



UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Biociências, Campus de Botucatu



ORGANIZAÇÃO GENÔMICA E ANÁLISE DA EXPRESSÃO DO
GENE HNRNP Q-LIKE (HETEROGENEOUS NUCLEAR
RIBONUCLEOPROTEIN Q-LIKE) RETROINSERIDO NO
CROMOSSOMO B DE *Astatotilapia latifasciata*

BIANCA DE OLIVEIRA CARMELLO

BOTUCATU – SP

2015

UNIVERSIDADE ESTADUAL PAULISTA

“Julio de Mesquita Filho”

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

ORGANIZAÇÃO GENÔMICA E ANÁLISE DA EXPRESSÃO DO
GENE HNRNP Q-LIKE (HETEROGENEOUS NUCLEAR
RIBONUCLEOPROTEIN Q-LIKE) RETROINSERIDO NO
CROMOSSOMO B DE *Astatotilapia latifasciata*

BIANCA DE OLIVEIRA CARMELLO

ORIENTADOR: CESAR MARTINS

COORIENTADOR: GUILHERME TARGINO VALENTE

Dissertação apresentada ao Instituto de
Biotecnologia, Câmpus de Botucatu, UNESP, para
obtenção do título de Mestre no Programa de
Pós- Graduação em Ciências Biológicas
(Genética).

**BOTUCATU – SP
2015**

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Carmello, Bianca de Oliveira.

Organização genômica e análise da expressão do gene hnRNP Q-like (Heterogeneous nuclear ribonucleoprotein Q-like) retroinserido no cromossomo B de *Astatotilapia latifasciata* / Bianca de Oliveira Carmello. - Botucatu, 2015

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: Cesar Martins

Coorientador: Guilherme Targino Valente

Capes: 20204000

1. Cromossomos. 2. Código genético. 3. Genética Expressão. 4. Genomas. 5. Reação em cadeia de polimerase.

Palavras-chave: Duplicação; Evolução; Genoma; Retrogene; Sequenciamento massivo.

AGRADECIMENTOS

Agradeço primeiramente à Deus por mais essa conquista e por me proteger sempre.

Aos meus pais por todo amor, apoio, incentivo, ajuda e compreensão em todas as horas, sem os quais não teria chegado até aqui.

À minha irmã Victória e ao meu namorado Raul pelo amor, amizade, companheirismo, alegria e palavras de conforto.

À minha família por todos os conselhos, por serem preocupados com minha vida pessoal e profissional e por orgulharem-se de todas as minhas conquistas, em especial às minhas avós Ercília e Herminia que me acompanharam a vida toda, me aconselhando e sendo prestativas o tempo todo.

Ao meu orientador Cesar Martins pela confiança, por me receber como aluna e todos os conhecimentos e conselhos transmitidos que contribuíram com a minha formação profissional.

Ao meu coorientador Guilherme Valente por toda ajuda e conhecimentos que foram essenciais para o meu trabalho.

Aos meus colegas de laboratório Adauto, Bruno, Diego, Diogo, Erica, Fábio, Farhan, Maryam, Natália, Rafael por toda amizade e aprendizado, em especial ao Adauto por toda ajuda e paciência.

Aos meus amigos de Botucatu e São Carlos pelas palavras amigas, conselhos, paciência e apoio em todas as minhas decisões.

Ao programa de Pós-graduação em genética pelo apoio e oportunidades fornecidas à minha formação acadêmica.

Ao departamento de Morfologia, ao instituto de Biociências e a UNESP de Botucatu pela estrutura para realização deste trabalho.

Aos professores membros da banca de qualificação e de defesa pelo seu tempo e contribuições para minha dissertação.

À FAPESP, CNPq e CAPES pelo auxílio financeiro.

Índice

Resumo	6
Abstract	7
1. Introdução	8
2. Objetivos	13
2.1 Objetivo Geral	13
2.2 Objetivos Específicos	13
3. Materiais e métodos	14
3.1 Material biológico	14
3.2 Análise dos reads paired-end no genoma B+	14
3.3 Seleção de regiões representativas do gene hnRNP Q-like	15
3.4 Extração de DNA	15
3.5 Obtenção de sequências	16
3.5.1 Reação de PCR	16
3.5.2 Montagem <i>de novo</i>	17
3.6 Análise das sequências obtidas	17
3.7 Busca por elementos transponíveis e domínios proteicos	17
3.8 Validação de genes de referência para análises de PCR quantitativa (qPCR)	18
3.9 PCR quantitativa (qPCR) em DNA genômico	18
3.10 PCR quantitativa (qPCR) em cDNA	19
3.11 Análise estatística	19
4. Resultados e discussões	20
The hnRNP Q-like gene is retroinserted in B chromosomes of the cichlid fish <i>Astatotilapia latifasciata</i>	21
5. Conclusões	33
6. Referências bibliográficas	34

Resumo

Cromossomos B, também conhecidos como supernumerários, são considerados elementos dispensáveis ao organismo. Porém, alguns estudos têm apresentado evidências de uma possível funcionalidade destes cromossomos extras. São encontrados em muitas espécies de eucariotos, entre elas, *Astatotilapia latifasciata*, ciclídeo africano que possui de 1 a 2 cromossomos B. Os ciclídeos têm sido muito utilizados como organismos modelo para estudos genéticos e genômicos. Análises genômicas em larga escala prévias envolveram sequenciamento por next generation (Illumina HiSeq sequencing) de genomas sem cromossomo B (B-) e com cromossomo B (B+) de *A. latifasciata* e uma forma variante do gene hnRNP Q-like (Heterogeneous nuclear ribonucleoprotein Q-like) foi identificada. Esta sequência apresenta um alto número de cópias e características de retrogene no cromossomo B. Diante disso, foi objetivo do presente trabalho compreender a organização genômica, identificar os mecanismos de retro-inserção e fazer análises de expressão desta sequência, possibilitando um melhor entendimento da origem, evolução e possíveis efeitos deste cromossomo na espécie estudada. A partir de análises de bioinformática foram identificadas cópias variantes do gene hnRNP Q-like no genoma B+ e as regiões mais representativas foram selecionadas, de acordo com o grau de mutações, para demais estudos. Análises de fragmentos sequenciados e qPCR confirmaram que o gene possui maior número de cópias e ausência de íntrons no cromossomo B. Evidências de resquícios de domínios da transcriptase reversa de elementos transponíveis sugerem que o gene foi retroinserido no cromossomo B. Dados de qPCR em cDNA mostraram que, apesar de possuir mais cópias no genoma B+, o gene hnRNP Q-like não é diferencialmente expresso nos genomas B+ e B-.

Abstract

B chromosomes, also known as supernumerary, are considered dispensable elements to organisms. However, some studies have presented evidences of a possible functionality of this extras chromosomes. They are observed in many eukaryote species, such as in the African cichlid, *Astatotilapia latifasciata*, which might have one or two B chromosomes. Cichlids have been used as model organisms to genetics and genomics studies. Previous genomic analyses based on next generation sequencing (Illumina HiSeq sequencing) were realized in the whole genome without B chromosome (B-) and with B chromosomes (B+) of *A. latifasciata* and a variant form of hnRNP Q-like (Heterogeneous nuclear ribonucleoprotein Q-like) gene was identified. This sequence presented a high copy number and characteristics of a retrogene in B chromosomes. Thus, the aim of this study consists in comprehend genomic organization, identify retroinsertion mechanisms and perform expression analyses of the hnRNP Q-like gene, making it possible a better understanding of origin, evolution and possible effects of this chromosomes in the studied species. Bioinformatics analyses identified variant copies of hnRNP Q-like gene in B+ genome and the higher and lower variable regions of the gene were selected, based in mutation rates, for further analysis. Analyses from sequenced fragments and qPCR confirmed gene has a higher number of copies and do not present introns in B chromosome. Evidences of transposable element relics with reverse transcriptase domains suggest gene was retroinserted in B chromosome. The data of qPCR in cDNA showed that the hnRNP Q-like gene is not differentially expressed in both genomes (with and without B chromosome).

1. Introdução

Cromossomos B, também conhecidos como supernumerários, acessórios, ou extras, são elementos dispensáveis encontrados em ampla diversidade de espécies de eucariotos, tais como animais, plantas e fungos. Acredita-se que muitos dos cromossomos supernumerários provavelmente originaram-se dos cromossomos do complemento A, mas seguiram seus próprios passos evolutivos (revisado por Beukeboom, 1994), e não recombina com membros do conjunto básico dos cromossomos do complemento A. A transmissão dos cromossomos B não segue o padrão Mendeliano clássico, apresentando comportamento variável durante a meiose e gametogênese. Possuem um mecanismo de acúmulo por meio da não-disjunção cromossômica na primeira divisão da meiose e segregação preferencial para o núcleo que formará os gametas (Jimenez et al. 1997).

Cromossomos B foram descritos em muitas espécies, entre elas, os peixes, incluindo representantes da família Cichlidae (Perciformes). A família Cichlidae representa a mais rica família dos vertebrados e possui mais de 3.000 espécies distribuídas na América Central e do Sul, África, Madagascar e sul da Índia. Apresentam uma fauna diversa que sofreu um rápido processo de radiação e especiação simpátrica, especialmente as espécies dos Grandes Lagos do leste da África, que têm atraído maior atenção. Esta família é composta por quatro subfamílias: Etroplinae, presentes em Madagascar e Sul da Ásia; Ptychochrominae, de Madagascar; Cichlinae, linhagens da região Neotropical; e Pseudocrenilabrinae, linhagens africanas. A subfamília Cichlinae foi reconhecida como monofilética e foi dividida em sete tribos: Astronotini, Chaetobranchini, Cichlasomatini, Cichlini, Geophagini, Heroini e Retroculini (Smith et al. 2008).

Estudos em ciclídeos têm promovido importantes descobertas em genética, evolução e especiação devido a rápida radiação adaptativa sofrida pelo grupo nos lagos do leste da África, onde centenas de espécies fortemente relacionadas, mas fenotipicamente diversas, se originaram a partir de uma ou poucas espécies. Esta radiação ocorreu em um curto período de tempo e os modelos clássicos de evolução não explicam este rápido processo, o que coloca os ciclídeos em destaque nas discussões sobre mecanismos de especiação (Kocher, 2004). Entre os ciclídeos, cromossomos B foram descritos em algumas espécies da América do Sul, como *Gymnogeophagus balzanii*, *Geophagus brasiliensis*, *Cichlasoma paranaensis*, *Crenicichla niederleini*, *Cichla monoculus*, *Cichla* sp. e *Crenicichla reticulata* (Poletto et al. 2010b). Entre os ciclídeos africanos, Bs foram descritos primeiramente em *Astatotilapia latifasciata* do Lago Nawampasa, um lago satélite do Lago Kyoga e em *Metriaclima*

lombardoi do lago Malawi no leste da África (Poletto et al. 2010a, 2010b). Bs foram também identificados em 12 espécies de ciclídeos africanos no Lago Victoria (Yoshida et al. 2011; Kuroiwa et al. 2014). Além da rápida radiação evolutiva, os ciclídeos africanos são ainda importantes devido à uma boa quantidade de recursos genéticos/genômicos (genomas sequenciados, transcriptomas, miRNomas, mapas genéticos, mapas cromossômicos e dados citogenéticos) disponíveis. Foram sequenciados recentemente cinco genomas de ciclídeos africanos, sendo estes a tilápia do Nilo (*Oreochromis niloticus*), *Neolamprologus brichardi/pulcher* (espécie de radiação antiga do Lago Tanganyika), *Metriaclima zebra* (radiação recente do Lago Malawi), *Pundamilia nyererei* (radiação muito recente do Lago Victoria) e *Astatotilapia burtoni* (espécie fluvial próxima ao Lago Tanganyika) (Brawand et al. 2014). As análises genômicas comparativas desses cinco novos genomas de ciclídeos mostraram uma acelerada evolução gênica, além de um acúmulo de duplicações gênicas, entre outras evidências associadas a rápida diversificação, corroborando a teoria de radiações adaptativas desse grupo (Brawand et al. 2014). Além disso, muitas destas espécies são reprodutivamente viáveis em laboratório (revisado por Kocher, 2004), o que facilita sua manutenção e realização de cruzamentos dirigidos para fins de análises diversas em genética.

Astatotilapia latifasciata apresenta espécimes sem nenhum, com um ou dois cromossomos B. Poletto et al. (2010a) e Fantinatti et al. (2011) realizaram análises mais detalhadas da estrutura dos cromossomos B presentes nesta espécie, pelo mapeamento de sequências de DNAs ribossomais, retrotransposons, elemento repetitivo SATA, BACs (Bacterial Artificial Chromosomes) ricos em sequências repetitivas e genes, fração genômica contendo DNA altamente repetitivo (*Cot*-1 DNA) e hibridização genômica comparativa (CGH). Estas análises mostraram a presença de uma grande quantidade de DNAs repetitivos nos cromossomos B desta espécie, sugerindo ainda a possível origem deste elemento a partir do primeiro par cromossômico do complemento. Além disso, estes autores sugerem que este cromossomo compartilha segmentos com cromossomos A contendo DNAr 18S. Em suma, sugere-se que o cromossomo supernumerário nesta espécie possa ter evoluído a partir de um pequeno fragmento cromossômico oriundo do genoma A, seguido pela invasão por várias famílias de DNA repetitivos (Poletto et al. 2010a; Fantinatti et al. 2011; Valente et al. 2014).

Os cromossomos B sempre foram estudados em nível citogenético, que geralmente forneciam informações de que estes cromossomos extras são ricos em elementos transponíveis, elementos repetitivos e DNAs satélites, e que possuem uma pequena quantidade de genes, além de serem considerados heterocromáticos. Havia muitas

dificuldades para definir a origem destes cromossomos dentro do genoma, visto que a maioria das conclusões eram baseadas em sequências repetitivas não codificantes e elementos transponíveis (Houben et al. 2013). Paralelamente às investigações em nível citogenético, outras abordagens baseadas em análises genômicas em larga escala estão sendo utilizadas para elucidar questões a cerca dos cromossomos B. Estas análises vêm demonstrando um cenário um pouco diferente do anterior, pôde-se inferir dados sobre as sequências genômicas, análises de expressão, análises evolutivas e estudos comparativos utilizando muitas espécies. Portanto, a partir dos dados de sequenciamento pôde-se obter dados mais informativos e precisos.

Zhou et al. (2012), utilizou dados do genoma total de *Drosophila* para inferir sobre aspectos de origem, evolução e conteúdo genômico do B. Abordagem utilizando dados genômicos em larga escala foi também empregada por Martis et al. (2012) na análise de cromossomos B de centeio, mostrando que os cromossomos acessórios desta espécie se originaram de diversos cromossomos do complemento A, seguido de captura adicional de sequências dos genomas A e mitocondrial e amplificação de sequências repetitivas. Banaei-Moghaddam et al. (2013), também analisaram o centeio e concluíram que os cromossomos B, nesta espécie, são ricos em sequências derivadas de genes, as quais são pseudogenes. Cerca de 15% destes fragmentos de pseudogenes presentes no cromossomo B são transcritos em um tipo de tecido e em genótipos específicos.

Análises genômicas em larga escala permitiram analisar genomas totais com (B+) e sem (B-) cromossomos B do peixe ciclídeo *Astatotilapia latifasciata* (Valente et al. 2014). Estas análises envolveram alinhamento de *reads* contra o genoma de referência do ciclídeo africano *Metriaclima zebra*, disponível no banco de dados BouillaBase (www.bouillabase.org). Foram identificadas regiões com diferença de cobertura entre os dois genomas, algumas das quais apresentaram uma média de cobertura 10.000 vezes maior no genoma B+ em relação ao genoma B-. Esses blocos genômicos foram anotados utilizando como base a anotação existente no genoma de referência e uma anotação de ontologia. Uma grande quantidade de genes (fragmentados ou intactos) foi encontrada no cromossomo B.

Entre as sequências gênicas identificadas presentes no cromossomo B de *A. latifasciata*, destaca-se o gene hnRNP Q-like (*heterogeneous nuclear ribonucleoprotein Q-like*) que possui alto número de cópias e apresenta características de retrogene (contém apenas regiões de similaridade a éxons) no genoma B+ (ver Figura 1, tópico 4 (Materials and methods)). Retrotransposição é o processo no qual RNAs mensageiros são reversamente transcritos em

DNA e inseridos em uma nova posição no genoma. Retrocópias não possuem muitas características genéticas de seus parentais, tais como íntrons e elementos regulatórios (Pan e Zhang, 2009).

A expressão de genes em eucariotos é regulada em diferentes níveis, desde a transcrição à modificações pós-traducionais. Uma das famílias mais caracterizadas de proteínas de ligação ao RNA é o grupo das *heterogeneous nuclear ribonucleoproteins* (hnRNPs), formado por aproximadamente vinte proteínas. Membros desta família exercem importante papel no controle da expressão gênica e metabolismo de RNAs mensageiros (Quaresma et al. 2009).

As hnRNPs são proteínas multifuncionais, participam de todos os aspectos cruciais do processamento do RNA, incluindo o *splicing* do pré-RNA, exportação do RNA, localização, tradução e estabilidade. Também participam na regulação do *splicing* alternativo (AS), onde sítios de *splicing* contidos no pré-RNA são preferencialmente selecionados ou reprimidos para produzir diferentes isoformas de RNA. Este processo contribui muito para a diversidade do proteoma (Huelga et al. 2012).

A hnRNP Q é uma proteína predominantemente nuclear que exibe uma organização modular e muitas funções relacionadas ao RNA (Quaresma et al. 2009). Muitos estudos mostram sua relação com os genes *Period* (Per) que são importantes elementos envolvidos com processos transcripcionais e traducionais que geram o ritmo circadiano endógeno (Pendergast et al. 2010). A regulação circadiana do RNA do mPer3 (Mouse Period) requer a função cooperativa de 5'e 3' UTRs (*Untranslated Regions*). hnRNP Q se liga a ambas UTRs desencadeando a tradução e acelerando o decaimento de RNA do mPer3 (Kim et al. 2011). O *knockdown* de hnRNP Q, em camundongos, diminuiu a ligação direta de hnRNP Q com a região 5'UTR do RNA de mPer1, o que causou uma diminuição nos níveis de mPer1 e um pequeno atraso em sua expressão, sem alterar os níveis de RNA mensageiro (Lee et al. 2012).

O hnRNP Q representa uma sub-família de hnRNPs, com três isoformas de proteínas chamadas hnRNP Q1-Q3, que são originadas a partir do *splicing* alternativo de um único gene (Mourelatos et al. 2001). Por sua vez o gene hnRNP Q-like apresenta três isoformas em *Homo sapiens* e duas isoformas em *Mus musculus* (<http://www.uniprot.org>). Possui características de retrogene nas espécies *Macaca mulatta*, *Homo sapiens* e *Rattus norvegicus* e são considerados pseudogenes de acordo com o RetrogeneDB

(<http://retrogenedb.amu.edu.pl>) (Kabza et al. 2014).

O processo de retrotransposição representa uma importante força evolutiva para a criação de novos genes denominados retrogenes, pois estes não possuem íntrons (Huang et al. 2009). Este processo está diretamente relacionado aos elementos transponíveis (TEs, do Inglês *Transposable Elements*), os quais compreendem um vasto conjunto de sequências de DNA, que possuem a habilidade de mover-se para novas posições nos genomas tanto diretamente, pelo mecanismo corta/copia e cola (transposons – classe II), ou indiretamente, através de um RNA intermediário (retrotransposons – classe I) (Fedoroff, 2012).

A classe I representa os retrotransposons, que são produtos de eventos de transcrição reversa. A movimentação é intermediada pelo RNA e é replicativa, mantendo a sequência original. O RNA é transcrito reversamente em um DNA por um TE que possui o domínio da transcriptase reversa. São separados em cinco categorias: LTR (*Long terminal repeats*), DIRS-like, Penelope-like (PLEs), LINEs (*Long interspersed nuclear element*) and SINEs (*Short interspersed nuclear element*) (Wicker et al. 2007). Os elementos transponíveis de classe II replicam-se por meio de um DNA intermediário que pode ser do tipo transposição replicativa ou conservativa. Na transposição conservativa, o elemento é retirado durante a transposição e o número de cópias não aumenta. No processo replicativo, o elemento é duplicado, produzindo aumento do número de cópias no genoma. São divididos em duas categorias: a primeira é caracterizada pela presença de TIRs (*Terminal Inverted Repeats*) de comprimentos variáveis e a segunda compreende os TEs que durante o processo de transposição, a replicação ocorre sem a clivagem da dupla-fita de DNA, havendo apenas o deslocamento de uma das fitas (Wicker et al. 2007).

A hipótese do presente trabalho, de que o gene hnRNP Q-like foi inserido no cromossomo B por mecanismos de retroinserção e posteriormente sofreu uma duplicação, se baseia na premissa de que o gene hnRNP Q-like apresenta características de retrogene e um maior número de cópias no cromossomo B. Considerando que a presença de uma sequência com características de retrogene no cromossomo B é uma nova descoberta, uma melhor caracterização estrutural e funcional deste gene é necessária. Deste modo, foi analisado o perfil transcricional do gene hnRNP Q-like e sua organização genômica, fornecendo evidências de mecanismos de retroinserção e duplicação gênica em organismos portadores de cromossomo B.

2. Objetivos

2.1 Objetivo Geral

O presente trabalho teve por objetivo analisar o gene hnRNP Q-like na espécie *Astatotilapia latifasciata*, com o propósito de compreender sua organização genômica e seu nível transcricional.

2.2 Objetivos específicos

- Avaliar os níveis de variabilidade nucleotídica do gene hnRNP Q-like dos genomas B+ e B-;
- Analisar a sequência gênica para identificar características de retrogene;
- Buscar elementos transponíveis em suas regiões flanqueadoras para compreender os mecanismos de retroinserção;
- Avaliar o número relativo de cópias do gene hnRNP Q-like em indivíduos B+ e B-;
- Quantificar os níveis de expressão de RNA mensageiro em diferentes tecidos de indivíduos B+ e B-;

3. Materiais e métodos

3.1 Material biológico

A espécie utilizada como modelo deste estudo foi o ciclídeo *Astatotilapia latifasciata*, natural dos lagos Kyoga e Nawampasa (lagos satélites ao lago Vitória, leste africano) (Steeves et al. 2010), amplamente difundida na aquariofilia. A partir de análises prévias de citogenética (Poletto et al. 2010a, 2010b; Fantinatti et al. 2011) foram formadas populações cruzantes de indivíduos portadores de cromossomos B que estão sendo mantidos no biotério de peixes do Laboratório Genômica Integrativa, UNESP/IB, Botucatu.

A identificação de indivíduos portadores e não portadores de cromossomo B foi realizada por PCR (*Polymerase Chain Reaction*), utilizando marcadores moleculares específicos para indivíduos B+ e B- (BEA Fantinatti, comunicação pessoal).

3.2 Análise dos reads paired-end no genoma B+

As análises realizadas no presente trabalho partiram de dados de sequenciamento de próxima geração (NGS) realizados utilizando plataforma Illumina. Para a obtenção destes dados de sequenciamento, DNA genômico dos indivíduos de interesse foram fragmentados a uma média de 500 pb usando um ultrasonicador S229 (Covaris Inc., Woburn, MA). Bibliotecas separadas para amostras B+ e B- foram construídas para cada indivíduo, sendo um indivíduo B+ e um indivíduo B-, usando o kit TruSeq DNA sample preparation ver.2 rev.C (Illumina Inc., San Diego, CA). Procedimento de sequenciamento paired-end (2x100 bp) foi realizado para cada biblioteca em lanes separados no sequenciador Illumina HiSeq 1000. Os dados genômicos gerados foram alinhados contra o genoma referência de *Metriaclima zebra* para fins de anotação gênica. Genomas totais de indivíduos B+ e B- foram gerados produzindo um total de 306.823.512 de *reads* para o indivíduo B+ e 401.017.570 de *reads* para o indivíduo B- (Valente et al. 2014). Estas informações estão disponíveis na base de dados Sacibase (www.sacibase.ibb.unesp.br).

O genoma B+ foi analisado com a utilização do software Tablet (1.12.12.05) (Milne et al. 2010), no qual os *reads paired-end* foram considerados individualmente e comparados à referência para identificação de cópias variantes do gene hnRNP Q-like no genoma B+.

3.3 Seleção de regiões representativas do gene hnRNP Q-like

O retrogene hnRNP Q-like apresenta doze regiões exônicas (Ver Figura 1, tópico 4 (Materials and Methods)) das quais onze foram analisadas. Com o intuito de selecionar as regiões mais representativas, no presente estudo, foi utilizado o software Tablet (1.12.12.05) (Milne et al. 2010) para comparar a presença ou ausência de mutações por posição no retrogene hnRNP Q-like em um indivíduo B- e um indivíduo B+, da espécie *Astatotilapia latifasciata*, tendo como genoma de referência o ciclídeo *Metriacroma zebra*, ambos disponíveis no BouillaBase (www.bouillabase.org).

A análise foi realizada nas sequências correspondentes aos éxons, a partir da seguinte equação:

$$MR1 = \frac{\text{Quantidade de mutações/Comprimento do éxon}}{\text{Média de cobertura}}$$

A quantidade de variações refere-se a ausência ou presença de mutações em cada posição (sendo que a ausência = 0 e presença = 1). Para normalização, esse valor foi dividido pelo comprimento da região exônica correspondente, e posteriormente dividido pela cobertura média de seu respectivo genoma (B- 30,99x e B+ 22,92x) de acordo com Valente et al. (2014). A partir de MR1 foi calculado o valor delta, que representa a diferença dos valores de MR1 de B+ e B- para cada uma das onze regiões relativas aos éxons.

$$\Delta = MR1 (\text{genoma B +}) - MR1 (\text{genoma B-})$$

As regiões selecionadas foram aquelas com maior variabilidade (maior delta) e menor variabilidade (menor delta), com o intuito de obter uma análise mais precisa.

3.4 Extração de DNA

O grupo experimental foi composto por amostras de indivíduos sem (B-) e com (B+) cromossomos B, incluindo machos e fêmeas para ambas condições.

Amostras de fígado e nadadeiras foram coletadas, fixadas em etanol 100% e armazenadas a -20°C para a extração de DNA total utilizando a metodologia fenol-clorofórmio descrita por Sambrook e Russel (2001). A análise qualitativa e quantitativa do DNA foi realizada por meio de eletroforese em gel de 1% de agarose e espectrofotometria.

3.5 Obtenção de sequências

Para a realização de estudos evolutivos, sequências do gene hnRNP Q-like foram obtidas por reação de PCR a partir das amostras de DNA B- e B+. Inicialmente, foram realizadas PCRs das regiões correspondentes aos éxons 2 e 10. Porém, este método não apresentou os resultados esperados, pois as sequências obtidas não representavam a diversidade das cópias variantes já detectadas via dados de NGS. Assim foi realizada a montagem *de novo* (ver descrição detalhada no item 3.5.2) a partir dos dados de NGS para obtenção do gene hnRNP Q-like.

3.5.1 Reação de PCR

Conjuntos de *primers* foram construídos para amplificar por PCR os segmentos exônicos do retrogene hnRNP Q-like, os quais foram selecionados por serem regiões representativas em função de mutações previamente analisadas (Tabela 1; Ver Figura 2 Resultados (4.1)). Os *primers* foram situados nas regiões dos éxons adjacentes, cobrindo toda a região do éxon de interesse e uma parte dos éxons mais próximos, para certificar que o éxon seria totalmente amplificado. Foram utilizadas amostras de seis indivíduos B+ e seis indivíduos B-, machos e fêmeas, para ambas as condições.

Éxon	Primer ID	Sequência Primer	Posição no genoma de <i>M. Zebra</i>
2 (B+)	Pic1-2 cF	5'GAGTGAATGACTGAAGCC	scaffold_3:9114625..9114642
	Pic2-3 aR	5'CTCAATCGCTCGATCATC	scaffold_3:9117657..9117674
2 (B-)	Primer F –Pico2	5'GCACATCACAATTGTTGCC	scaffold_3:9116465..9116483
	Primer R –Pico2	5'CTGTAGCACTGCAGTACCCATT	scaffold_3:9116819..9116840
10	Pic9-10 cF	5'TTTAGCCAACACGGCAAGC	scaffold_3:9120036..9120054
	Pic10-11 aR	5'AATAACCTCCTCTACCGC	scaffold_3:9120551..9120570

Tabela 1: Conjuntos de primers projetados a partir dos dados genômicos para animais B- e B+ para regiões referentes aos éxons 2 e 10 do retrogene hnRNP Q-like de *A. latifasciata*.

Os *amplicons* de DNA obtidos foram clonados em vetores plasmidiais utilizando o kit pGEM-T Vector System (Promega) de acordo com as especificações do fabricante. Para verificar a presença dos insertos nos plasmídeos recombinantes foi realizada a reação de PCR utilizando o conjunto de primers M13, que flanqueiam a região de inserção no plasmídio.

Os clones recombinantes de cada região foram submetidos ao sequenciamento pelo aparelho ABI Prism 3100 DNA sequencer (Perkin- Elmer) utilizando o kit ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction (Perkin- Elmer).

3.5.2 Montagem *de novo*

A partir dos resultados dos sequenciamentos dos genomas B+ e B-, sendo uma amostra B+ e outra B-, os *reads* referentes ao gene hnRNP Q-like foram extraídos utilizando o software SAMtools (Li et al. 2009). Foi realizada, separadamente, em cada genoma, a montagem *de novo* por meio dos softwares Velvet (Zerbino et al. 2008) e Oases (Schulz et al. 2012). As sequências resultantes foram alinhadas contra o genoma de referência por meio do software Bowtie2 (Langmead e Salzberg 2012) para confirmar se as sequências eram correspondentes as regiões de interesse.

3.6 Análise das sequências obtidas

As sequências obtidas foram analisadas utilizando o software Geneious Pro 4.8.5. (Drummond et al. 2009). Todas as sequências, B+ e B-, foram alinhadas contra o genoma de referência. Aquelas que possuem mutações diferentes das sequências B- e do genoma de referência, foram classificadas como B específicas.

O grau de saturação das sequências foi determinado através do *software* DAMBE (Xia et al. 2001), para fins de análises filogenéticas posteriores.

3.7 Busca por elementos transponíveis e domínios proteicos

Uma busca por transposable elements (TEs) foi realizada nas regiões flangeadoras (cerca de 10.000pb) do gene hnRNP Q-like, as quais foram usadas como *query* contra o banco de dados do Repbase (Jurka et al. 2005) pertencente ao Genetic Information Research Institute (GIRI) (<http://www.girinst.org>).

As sequências com similaridade maior que 80% aos elementos transponíveis foram analisadas utilizando a ferramenta assembly pertencente ao *software* Geneious Pro 4.8.5 (Drummond et al. 2009), e foi realizada uma busca por ORFs (*Open Reading Frames*) cujas sequências de aminoácidos foram submetidas ao banco de dados Pfam (Finn et al. 2010).

3.8 Validação de genes de referência para análises de PCR quantitativa (qPCR)

A eficiência dos primers utilizados em qPCR foi calculada usando o programa *LingRegPCR* (Ramakers et al. 2003). O gene mais apropriado para ser utilizado como gene de referência foi selecionado com uso da ferramenta *RefFinder* (Xie et al. 2011).

3.9 PCR quantitativa (qPCR) em DNA genômico

Para avaliar o número relativo de cópias do gene hnRNP Q-like, foi realizada a reação de PCR em tempo real (PCR quantitativa), com DNA genômico de vinte e quatro indivíduos B+ e B-, machos e fêmeas. Os *primers* foram construídos de acordo com as recomendações de Bustin et al. 2009, utilizando-se as ferramentas *do* Geneious Pro 4.8.5 (Drummond et al. 2009) e *PCR Primer Stats* (Stothard, 2000). As regiões amplificadas foram as correspondentes aos éxons 2 e 10 (Tabela 2). Foram construídos primers com mutações específicas do genoma B+, para avaliar o número de cópias existentes apenas no cromossomo B, porém estes primers apresentaram uma eficiência muito baixa, impossibilitando as realização de reações de qPCR. Portanto, os primers utilizados nas análises não apresentam as mutações específicas do cromossomo B e amplificam apenas as cópias canônicas presentes nos dois genomas. Inicialmente, foram realizados testes utilizando triplicatas e duplicatas, a partir da comparação dos resultados, todas as análises foram desenvolvidas a partir de duplicatas pois apresentaram uma menor variação, fornecendo dados mais precisos.

O valor de GDR (*gene dosage ratio*) foi obtido usando o método de quantificação relativa (ΔC_t), onde o número relativo de cópias do gene é calculado como $2^{-\Delta C_t}$, sendo $\Delta C_t = C_{t_{\text{target}}} - C_{t_{\text{reference}}}$ (Bel et al. 2011). O gene de referência utilizado foi o gene autossômico de cópia simples, hypoxanthine phosphoribosyltransferase (Hprt). A reação de qPCR foi realizada no StepOne Real-Time PCR Systems (Life Technologies, Carlsbad, CA, USA). O gene alvo e o de referência foram analisados simultaneamente em duplicatas. Os ciclos foram, 95 °C por 10 minutos; 45 ciclos de 95 °C por 15 segundos; e 60 °C por 1 minuto.

Éxon	Primer ID	Sequência Primer	Posição no genoma de <i>M. Zebra</i>
2	PIC2-RT-F	5'GGAAATGGTCCAGAAGAACC	scaffold_3:9116556..9116576
	PIC2-TR-R	5'CTGTGGTAAACCAGCTTC	scaffold_3:9116624..9116642
10	PIC10 –TR-F	5'GGAGAGCACATTGAAATTG	scaffold_3:9120259..9120276
	PIC10-TR-R	5'GTGTTTTAGCGGCTTGTCTC	scaffold_3:9120323..9120342

Tabela 2: Conjuntos de primers construídos a partir dos dados genômicos para regiões referentes aos éxons 2 e 10 do retrogene hnRNP Q-like de *A. latifasciata*.

3.10 PCR quantitativa (qPCR) em cDNA

Foram utilizadas vinte e quatro amostras de músculo, vinte e três amostras de cérebro e vinte e três amostras de coração de indivíduos B+ e B-, machos e fêmeas. A extração de RNA foi realizada utilizando o protocolo de TRIzol (Life Technologies, EUA). A quantificação do RNA extraído foi feita utilizando o espectrofotômetro NanoVue e sua integridade foi avaliada utilizando o 2100 Bioanalyzer.

O RNA obtido foi tratado com a enzima DNase, utilizando a DNase I - Amplification Grade (Life Technologies), e posteriormente foi realizada a reação de transcrição reversa (RT) utilizando-se o kit High Capacity (Life Technologies). As regiões amplificadas foram as correspondentes aos éxons 2 e 10 (Tabela 1). O gene de referência utilizado foi o UBCE (*ubiquitin-conjugatin enzyme*) que foi validado em trabalhos prévios com peixes ciclídeos (Deloffre et al. 2012; Yang et al. 2013). As taxas de expressão foram mensuradas, no sistema Step One Plus (Life Technologies, Carlsbad, CA, USA), utilizando o kit GoTaq qPCR Master Mix (Promega). Os ciclos foram: 10 minutos a 95°C, 40 ciclos de desnaturação a 95°C por 15 segundos e anelamento e extensão a 55°C por 1 minuto.

A normalização dos dados foi realizada usando o programa Q-Gene (Muller et al. 2002, Simon 2003).

3.11 Análise estatística

A análise estatística foi realizada pelo programa *GraphPad Prism 5* (GraphPad Software Inc., San Diego, CA, EUA). A distribuição dos dados foi não-paramétrica e o teste estatístico o teste U de Mann-Whitney foi aplicado. O nível de significância estatística usado foi de 5% ($p < 0.05$).

4. Resultados e discussão

O presente tópico será apresentado na forma de manuscrito a ser submetido a revista científica da área de genética

The hnRNP Q-like gene is retroinserted in B chromosomes of the cichlid fish *Astatotilapia latifasciata*

Carmello BO¹, Cardoso AL¹, Marques DF¹, Valente GT², Martins C¹

¹Department of Morphology, Institute of Biosciences, UNESP – Sao Paulo State University, Botucatu, São Paulo, 18618-970, Brazil

²Department of Bioprocess and Biotechnology, Agronomic Science School, UNESP – Sao Paulo State University, Botucatu, São Paulo, 18610-307, Brazil

Corresponding author – Dr. Cesar Martins, Department of Morphology, Institute of Biosciences, Sao Paulo State University, Botucatu, Sao Paulo, 18618-970, Brazil. Phone: 55 14 38800462, E-mail cmartins@ibb.unesp.br, cmartinsunesp@gmail.com

Abstract

B chromosomes, also known as supernumerary, are considered dispensable elements and are observed in many eukaryote species, such as in the African cichlid, *Astatotilapia latifasciata*, which might have one or two B chromosomes. Previous genomic analyses based on next generation sequencing (Illumina HiSeq sequencing) were realized in the whole genome without B chromosome (B-) and with B chromosomes (B+) and a variant form of hnRNP Q-like gene (named hnRNP Q-B) was identified. This sequence presented a high copy number and characteristics of a retrogene in B chromosomes. Thus, the aim of this study consists in comprehend genomic organization, identify retroinsertion mechanisms and perform expression analyses of the hnRNP Q-like gene. Bioinformatics analyses identified variant copies of hnRNP Q-like gene in B+ genome and the higher and lower variable regions of the gene were selected, based in mutation rates, for further analysis. Analyses from sequenced fragments and qPCR confirmed gene has a higher number of copies and do not present introns in B chromosome. Evidences of transposable element relics with reverse transcriptase domains suggest gene was retroinserted in B chromosome. The data of qPCR in cDNA was analyzed and results showed the canonical hnRNP Q-like copy is not differentially expressed in both genomes (with and without B chromosome), showing that B+ genome has more copies of hnRNP Q-like gene, but they are not expressed.

Introduction

B chromosome, also known as supernumerary, accessory or extra, are dispensable elements found in many eukaryote species, such as animals, plants and fungi. Transmission of B chromosome does not follow Mendelian laws of inheritance. It is believed many of supernumerary chromosomes probably were originated from the chromosomes of the A complement, but followed their own evolutionary steps without recombination with members from A complement (reviewed by Beukeboom, 1994).

B chromosomes were described in many fish species, including representatives of the cichlid family (Perciformes), which is the richest family of vertebrates and has more than 3,000 species distributed in Central and South America, Africa, Madagascar and South India. They present a diverse fauna that suffered a rapid process of sympatric speciation, especially species from the Great Lakes of East Africa, which have attracted more scientific attention (Kocher, 2004). Cichlids are good models for genetic analysis because their evolutionary radiation. There are genomic resources (whole genomes sequenced, transcriptomes, miRNomes, genetic maps, chromosome maps) and many of these species are reproductively viable in laboratory (reviewed by Kocher, 2004). Among African cichlids, B chromosomes were first described in *Astatotilapia latifasciata* from Lake Nawampasa, a satellite lake of the Lake Kyoga (part of Lake Victoria system) and in *Metriaclima lombardoi* from Lake Malawi in East Africa (Poletto et al. 2010a, 2010b). Later, Bs were identified in twelve species of African cichlids of Lake Victoria (Yoshida et al. 2011; Kuroiwa et al. 2014).

Astatotilapia latifasciata can harbor 0, 1 or 2 B chromosomes that have been the subject of previous classical and molecular cytogenetics studies. It is suggested supernumerary chromosome in this species might have evolved from a small chromosome fragment from A complement followed by invasion of many families of repetitive DNA (Poletto et al. 2010a; Fantinatti et al. 2011). Besides the investigations in cytogenetic level, large scale genomic analyses were conducted involving next generation sequencing (NGS) of whole genomes of *A. latifasciata* with B chromosome (B+) and without B chromosome (B-) and a microdissected B. Sequenced reads were aligned to the *Metriaclima zebra* reference assembly, available in BouillaBase database (www.bouillabase.org). Regions differentially covered were identified, some of that presented a mean coverage of 10,000x higher in B+ genome compared to B- genome and were referred as B chromosome “blocks”. This study showed B chromosome of *A. latifasciata* contains thousands of sequences duplicated from essentially every chromosome in the ancestral karyotype, including transposable elements

and genes. Although most genes on the B chromosome are fragmented, a few are largely intact. Among the identified B-sequences, hnRNP Q-like (*heterogeneous nuclear ribonucleoprotein Q-like*) gene was highlighted due to its duplicated number of copies and characteristics of retrogene (Valente et al. 2014).

Considering the presence of a retrogene-like sequence in B chromosomes is a new finding, a better structural and functional characterization of this gene is necessary. In this way, we analyzed the hnRNP Q-like transcriptional profiles and genomic organization, providing evidences of the retroinsertion and gene duplication mechanisms of this retrogene in B carrier organisms.

Materials and Methods

Genomic analyses

Illumina HiSeq sequencing data of whole B⁺ and B⁻ genomes were previously generated for the African cichlid *Astatotilapia latifasciata*. The generated data contain 306,823,512 reads for the B⁺ sample and 401,017,510 reads for the B⁻ sample. The focus of the analysis was directed to a specific region of scaffold 3 containing a retrogene-like sequence derived from the hnRNP Q-like gene (Valente et al. 2014). There are twelve regions corresponding to exons of the hnRNP Q-like gene in the B chromosome (Figure 1). The reads and alignments against the reference genome of *Metriaclicma zebra* cichlid are available in Sacibase (www.sacibase.ibb.unesp.br).

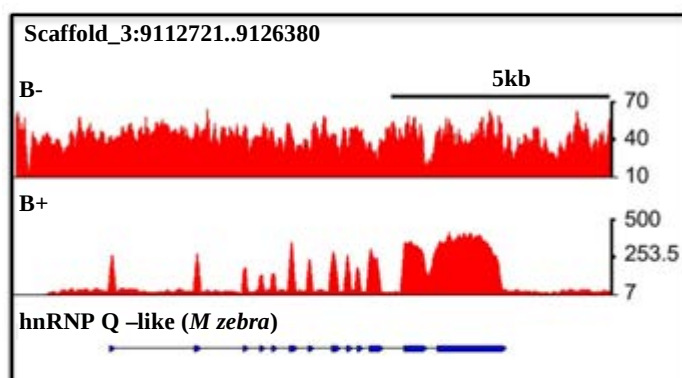


Figure 1: Comparative analysis of hnRNP Q-like gene in B⁺ and B⁻ genomes based in coverage of reads obtained by NGS. Scale in the right indicates number of reads obtained for specific regions. Regions correspondents to exons are indicated below. Modified from Valente et al. (2014).

The B+ paired-end reads were analyzed individually and compared to reference genome (*M. zebra*) to identify the variants copies of hnRNP Q-like gene in B+ genome, using Tablet (1.12.12.05) (Milne et al. 2010). The next step was directed to identification of nucleotide variation in the exonic regions of hnRNP Q-like gene. Comparative analysis based on Tablet software (1.12.12.05) (Milne et al. 2010) focus in the identification of presence or absence of mutation per position in hnRNP Q-like in B+ and B- genomes of *A. latifasciata*, using as reference the genome of *M. zebra*. The presence or absence of point mutations was counted as 1 or 0, respectively. For normalization, the total number of point mutations was divided by the correspondent exonic region length, and this result was divided by the mean read coverage of each correspondent genome (B- = 30,99x and B+ =22,92x coverage). The result was named as MR1 and the delta was calculated by difference between MR1 of B+ and B- for each region. Regions with higher variability (higher delta) and lower variability (lower delta) were selected as the most representatives ones.

DNA samples

DNA samples from B+ and B- individuals were extracted from tissues using the phenol–chloroform method (Sambrook and Russel, 2001) and submitted to polymerase chain reaction (PCR), using specific primers for the B chromosome identification (BEA Fantinatti, personal communication). Animal samples were obtained from the fish facility of the Integrative Genomics Laboratory, Sao Paulo State University, Botucatu, Brazil.

Obtaining sequences: PCR and de novo assembly

For phylogenetic studies, exonic segments of hnRNP Q-like retrogene were amplified by PCR using primers based in representative regions with high mutation levels previously identified. The amplicons were cloned in pGEM-T Vector System (Promega), and the 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).

Additionally, sequences of the hnRNP Q-like gene were obtained from the whole genome read datasets of B+ and B- genomes. The reads of hnRNP Q-like, from both genomes, were extracted using SAMtools (Li et al. 2009) and then were assembled using *de novo* approaches with Velvet (Zerbino et al. 2008) and Oases (Schulz et al. 2012). The contigs were aligned to the *M. zebra* reference genome using Bowtie2 (Langmead and

Salzberg 2012) software. Additionally, with Geneious Pro 4.8.5 (Drummond et al. 2009) software, the B specific mutations were identified and were used to classify contigs as belonging to A or B chromosomes for further analyses. Each sequence was aligned against reference genome and the result of the alignments was submitted to DAMBE (Xia et al. 2001) software to verify the nucleotide saturation to begin the phylogenetic studies.

Search for TEs

Flanking regions (FR) of the hnRNP Q-like gene were used as queries to search for transposable elements (TEs) against the Repbase (Jurka et al. 2005) database at the Genetic Information Research Institute (Giri) (<http://www.girinst.org/repbase/>). A search for open reading frames (ORFs) was performed using Geneious Pro 4.8.5 (Drummond et al. 2009), in an attempt to identify the potential protein coding regions of the FRs. ORFs correspondents to sequences of TEs were submitted to Pfam (Finn et al. 2010) database to obtain more information about the TEs.

qPCR – Genomic DNA

To evaluate the relative copy number of hnRNP Q-like gene, the quantitative real-time PCR (qPCR) was performed with genomic DNA from B + and B- individuals. The value of GDR (gene dose ratio) was obtained using the method of relative quantification (ΔC_t), the relative gene copy number was calculated as $2^{-\Delta C_t}$, where $\Delta C_t = C_{t_{\text{target}}} - C_{t_{\text{reference}}}$ (Bel et al. 2011). The single-copy autosomal gene, hypoxanthine phosphoribosyltransferase (Hprt) was used as reference. qPCR was performed in StepOne Real-Time PCR Systems (Life Technologies, Carlsbad, CA, USA). The target and the reference gene were analyzed simultaneously in duplicates. The cycles were 95 ° C for 10 minutes; 45 cycles of 95 ° C for 15 seconds; and 60 ° C for 1 minute. Primers efficiency was calculated by the *LingRegPCR* program (Ramakers et al. 2003).

qPCR – cDNA (Complementary DNA)

The qPCR was performed in samples of cDNA, of muscle, brain and heart of B+ and B- individuals, males and females. The extraction was done using the TRIzol protocol (Life Technologies, EUA). The regions amplified were that correspondents to exons 2 and 10.

The UBCE (*ubiquitin-conjugatin enzyme*) gene was used as reference. qPCR was performed in StepOne™ Real-Time PCR Systems (Life Technologies, Carlsbad, CA, USA). The target and the reference gene were analyzed simultaneously in duplicates. The cycles were 95 °C for 10 minutes; 40 cycles of 95 °C for 15 seconds; and 55 °C for 1 minute. Primers efficiency was calculated by the *LingRegPCR* program (Ramakers et al. 2003). The normalization of data was performed using the program Q-Gene (Muller et al. 2002, Simon 2003).

Results

Genomic analyses

Analyses from paired-end reads from B+ genome identified one canonical hnRNP Q-like sequence identical to the A genome copy of hnRNP Q-like gene, and several variant copies. Nucleotide diversity analyses of hnRNP Q-like gene sequences showed B+ genome presents higher mutation ratios compared to B- genome (Figure 2). The nucleotide diversity was analyzed independently for each of the hnRNP Q exons. The attention was directed to the exons 2 and 10, that presented lower and higher delta values, respectively (Figure 2).

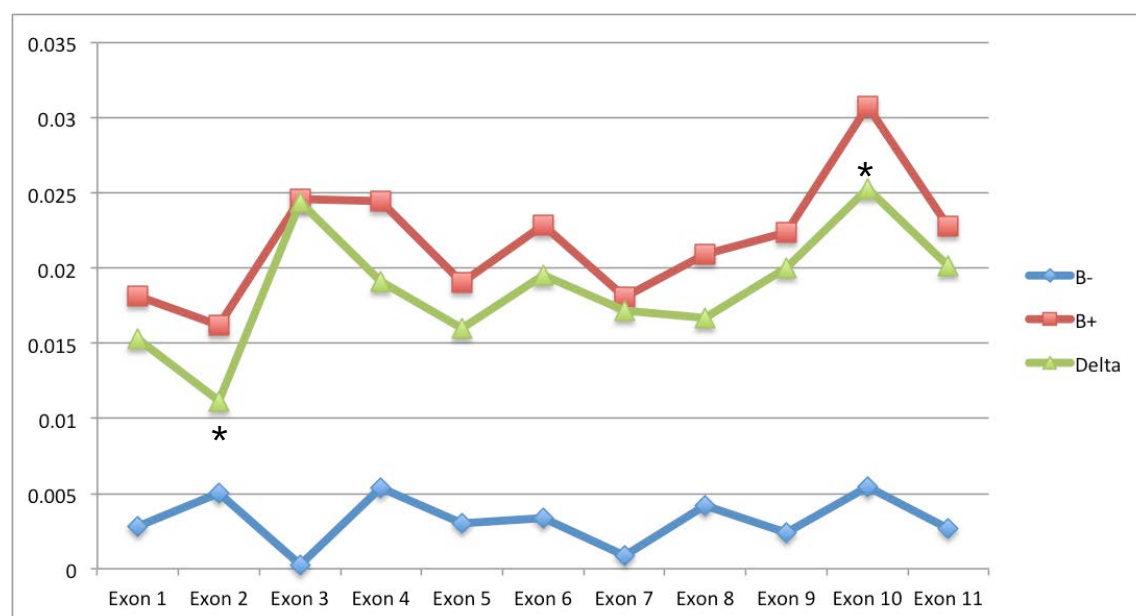


Figure 2: Nucleotide diversity (axis x) of exons (axis y) from hnRNP Q-like gene in B+ (red line) and B- (blue line) genome. Most representative regions (*) with lower and higher delta values.

Sequences correspondent to exons 2 and 10 (and flanking exons of both regions) were obtained by PCR, cloned and sequenced, which showed absence of introns in B+ sequences. In exon 2, PCR with the same set of primers failed to amplify a fragment in B- genome, probably, because of the presence of a long intron (~3.000pb, based in the *M. zebra* genome). New primers were constructed nearest to exon 2 region, and a fragment of ~380pb was amplified. Exon 10 was amplified in both genomes (Figure 3).

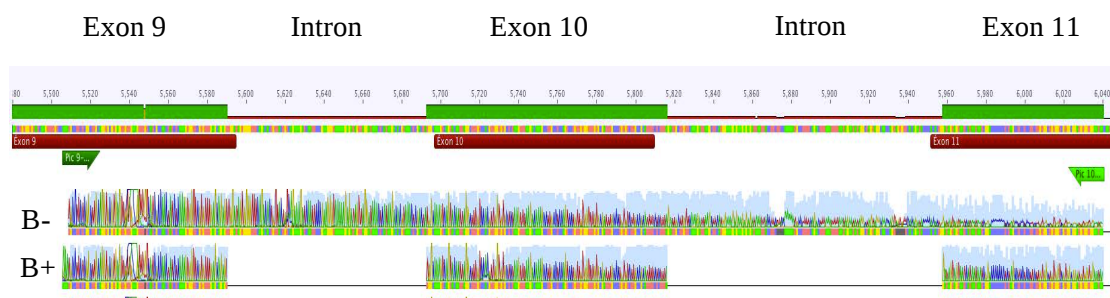


Figure 3: Genomic organization of a region involving the exons 9, 10 and 11 of hnRNP Q-like gene. Alignment indicates the presence of introns in B- sequence and the absence of introns in B+ sequences.

Sequences generated from PCR did not present enough mutations for phylogenetic studies, so it was done the *de novo* assembly that provides a high number of sequences and can recuperate mutations that do not exist in the genome of reference.

Results from *de novo* assembling of B- genome recovered only one contig corresponding to whole hnRNP Q-like gene. For B+ genome, many contigs were obtained, probably due mutation in hnRNP Q-like copies of B chromosome. Prior to define if contigs belongs to A or B genomes, they were filtered based in specific mutations that are not detected in reference genome neither at B- genome. Out of 165 contigs, 65 were classified as B specific (Figure 4).

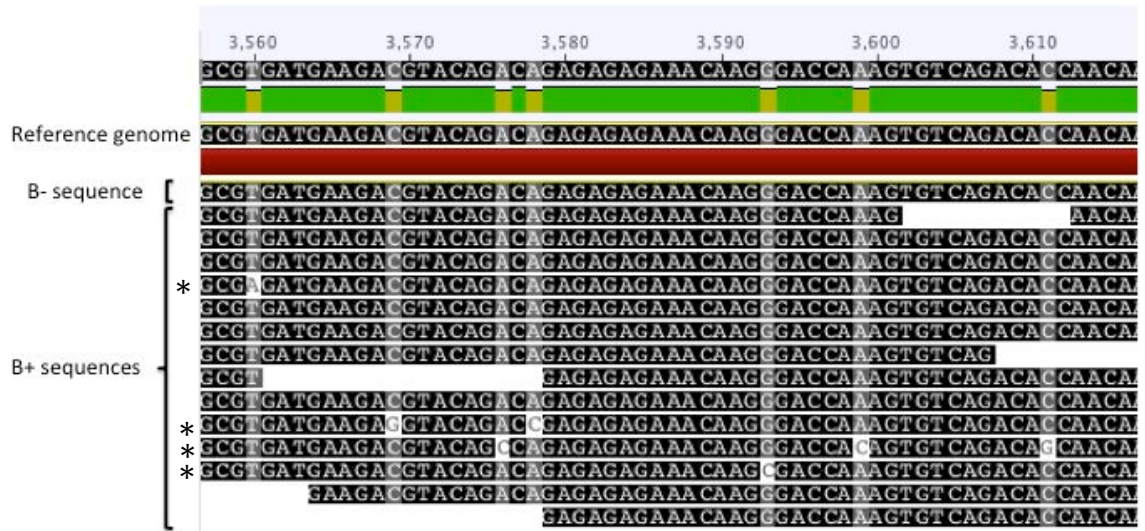


Figure 4: Alignment of resultants contigs from de novo assembly, of the region correspondent to exon 4, against reference genome, sequence of B- genome and variants copies (*) and B+ sequences of B+ genome.

Alignment of this contigs was exported to DAMBE (Xia et al. 2001), showing sequences have a high level of saturation, indicating the absence of phylogenetic signals. Thus, phylogenetic analysis does not make any sense.

A search for transposable elements (TEs) found evidences of TEs relics at low level of integrity and some of that have reverse transcriptase nucleotide sequences (Table 1).

Region	Name	Class	Sim	Pos/Mm:Ts	Score
Upstream	L1-112_DR	NonLTR/L1	0.8148	1.6000	241
	L1-3_STu	NonLTR/L1	0.8478	2.0000	242
	Mariner-8_SSa	DNA/Mariner	0.8038	2.1667	653
	hAT-N27_LMi	DNA/hAT	0.8537	1.3333	201
	DONG_FR2	NonLTR/R4	0.8624	1.6250	637
	Harbinger 9_PGr	DNA/Harbinger	0.8158	1.4000	217
Downstream	Gypsy-63_PAb-I	LTR/Gypsy	0.875	1.3333	219
	TC1_FR1	DNA/Mariner	0.8083	1.65	2919
	I-8_DBp	NonLTR/I	0.8163	1.75	209

Table 1: Transposable elements with similarity higher than 80% detected in the search in the flanking regions of hnRNP Q-like. Sim, means value of similarity between 2 aligned fragments; Pos. is the ratio of positives to alignment length; Mm:Ts, is the ratio of mismatches to transitions in nucleotide alignment; Score, is alignment score obtained from blast.

The relative copy number of hnRNP Q-like gene was estimated by qPCR showing higher GDR in B+ individuals compared to B- individuals. The B+ have the double of copies of hnRNP Q-like gene, in exons 2 and 10, compared to B- (Figure 5).

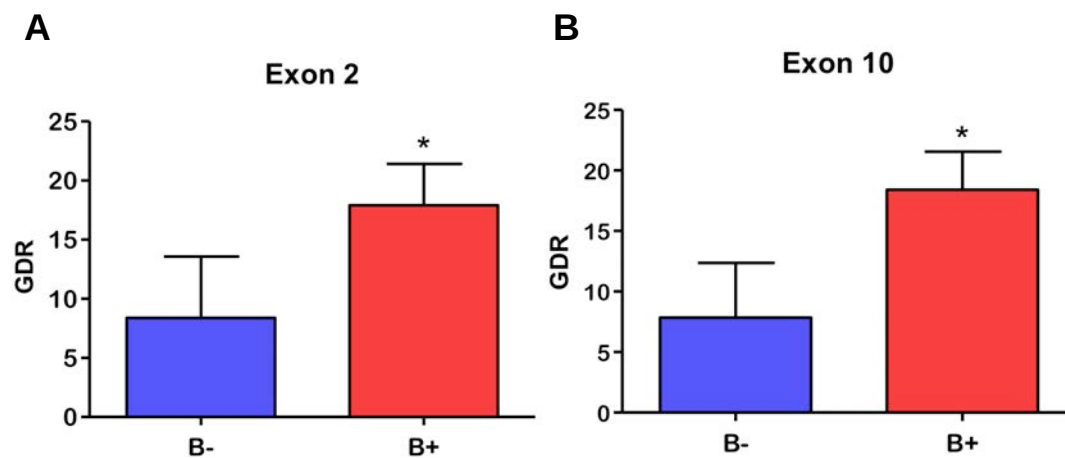


Figure 5. GDR estimated by qPCR of hnRNP Q-like gene in exons 2 (A) and 10 (B) in B- and B+ genomes. (*) GDR statistically significant. N = 24.

Expression analyses

The expression levels were quantified by qPCR, in samples of cDNA of muscle (Figure 6), brain (Figure 7) and heart (Figure 8) of B+ and B- individuals. There is no difference statistically significant, between B+ and B- samples.

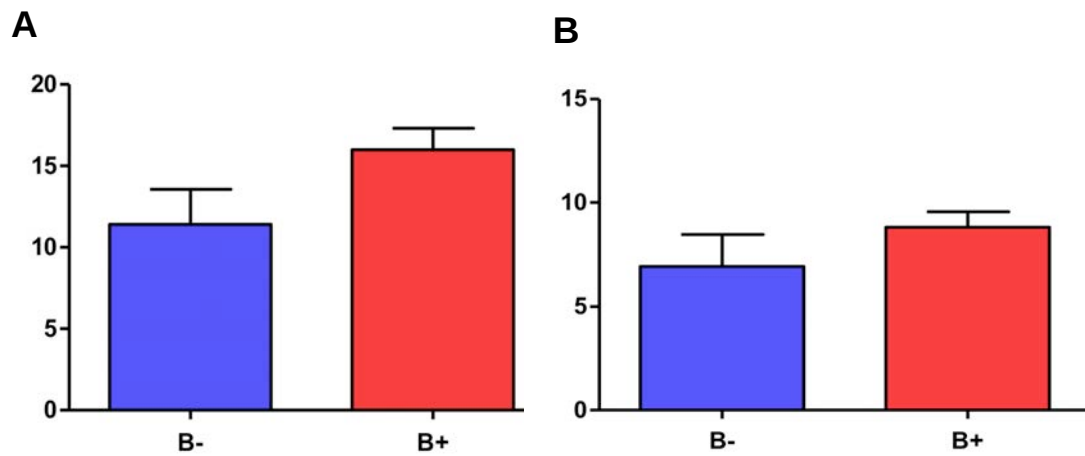


Figure 6. Expression level of hnRNP Q-like gene in exons 2 (A) and 10 (B) in B- and B+ genomes in muscle samples. N = 24.

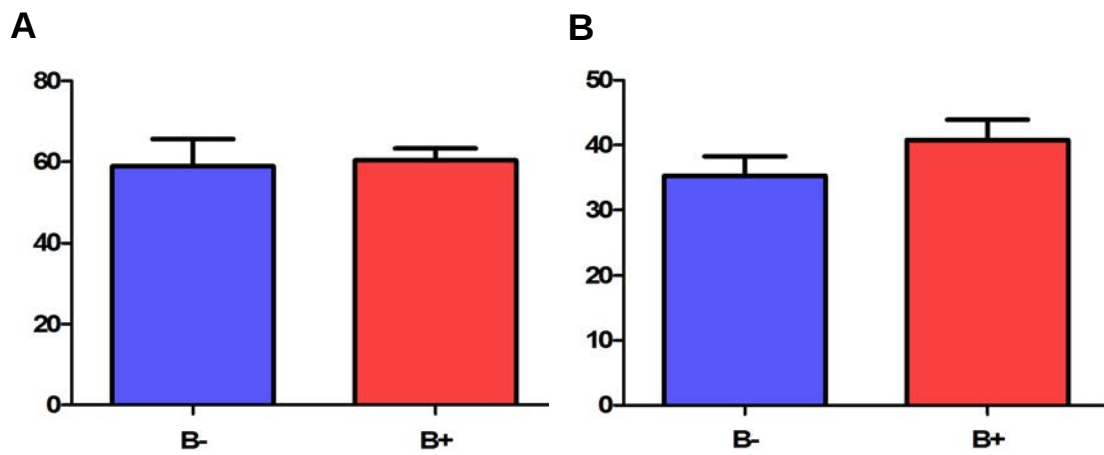


Figure 7. Expression level of hnRNP Q-like gene in exons 2 (A) and 10 (B) in B- and B+ genomes in brain samples. N = 24.

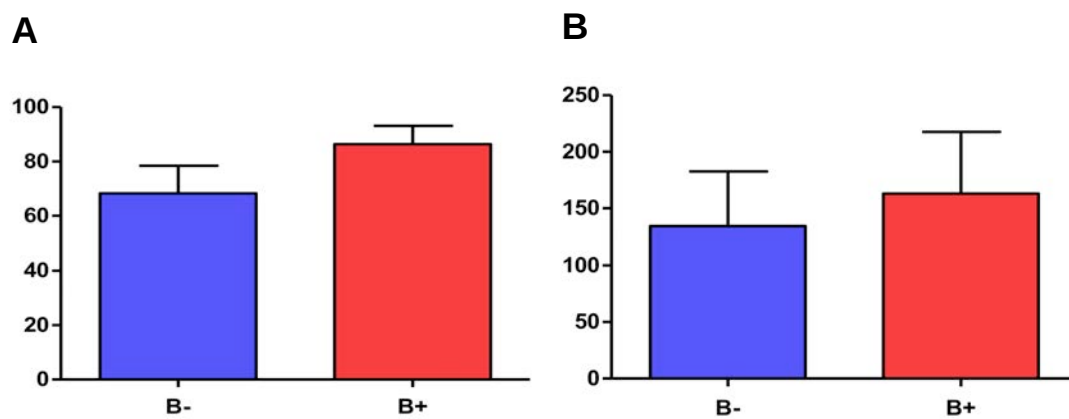


Figure 8. Expression level of hnRNP Q-like gene in exons 2 (A) and 10 (B) in B- and B+ genomes in heart samples. N = 23.

Discussion

Retroposed copies lack many of their parental genes genetic features, such as introns and regulatory elements (Pan and Zhang, 2009). Analysis of amplified fragments showed hnRNP Q-like gene does not have introns in B chromosome, supporting the hypothesis that the gene was retroinserted in this element. In an attempt to identify the retroinsertion mechanism a search for transposable elements was conducted. Only small regions similar to transcriptase reverse domain of TEs were found, which indicate sequences belong to TEs of class I. Class I represents retrotransposons, which use reverse transcription to duplicate and spread in the genome. The movement is mediated by RNA, which is transcribed from a genomic copy, then reverse-transcribed into DNA by a TE-encoded reverse transcriptase (RT) (Wicker et al. 2007). TEs of class I are separated into categories, those TEs found flanking hnRNP Q-like copies of the B chromosome represent two of them, LINEs (long interspersed elements) that contain the coding sequence of reverse transcriptase (Pan and Zhang, 2009) and retrotransposons which have long terminal repeats (LTR). These terminal repeats mark the insertion point into a host DNA. However most of TE sequences detected were degraded over the time, and are no longer clearly recognizable (Henikoff et al. 1997). Our analysis detected relics of transposable elements TR domains flanking the B chromosome copy of hnRNP Q-like gene, suggesting retroinsertion was the mechanism for the gene copy movement from the A to the B genome. Previous analysis of B⁺ and B⁻ whole genomes (Valente et al. 2014) identified a high quantity of fragmented genes in B chromosome. After migrating to B chromosome, the most part of genes seems to have entered in stage of degeneration. Ancient TE insertions have often suffered mutations and subsequent insertions and rearrangements both reduce similarity of the element to known families and interrupt their continuous genome activity (Wicker et al. 2007).

Analyses of bioinformatics showed copies in B⁺ genome have more mutations than that in B⁻ genome. Variant copies, from B chromosome, obtained from de novo assembly, presented high levels of saturation, due high ratio of mutation, and consequently, are not informative about their evolutionary process (Posada, 2003). The high level of variation detected in the B copies indicates the retroinsertion occurred a long time ago, probably during the origin of this chromosome.

Quantitative PCR of genomic DNA showed B⁺ genome has the double of copies of hnRNP Q-like gene compared to B⁻ genome, in the regions correspondents to exons 2 and

10, corroborating with previous bioinformatics analyses (Valente et al. 2014). Possibly, the gene hnRNP Q-like was retroinserted in B chromosome and then duplicated.

Expression analysis of hnRNP Q-like in B⁺ and B⁻ genomes based on qPCR was performed and no differential expression between the samples was detected. According to RetrogeneDB (<http://retrogenedb.amu.edu.pl>), the hnRNP Q-like gene has retrogene features in *Macaca mulatta*, *Homo sapiens* e *Rattus norvegicus* and is considered a pseudogene (Kabza et al. 2014). This can explain the absence of significant difference in gene expression, because the retrogene in B chromosome is no longer functional. Retroposed gene copies are often believed to be pseudogenes because they lack regulatory regions and, as a consequence, they will often degenerate (Bai et al. 2007).

The results show hnRNP Q-like gene was retroinserted in B chromosome of *A. latifasciata* probable as a consequence of TEs action. Furthermore, hnRNP Q-like gene was duplicated in B chromosome, but these copies are no longer functional, indicating it is a pseudogene. This is the first report of retroinsertion of genes contributing to the B chromosome genome content.

References

As referências deste artigo encontram-se reunidas no item “5. Referências bibliográficas”

5. Conclusões

As análises da sequência nucleotídica do gene hnRNP Q-like mostraram que o genoma B⁺ possui cópias canônicas idênticas às cópias do genoma B⁻, assim como diversas cópias variantes. Também foi verificado que o genoma B⁺ apresenta maior variabilidade nucleotídica quando comparado ao genoma B⁻, confirmando a presença das cópias variantes.

As sequências amplificadas por reação de PCR não apresentam íntrons quando presentes no cromossomo B, indicando que o hnRNP Q-like é de fato um retrogene. Análises das regiões flangeadoras do gene mostraram a existência de resquícios de elementos transponíveis, sugerindo que o gene foi reinserido no cromossomo B por mecanismos de retroinserção. A partir das reações de PCR quantitativa em DNA genômico, foi verificado

que o número de cópias relativo é o dobro nos genomas B⁺ em relação ao B⁻, entretanto essas sequências não são expressas.

A partir dos resultados do presente trabalho, conclui-se que o gene hnRNP Q-like foi retroinserido no cromossomo B de *Astatotilapia latifasciata* provavelmente como consequência da ação de elementos transponíveis, e pode ter sofrido duplicações posteriores no cromossomo B, tornando-se um pseudogene. Este é o primeiro trabalho sobre retroinserção de genes em cromossomos B.

6. Referências bibliográficas

- Bai Y, Casola C, Feschotte C, Betran E (2007). Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biology*. 8: R11.
- Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A (2013). Formation and expression of pseudogenes on the B chromosome of rye. *Plant Cell*. 25: 2536-44.
- Bel Y, Ferré J, Escrìche B (2011). Quantitative real-time PCR with SYBR Green detection to assess gene duplication in insects: study of gene dosage in *Drosophila melanogaster* (Diptera) and in *Ostrinia nubilalis* (Lepidoptera). *BMC Research. Notes*. 4:84.
- Beukeboom LW (1994). Bewildering Bs: an impression of the 1st B-Chromosome Conference. *Heredity*. 73: 328-336.
- Brawand D, Wagner CE, Li YI, Malinsky M, Keller I et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*. 513: 375–381.
- Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2009). Geneious v4.8.5 [Internet]. Available from <http://www.geneious.com>.
- Fantinatti BEA, Mazzuchelli J, Valente GT, Cabral-de-Mello DC, Martins C (2011). Genomic content and new insights on the origin of the B chromosome of the cichlid fish *Astatotilapia latifasciata*. *Genetica*. 139: 1273-1282.
- Fedoroff NV (2012). Transposable elements, epigenetics, and genome evolution. *Science*. 338: 758-767

- Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE et al. (2010). The Pfam protein families database. *Nucleic Acids Res* 38 (Database issue) :D211–D222.
- Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*. 278, 609.
- Houben A, Banaei-Moghaddam AM, Klemme S (2013). Biology and evolution of B chromosomes. In: Leitch IJ, editor. *Plant Genome Diversity*. Volume 2. Berlin: Springer. p. 149-165.
- Huang C-J, Lin W-Y, Chang C-M, Choo K-B (2009). Transcription of the rat testis-specific Rtdpoz-T1 and -T2 retrogenes during embryo development: co-transcription and frequent exonisation of transposable element sequences. *BMC Mol Biol*. 10(1):74.
- Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, et al. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep*. 1: 167–178.
- Jiménez MM, Romera F, González-Sánchez M, Puertas MJ (1997). Genetic control of the rate of transmission of rye B chromosomes. III. Male meiosis and gametogenesis. *Heredity*. 78: 636–644.
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110:462–467.
- Kabza M, Ciomborowska J, Makołowska I (2014). RetrogeneDB - a database of animal retrogenes. *Molecular Biology Evolution*. 31(7):1646–1648.
- Kim DY, Kwak E, Kim SH, Lee KH, Woo KC, et al. (2011). hnRNP Q mediates a phase-dependent translation-coupled mRNA decay of mouse Period3. *Nucleic Acids Res*. 39: 8901–8914.
- Kocher TD (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*. 5: 288- 298.
- Kuroiwa A, Terai Y, Kobayashi N, Yoshida K, Suzuki M, Nakanishi A, Matsuda Y, Watanabe M, Okada N (2014). Construction of chromosome markers from the Lake Victoria cichlid *Paralabidochromis chilotes* and their application to comparative mapping. *Cytogenet Genome Res*. 142:112–120.

- Langmead B, Salzberg S (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9:357-359.
- Lee K-H, Woo K-C, Kim D-Y, Kim T-D, Shin J, Park SM, Jang SK, Kim KT (2012). Rhythmic interaction between *Period1* mRNA and hnRNP Q leads to circadian time-dependent translation. *Mol Cell Biol*. 32: 717–728.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009). 1000 Genome project data processing subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*. 25: 2078-9.
- Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Šimková H, et al. (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci USA*. 109: 13343-13346.
- Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F and Marshall D (2010). Tablet – next generation sequence assembly visualization. *Bioinformatics*. 26(3): 401-402.
- Mourelatos Z, Abel L, Yong J, Kataoka N, Dreyfuss G (2001). SMN interacts with a novel family of hnRNP and spliceosomal proteins. *EMBO J*. 20 5443–5452.
- Muller PY, Janovjak H, Miserez AR, Dobbie Z (2002). Processing of gene expression data generated by quantitative real-time RT-PCR. *BioTechniques*. 32(6):1372-4, 1376, 1378-9.
- Pan D, Zhang L (2009). Burst of young retrogenes and independent retrogene formation in mammals. *PLoS ONE*. 4(3): e5040.
- Pendergast JS, Friday RC, Yamazaki S (2010). Distinct functions of *Period2* and *Period3* in the mouse circadian system revealed by *in vitro* analysis. *PLoS One*. 5(1). e8552.
- Poletto AB, Ferreira IA, Cabral-de-Mello DC, Nakajima RT, Mazzuchelli J, Ribeiro HB, Venere PC, Nirchio M, Kocher TD, Martins C (2010b). Chromosome differentiation patterns during cichlid fish evolution. *BMC Genetics*. 11: 50.
- Poletto AB, Ferreira IA, Martins C (2010a). The B chromosome of the cichlid fish *Haplochromis obliquens* harbors 18S rRNA genes. *BMC Genetics*. 11: 1.
- Posada D (2003). Selecting models of evolution. In: Salemi, M., Vandamme, A.M. *The phylogenetic handbook: a practical approach to dna and protein phylogeny*. Cambridge University Press, Cambridge. Pp. 256-282.

- Quaresma AJ, Bressan GC, Gava LM, Lanza DC, Ramos CH, Kobarg J (2009). Human hnRNP Q re-localizes to cytoplasmic granules upon PMA, thapsigargin, arsenite and heat-shock treatments. *Exp Cell Res*. 315: 968–980.
- Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003). Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neuroscience Letters*. 339(1): 62-66.
- Sambrook J, Russel DW (2001). *Molecular cloning. A laboratory manual*, 3rd ed. ColdSpring.
- Schulz MH, Zerbino DR, Vingron M, Birney E (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. 28(8):1086-92.
- Simon P (2003). Q-gene: processing quantitative real-time RT-PCR data. *Bioinformatics*. 19(11):1439-1440.
- Smith WmL, Chakrabarty P, Sparks JS (2008). Phylogeny, taxonomy, and evolution of neotropical cichlids (Teleostei: Cichlidae: Cichlinae). *Cladistics*. 24: 625–641.
- Steeves G, Hansen D, Lamboj A (2010). *Cichlids of Africa. Vol. I: Haplochromines*. Grandã Publishing. pp 23-25.
- Stothard P (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*. 28:1102-1104.
- Valente GT, Conte MA, Fantinatti BEA, Cabral-de-Mello DC, Carvalho RF, Vicari MR, Kocher TD, Martins C (2014). Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Mol Biol Evol*. 31:2061-72.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8(12):973-82.
- Xia X, Xie Z (2001). DAMBE: software package for data analysis in molecular biology and evolution. *J Heredity*. 92(4): 371-3.
- Xie F, Sun G, Stiller JW, Zhang B (2011). Genome-wide functional analysis of the cotton transcriptome by creating an integrated est database. *PLoS One*. 6(11):e26980.
- Yoshida K, Terai Y, Mizoiri S, Aibara M, Nishihara H, Watanabe M, Kuroiwa A, Hirai H,

- Hirai Y, Matsuda Y. et al. (2011). B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes. *PLoS Genetics*. 7: e1002203.
- Zerbino DR , Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*. 18: 821-829.
- Zhou Q, Zhu H, Huang Q, Xuan Z, Zhang G, Zhao L, Ding Y, Roy S, Vicoso B, Ruan J, Zhang Y, Zhao R, Mu B et al. (2012). Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics*. 22: 13-109.