the sequences
GenBank
Pfam list
Blast
FOXL2 sequences
Genomes
BouillaBase
JGI
Ensembl
Blast
PCR (genome walking)
Exhaustive search
2nd step: phylogeny
Phylogenetic reconstruction
Forkhead sequences
FOXL2 sequences
3rd step: synteny analysis
FOXL2 syntenic region analysis

(B)

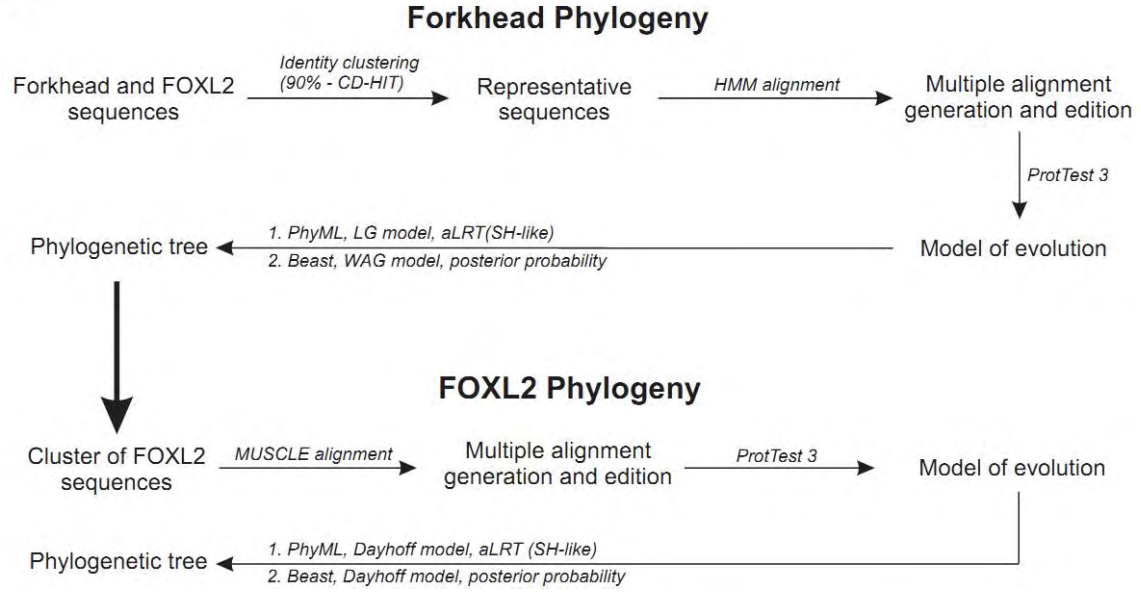

Forkhead Phylogeny
Forkhead and FOXL2 sequences
Identity clustering (90% - CD-HIT)
Representative sequences
HMM alignment
Multiple alignment generation and edition
ProtTest 3
Phylogenetic tree
1. PhyML, LG model, aLRT(SH-like)
2. Beast, WAG model, posterior probability
Model of evolution
FOXL2 Phylogeny
Cluster of FOXL2 sequences
MUSCLE alignment
Multiple alignment generation and edition
ProtTest 3
Model of evolution
Phylogenetic tree
1. PhyML, Dayhoff model, aLRT (SH-like)
2. Beast, Dayhoff model, posterior probability

Figure 2. Phylogenetic relationship of the forkhead domain (A) and FOXL2 (B) sequences. The letters in (A) represent each different member of the forkhead family (FOXA to FOXR). The branch colors indicate the consensus statistical support from PhyML and BEAST analysis: blue (<70%) and red (≥70%). The scale bars indicate the average number of amino acid substitutions per site. In (B), the first and second branch values represent the aLRT (SH-like) and the posterior probability from the PhyML and BEAST programs, respectively. The clades of each FOXL2 form are indicated by right side bars. The highlighted clades represent chondrichthyans (blue), teleosts (green), coelacanth (yellow) and tetrapods (red). The asterisks represent the procedure used to obtain each sequence in the phylogeny: accession number from GenBank (*), genome search (**), and genome walking (***). The scale bars indicate the average number of amino acid substitutions per site.

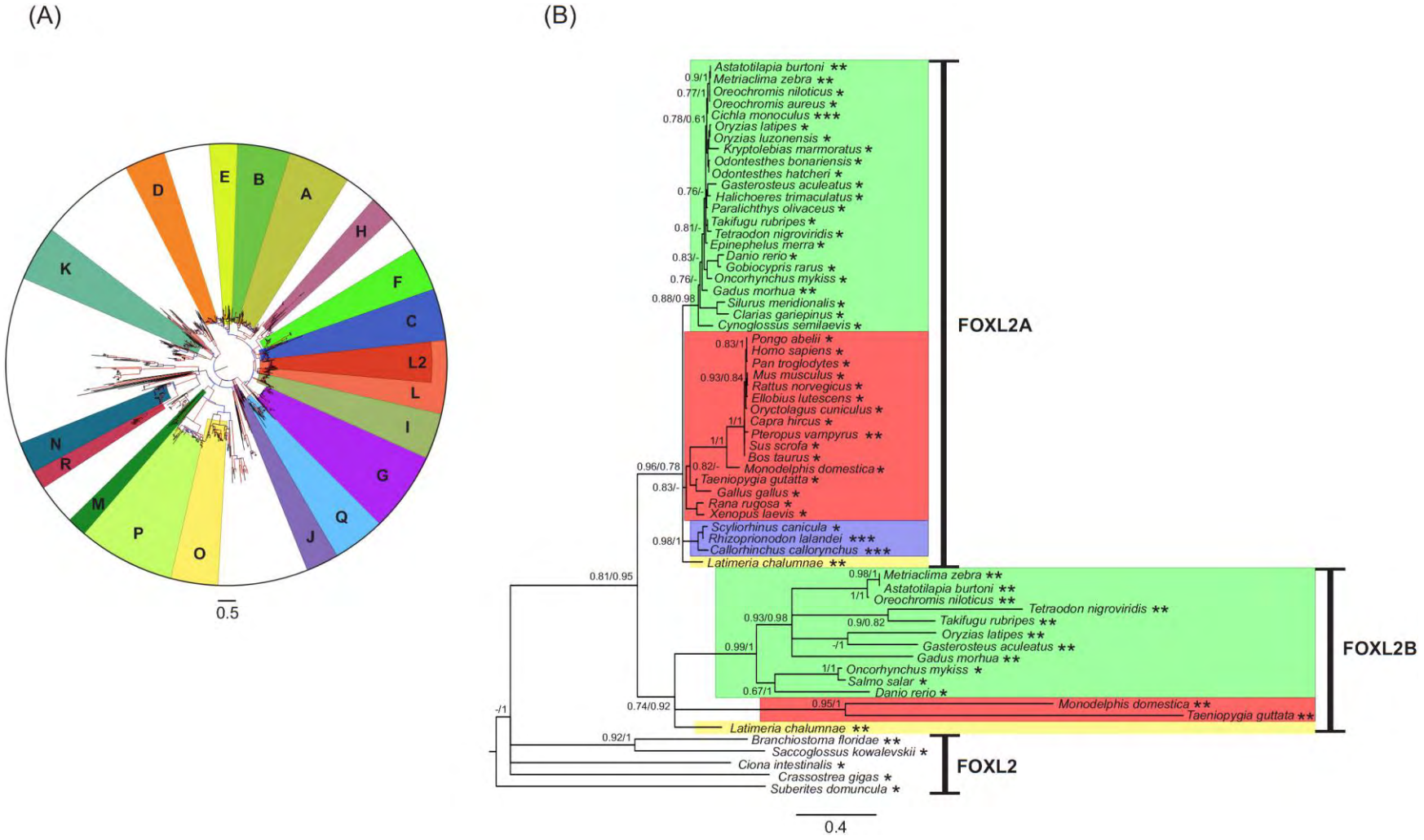

(A)
D
E
B
A
H
K
F
C
L2
L
I
N
R
G
M
Q
P
O
J
0.5
(B)
Astatotilapia burtoni **
Metriaclima zebra **
Oreochromis niloticus *
Oreochromis aureus *
Cichla monoculus ***
Oryzias latipes *
Oryzias luzonensis *
Kryptolebias marmoratus *
Odontesthes bonariensis *
Odontesthes hatcheri *
Gasterosteus aculeatus *
Halichoeres trimaculatus *
Paralichthys olivaceus *
Takifugu rubripes *
Tetraodon nigroviridis *
Epinephelus merra *
Danio rerio *
Gobiocypris rarus *
Oncorhynchus mykiss *
Gadus morhua **
Silurus meridionalis *
Clarias gariepinus *
Cynoglossus semilaevis *
Pongo abelii *
Homo sapiens *
Pan troglodytes *
Mus musculus *
Rattus norvegicus *
Ellobius lutescens *
Oryctolagus cuniculus *
Capra hircus *
Pteropus vampyrus **
Sus scrofa *
Bos taurus *
Monodelphis domestica *
Taeniopygia gutatta *
Gallus gallus *
Rana rugosa *
Xenopus laevis *
Scyliorhinus canicula *
Rhizoprionodon lalandei ***
Callorhinchus callorynchus ***
Latimeria chalumnae **
FOXL2A
Metriaclima zebra **
Astatotilapia burtoni **
Oreochromis niloticus **
Tetraodon nigroviridis **
Takifugu rubripes **
Oryzias latipes **
Gasterosteus aculeatus **
Gadus morhua **
Oncorhynchus mykiss *
Salmo salar *
Danio rerio *
Monodelphis domestica **
Taeniopygia guttata **
Latimeria chalumnae **
FOXL2B
Branchiostoma floridae **
Saccoglossus kowalevskii *
Ciona intestinalis *
Crassostrea gigas *
Suberites domuncula *
FOXL2
0.4

Figure 3. Schematic representation of the syntenic regions using *FOXL2A* (A) and *FOXL2B* (B) as reference genes. The transcriptional orientation of the genes is depicted by the direction of the arrows. The observed genes represent only the common markers for the analyzed species and not the complete syntenic block composition.

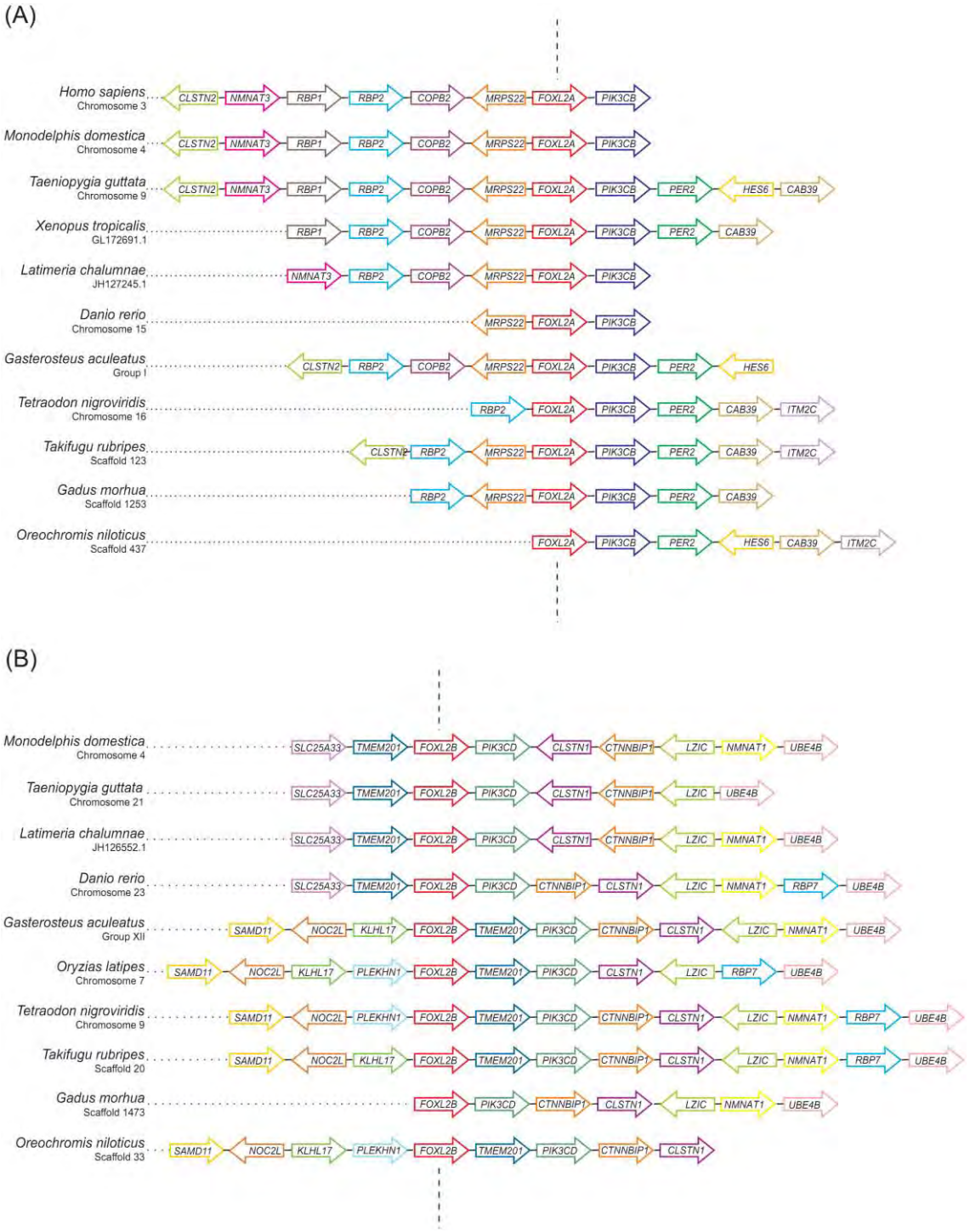
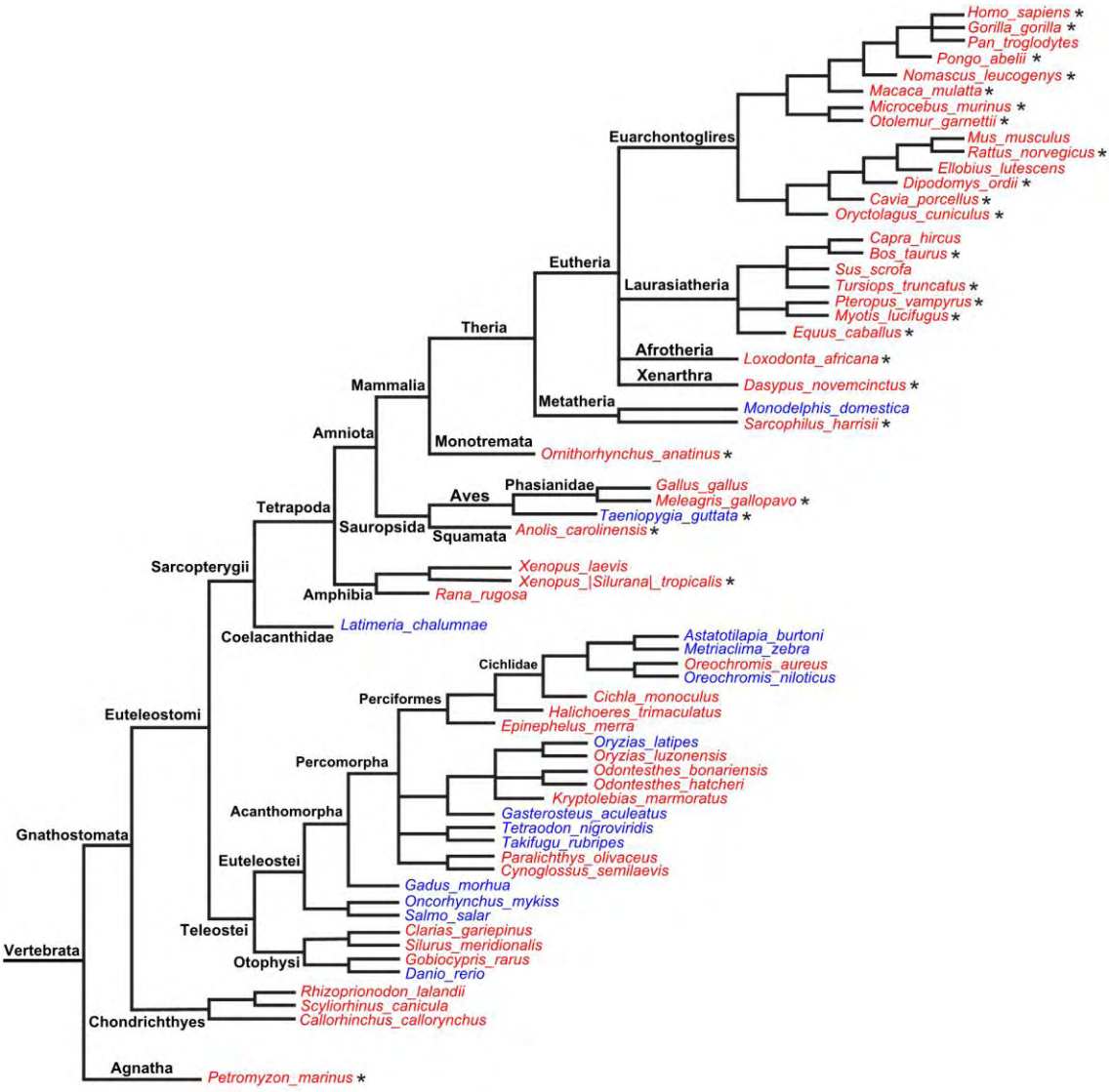

(A)
Homo sapiens
Chromosome 3
CLSTN2 NMNAT3 RBP1 RBP2 COPB2 MRPS22 FOXL2A PIK3CB
Monodelphis domestica
Chromosome 4
CLSTN2 NMNAT3 RBP1 RBP2 COPB2 MRPS22 FOXL2A PIK3CB
Taeniopygia guttata
Chromosome 9
CLSTN2 NMNAT3 RBP1 RBP2 COPB2 MRPS22 FOXL2A PIK3CB PER2 HES6 CAB39
Xenopus tropicalis
GL172691.1
RBP1 RBP2 COPB2 MRPS22 FOXL2A PIK3CB PER2 CAB39
Latimeria chalumnae
JH127245.1
NMNAT3 RBP2 COPB2 MRPS22 FOXL2A PIK3CB
Danio rerio
Chromosome 15
MRPS22 FOXL2A PIK3CB
Gasterosteus aculeatus
Group I
CLSTN2 RBP2 COPB2 MRPS22 FOXL2A PIK3CB PER2 HES6
Tetraodon nigroviridis
Chromosome 16
RBP2 FOXL2A PIK3CB PER2 CAB39 ITM2C
Takifugu rubripes
Scaffold 123
CLSTN2 RBP2 MRPS22 FOXL2A PIK3CB PER2 CAB39 ITM2C
Gadus morhua
Scaffold 1253
RBP2 MRPS22 FOXL2A PIK3CB PER2 CAB39
Oreochromis niloticus
Scaffold 437
FOXL2A PIK3CB PER2 HES6 CAB39 ITM2C
(B)
Monodelphis domestica
Chromosome 4
SLC25A33 TMEM201 FOXL2B PIK3CD CLSTN1 CTNNBIP1 LZIC NMNAT1 UBE4B
Taeniopygia guttata
Chromosome 21
SLC25A33 TMEM201 FOXL2B PIK3CD CLSTN1 CTNNBIP1 LZIC UBE4B
Latimeria chalumnae
JH126552.1
SLC25A33 TMEM201 FOXL2B PIK3CD CLSTN1 CTNNBIP1 LZIC NMNAT1 UBE4B
Danio rerio
Chromosome 23
SLC25A33 TMEM201 FOXL2B PIK3CD CTNNBIP1 CLSTN1 LZIC NMNAT1 RBP7 UBE4B
Gasterosteus aculeatus
Group XII
SAMD11 NOC2L KLHL17 FOXL2B TMEM201 PIK3CD CTNNBIP1 CLSTN1 LZIC NMNAT1 UBE4B
Oryzias latipes
Chromosome 7
SAMD11 NOC2L KLHL17 PLEKHN1 FOXL2B TMEM201 PIK3CD CLSTN1 LZIC RBP7 UBE4B
Tetraodon nigroviridis
Chromosome 9
SAMD11 NOC2L PLEKHN1 FOXL2B TMEM201 PIK3CD CTNNBIP1 CLSTN1 LZIC NMNAT1 RBP7 UBE4B
Takifugu rubripes
Scaffold 20
SAMD11 NOC2L KLHL17 FOXL2B TMEM201 PIK3CD CTNNBIP1 CLSTN1 LZIC NMNAT1 RBP7 UBE4B
Gadus morhua
Scaffold 1473
FOXL2B PIK3CD CTNNBIP1 CLSTN1 LZIC NMNAT1 UBE4B
Oreochromis niloticus
Scaffold 33
SAMD11 NOC2L KLHL17 PLEKHN1 FOXL2B TMEM201 PIK3CD CTNNBIP1 CLSTN1

Figure 4. Summary of the newest evolutionary scenario for *FOXL2* copies in vertebrates. The schematic representation of the phylogenetic relationships among the analyzed groups is exhibited, according to the Interactive Tree of Life (iTOL) (Letunic and Bork 2011). Taxa are labeled as carrying (blue) or missing (red) the *FOXL2B* copy. The asterisks (*) represent the species in which an exhaustive search for the *FOXL2B* copy was performed. The length of each branch is not scaled to any evolutionary distance measurement.


Homo_sapiens *
Gorilla_gorilla *
Pan_troglodytes
Pongo_abelii *
Nomascus_leucogenys *
Macaca_mulatta *
Microcebus_murinus *
Otolemur_garnettii *
Euarchontoglires
Mus_musculus
Rattus_norvegicus *
Ellobius_lutescens
Dipodomys_ordii *
Cavia_porcellus *
Oryctolagus_cuniculus *
Capra_hircus
Bos_taurus *
Sus_scrofa
Tursiops_truncatus *
Pteropus_vampyrus *
Myotis_lucifugus *
Equus_caballus *
Eutheria
Laurasiatheria
Afrotheria
Loxodonta_africana *
Xenarthra
Dasypus_novemcinctus *
Theria
Metatheria
Monodelphis_domestica
Sarcophilus_harrisii *
Mammalia
Amniota
Monotremata
Ornithorhynchus_anatinus *
Phasianidae
Gallus_gallus
Meleagris_gallopavo *
Aves
Taeniopygia_guttata *
Tetrapoda
Sauropsida
Squamata
Anolis_carolinensis *
Sarcopterygii
Xenopus_laevis
Xenopus_|Silurana|_tropicalis *
Amphibia
Rana_rugosa
Latimeria_chalumnae
Coelacanthidae
Astatotilapia_burtoni
Metriaclima_zebra
Oreochromis_aureus
Oreochromis_niloticus
Cichlidae
Perciformes
Cichla_monoculus
Halichoeres_trimaculatus
Epinephelus_merra
Euteleostomi
Oryzias_latipes
Oryzias_luzonensis
Odontesthes_bonariensis
Odontesthes_hatcheri
Percomorpha
Kryptolebias_marmoratus
Acanthomorpha
Gasterosteus_aculeatus
Tetraodon_nigroviridis
Takifugu_rubripes
Gnathostomata
Euteleostei
Paralichthys_olivaceus
Cynoglossus_semilaevis
Gadus_morhua
Oncorhynchus_mykiss
Salmo_salar
Teleostei
Clarias_gariepinus
Silurus_meridionalis
Vertebrata
Otophysi
Gobiocypris_rarus
Danio_rerio
Rhizoprionodon_lalandii
Scyliorhinus_canicula
Chondrichthyes
Callorhinchus_callorynchus
Agnatha
Petromyzon_marinus *

Supplementary Material Captions

Os materiais suplementares se encontram reunidos no CD que acompanha esta dissertação.

Supplementary Material 1. Information about each retrieved protein sequence from Genbank using the Pfam list of accession numbers for the forkhead domain.

Supplementary Material 2. Forkhead alignment of the sequences used in the domain phylogeny.

Supplementary Material 3. (A) Information about the acquisition of each FOXL2 sequence. The nomenclature suggested by this work was used in this table to refer to each gene copy. (B) The FOXL2 coding region and protein sequences obtained from genome searches. The forkhead domain, the size and location of which was determined from the Pfam database, is underlined.

Supplementary Material 4. The region comprehended by the FOXL2 branch in the domain phylogeny. Thicker lines: statistical support > 70% in maximum likelihood and Bayesian analysis. Blue color, the diverged duplicated copies (FOXL2B); red color, the sequence of *Petromyzon marinus*.

Supplementary Material 5. FOXL2 multiple alignment: interleaved (A) and sequential (B) formats.

Supplementary Material 6. Predicted genes from the FOXL2 syntenic region in *L. chalumnae* and *O. niloticus*. (A) Table of the identified genes with their corresponding ID from Ensembl and identities or E-values from blast searches. (B) The sequences of each predicted gene.

Supplementary Material 7. Figures of the syntenic regions obtained from the Ensembl browser (http://www.ensembl.org/) during the exhaustive search for *FOXL2B* copies.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans. Automatic Control* 19:716–723.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

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

Baron D, Cocquet J, Xia X, Fellous M, Guiguen Y, Veitia RA. 2004. An evolutionary and functional analysis of FoxL2 in rainbow trout gonad differentiation. *J. Mol. Endocrinol.* 33:705-715.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433-438.

Bravieri R, Shiyanova T, Chen TH, Liao XB. 1997. Different DNA contact schemes are used by two winged helix proteins to recognize a DNA binding sequence. *Nucleic Acids Res.* 25:2888-2896.

Bridges CB. 1935. Salivary chromosome maps with a key to the banding of the chromosomes of Drosophila melanogaster. *J. Hered.* 26:60-64.

Brissette JL, Li J, Kamimura J, Lee D, Dotto GP. 1996. The product of the mouse nude locus Whn, regulates the balance between epithelial cell growth and differentiation. *Genes Dev.* 10:2212-2221.

Brown JR, Auger KR. 2011. Phylogenomics of phosphoinositide lipid kinases: perspectives on the evolution of second messenger signaling and drug discovery. *BMC Evol Biol* 11:4.

Carlsson P, Mahlapuu M. 2002. Forkhead transcription factors: key players in development and metabolism. *Dev. Biol.* 250:1-23.

Clark KL, Halay ED, Lai ES, Burley SK. 1993. Co-Crystal Structure of the Hnf-3/Fork Head DNA-Recognition Motif Resembles Histone-H5. *Nature* 364:412-420.

Cocquet J, De Baere E, Gareil M, Pannetier M, Xia X, Fellous M, Veitia RA. 2003. Structure, evolution and expression of the FOXL2 transcription unit. *Cytogenet Genome Res* 101:206-211.

Cocquet J, Pailhoux E, Jaubert F, et al. 2002. Evolution and expression of FOXL2. *J. Med. Genet.* 39:916-921.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics* 9:938-950.

Crisponi L, Deiana M, Loi A, et al. 2001. The putative forkhead transcription factor FOXL2 is mutated in blepharophimosis/ptosis/epicanthus inversus syndrome. *Nat. Genet.* 27:159-166.

Crow KD, Stadler PF, Lynch VJ, Amemiya C, Wagner GP. 2006. The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Mol. Biol. Evol.* 23:121-136.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164-1165.

De Baere E, Dixon MJ, Small KW, et al. 2001. Spectrum of FOXL2 gene mutations in blepharophimosis-ptosis-epicanthus inversus (BPES) families demonstrates a genotype-phenotype correlation. *Hum. Mol. Genet.* 10:1591-1600.

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

Deng C, Cheng CH, Ye H, He X, Chen L. 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc. Natl. Acad. Sci. U. S. A.* 107:21593-21598.

Dottori M, Gross MK, Labosky P, Goulding M. 2001. The winged-helix transcription factor Foxd3 suppresses interneuron differentiation and promotes neural crest cell fate. *Development* 128:4127-4138.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A. 2009. Geneious v4.8.5, Available from http://www.geneious.com.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol* 8:114.

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

Felsenstein J. 2004. Inferring Phylogenies: Sinauer Associates.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39:W29-W37.

Finn RD, Mistry J, Tate J, et al. 2010. The Pfam protein families database. *Nucleic Acids Res* 38:D211-222.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99-113.

Flicek P, Amode MR, Barrell D, et al. 2011. Ensembl 2011. *Nucleic Acids Res.* 39:D800-D806.

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

Freeman S, Herron JC. 2009. Análise evolutiva. Porto Alegre: Artmed.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27:221-224.

Gray GS, Fitch WM. 1983. Evolution of Antibiotic-Resistance Genes - the DNA-Sequence of a Kanamycin Resistance Gene from Staphylococcus-Aureus. *Mol. Biol. Evol.* 1:57-66.

Gross DN, van den Heuvel APJ, Birnbaum MJ. 2008. The role of FoxO in the regulation of metabolism. *Oncogene* 27:2320-2336.

Guerrero-Estévez S, Moreno-Mendoza N. 2009. Sexual determination and differentiation in teleost fish. *Rev. Fish Biol. Fish*. 20(1):101-121.

Guindon S, Gascuel O. 2003. PhyML: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696-704.

Harrison CJ, Langdale JA. 2006. A step by step guide to phylogeny reconstruction. *Plant J.* 45:561-572.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157-1164.

Hennig W. 1966. Phylogenetic Systematics: University of Illinois Press.

Higgis PG, Attwood TK. 2005. Bioinformatics and molecular evolution. United Kingdom: Blackwell Science Ltd.

Hillis DM. 1996. Inferring complex phylogenies. *Nature* 383:130-131.

Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* 59:190-203.

Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Dev. Suppl.*:125-133.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* 11:97-108.

Jaillon O, Aury JM, Wincker P. 2009. "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. *C R Biol* 332:241-253.

Jaubert F, Galmiche L, Lortat-Jacob S, Fournet JC, Fellous M. 2011. Foxl-2 in gonad development and pathology. *Arkh. Patol.* 73:10-13.

Jiang W, Yang Y, Zhao D, Liu X, Duan J, Xie S, Zhao H. 2011. Effects of sexual steroids on the expression of foxl2 in Gobiocypris rarus. *Comp. Biochem. Physiol. B. Biochem. Mol. Biol.* 160:187-193.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428:617-624.

Lehmann OJ, Sowden JC, Carlsson P, Jordan T, Bhattacharya SS. 2003. Fox's in development and disease. *Trends Genet.* 19:339-344.

Lemey P, Salemi M, Vandamme AM. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. New York, US: Cambridge University Press.

Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39:W475-478.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.

Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet.* 21:602-607.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics* 4:865-875.

Maddison WP, Maddison DR. 2001. Mesquite: a modular system for evolutionary analysis. Version 2.75. Available from: http://mesquiteproject.org.

Mahlapuu M, Ormestad M, Enerback S, Carlsson P. 2001. The forkhead transcription factor Foxf1 is required for differentiation of extra-embryonic and lateral plate mesoderm. *Development* 128:155-166.

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

Moumne L, Dipietromaria A, Batista F, Kocer A, Fellous M, Pailhoux E, Veitia RA. 2008. Differential aggregation and functional impairment induced by polyalanine expansions in FOXL2, a transcription factor involved in cranio-facial and ovarian development. *Hum. Mol. Genet.* 17:1010-1019.

Nakae J, Kitamura T, Kitamura Y, Biggs WH, 3rd, Arden KC, Accili D. 2003. The forkhead transcription factor Foxo1 regulates adipocyte differentiation. *Dev Cell* 4:119-129.

Nehls M, Pfeifer D, Schorpp M, Hedrich H, Boehm T. 1994. New Member of the Winged-Helix Protein Family Disrupted in Mouse and Rat Nude Mutations. *Nature* 372:103-107.

Nicholas KB, Nicholas Jr HB, Deerfield DW. 1997. GeneDoc: Analysis and visualization of genetic variation. *EMBERNEW.NEWS* 4.

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

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* 10:1890-1902.

Rambaut A, Drummond AJ. 2007. Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer.

Sambrook J, Russel DW. 2001. Molecular Cloning: A Laboratory Manual. New York: Cold Spring Harbor Laboratory Press.

Semon M, Wolfe KH. 2007. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17:505-512.

Serebrowsky AS. 1938. Genes scute and achaete in Drosophila melanogaster, and a hypothesis of gene divergency. *Cr Acad Sci Urss* 19:77-81.

Simmons MP, Ochoterena H, Freudenstein JV. 2002. Amino acid vs. nucleotide characters: challenging preconceived notions. *Mol. Phylogenet. Evol.* 24:78-90.

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

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome biology* 8:R50.

Treier M, Gleiberman AS, O'Connell SM, Szeto DP, McMahon JA, McMahon AP, Rosenfeld MG. 1998. Multistep signaling requirements for pituitary organogenesis in vivo. *Genes Dev.* 12:1691-1704.

Uhlenhaut NH, Jakob S, Anlag K, et al. 2009. Somatic sex reprogramming of adult ovaries to testes by FOXL2 ablation. *Cell* 139:1130-1142.

Uhlenhaut NH, Treier M. 2006. Foxl2 function in ovarian development. *Mol. Genet. Metab.* 88:225-234.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* 35:W71-W74.

Veitia RA. 2010. FOXL2 versus SOX9: a lifelong "battle of the sexes". *Bioessays* 32:375-380.

von Schalburg KR, Yasuike M, Yazawa R, et al. 2011. Regulation and expression of sexual differentiation factors in embryonic and extragonadal tissues of Atlantic salmon. *BMC Genomics* 12:31.

Weigel D, Jurgens G, Kuttner F, Seifert E, Jackle H. 1989. The Homeotic Gene Fork Head Encodes a Nuclear-Protein and Is Expressed in the Terminal Regions of the Drosophila Embryo. *Cell* 57:645-658.

Wolfe K. 2000. Robustness - it's not where you think it is. *Nat. Genet.* 25:3-4.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708-713.

Wotton KR, French KE, Shimeld SM. 2007. The developmental expression of foxl2 in the dogfish Scyliorhinus canicula. *Gene expression patterns : GEP* 7:793-797.

Yang Z. 2006. Computational Molecular Evolution. Great Britain: Oxford University Press.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18:292–298.