

UNIVERSIDADE ESTADUAL PAULISTA

“Júlio de Mesquita Filho”

INSTITUTO DE BIOCÊNCIAS DE BOTUCATU

**CARACTERIZAÇÃO E GENÔMICA COMPARATIVA DE PEQUENOS RNAS
NUCLEARES EM PEIXES DA ORDEM CHARACIFORMES**

AMANDA BUENO DA SILVA

Botucatu

2025

UNIVERSIDADE ESTADUAL PAULISTA

“Júlio de Mesquita Filho”

INSTITUTO DE BIOCÊNCIAS DE BOTUCATU

AMANDA BUENO DA SILVA

**CARACTERIZAÇÃO E GENÔMICA COMPARATIVA DE PEQUENOS RNAS
NUCLEARES EM PEIXES DA ORDEM CHARACIFORMES**

Dissertação apresentada ao Programa de
Pós-graduação em Ciências Biológicas
(Genética) – Instituto de Biociências,
UNESP, *campus* de Botucatu, para a
obtenção do título de Mestre.

Orientador: Ricardo Utsunomia

Botucatu

2025

D229c da Silva, Amanda Bueno
Caracterização e genômica comparativa de pequenos
RNAs nucleares em peixes da ordem Characiformes /
Amanda Bueno da Silva. -- Botucatu, 2025
59 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista
(UNESP), Instituto de Biociências, Botucatu
Orientador: Ricardo Utsunomia

1. Sequência de RNA. 2. Genoma. 3. Elementos
transponíveis. 4. Evolução animal. I. Título.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE AMANDA BUENO DA SILVA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA), DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 30 dias do mês de setembro do ano de 2025, às 9h, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de AMANDA BUENO DA SILVA, intitulada **Caracterização e genômica comparativa de pequenos RNAs nucleares em peixes da ordem Characiformes**, sob orientação do Prof. Dr. Ricardo Utsunomia. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. RICARDO UTSUNOMIA (Orientador(a) - Participação Virtual) do(a) Faculdade de Ciências - Bauru / Unesp - Bauru, Pós-Doutorando DUILIO MAZZONI ZERBINATO DE ANDRADE SILVA (Participação Virtual) do(a) National Institute of Health, Pós-Doutorando FRANCISCO DE MENEZES CAVALCANTE SASSI (Participação Virtual) do(a) Centro de Ciências Biológicas e da Saúde / Southwest University (Chongqing - China). Após a exposição pela mestranda e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, a discente recebeu o conceito final: Aprovado . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. RICARDO UTSUNOMIA



Documento assinado digitalmente

RICARDO UTSUNOMIA

Data: 06/10/2025 16:11:35-0300

Verifique em <https://validar.itl.gov.br>

Agradecimentos

Meus sinceros agradecimentos à Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), pela formação acadêmica proporcionada, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Brasil), pelo suporte financeiro que viabilizou esta etapa.

Agradeço ao Prof. Dr. Ricardo Utsunomia e ao Prof. Dr. Fábio Porto Foresti por terem me acolhido desde a Iniciação Científica e possibilitado o início da minha trajetória na pesquisa. Obrigada pela orientação e por todos os ensinamentos ao longo desse percurso.

Aos colegas do Laboratório de Genômica e Conservação de Peixes, meu muito obrigada, especialmente a Henrique, Vitor, Zeni, Gabi, Nat, Hélio e Goes. Vocês foram essenciais na minha vida durante esses dois anos. Levarei comigo cada café, almoço, conversa e desabafo que compartilhamos nesse período.

Ao meu parceiro, Rafael, expresso minha profunda gratidão por todo o suporte e incentivo desde o início da minha trajetória em Bauru. Você nunca mediu esforços para me dar apoio e sua companhia foi essencial para que eu seguisse confiante.

À minha avó, a pessoa mais generosa que já conheci, dedico todo meu amor. Foi ela quem construiu uma família grande e unida, que sempre me fortaleceu e encorajou a perseguir meus sonhos. A cada familiar que, de alguma maneira, esteve presente nessa caminhada, deixo meu reconhecimento sincero.

Por fim, à minha mãe, Sandra, base da minha vida e força nos momentos desafiadores, entrego minha mais profunda gratidão. Seu amor, cuidado e confiança sustentaram cada passo dado até aqui. Nada seria possível sem você.

Resumo

Os pequenos RNAs nucleares (snRNAs) são RNAs não codificantes essenciais no *splicing* e em outros processos nucleares, incluindo maturação de RNAs ribossomais e processamento de RNAs transportadores. Esses genes, organizados em famílias multigênicas, podem evoluir por modelos de evolução em concerto ou *birth-and-death*, além de serem alvos de transposição via elementos móveis, mostrando alta plasticidade genômica. Embora a localização cromossômica dos snRNAs seja relativamente bem descrita em peixes, pouco se conhece sobre seus padrões estruturais e dinâmicas evolutivas em Characiformes, um dos grupos mais diversos de peixes neotropicais. Parte dessa lacuna se deve ao fato de que informações sobre os genes U estão comumente restritas às regiões codificantes e pouco se sabe sobre sua organização e evolução. Com isso, o objetivo inicial do estudo foi caracterizar os pequenos RNAs nucleares em peixes Characiformes utilizando dados de sequenciamento de *short-reads* pareados e *pipelines* de montagem baseados em NOVOPlasty e Bowtie2. As análises resultaram na identificação de dois padrões distintos de organização do snRNA U1. A classe U1-A, associada ao U6, foi encontrada em todas as espécies analisadas, enquanto a classe U1-B ocorreu em arranjo tandem com U1 em espécies das famílias Characidae e Crenuchidae. Os resultados evidenciaram sintonia recorrente entre os snRNAs U2 e U5, enquanto U4 ocorre isoladamente em todas as espécies analisadas. Elementos PSE estão presentes em todos os snRNAs e a TATA-box é restrita a U6. Os pequenos RNAs nucleares U1, U2, U4 e U5 estão ativamente transcritos, enquanto U6 apresenta menor nível de expressão. Todos os espaçadores intergênicos, com exceção do gene U4, são transcritos. Análises de genomas completos em *Astyanax mexicanus* e *Hoplias malabaricus* confirmaram os padrões mostrados pelos *short-reads*, com grandes *arrays* multigênicos de snRNAs frequentemente associados a domínios típicos de elementos transponíveis (TEs). Este estudo fornece a primeira caracterização abrangente da organização de snRNAs em Characiformes, evidenciando padrões específicos de associação entre genes e possíveis mecanismos de dispersão mediada por elementos transponíveis, contribuindo para a compreensão da dinâmica genômica e evolução de famílias multigênicas em peixes neotropicais.

Palavras-chave: sequência de RNA; genoma; elementos transponíveis; evolução animal.

Abstract

Small nuclear RNAs (snRNAs) are non-coding RNAs essential for splicing and other nuclear processes, including ribosomal RNA maturation and transfer RNA processing. These genes, organized in multigene families, may evolve through concerted evolution or birth-and-death models, in addition to being targets of transposition by mobile elements, which highlights their high genomic plasticity. Although the chromosomal location of snRNAs is relatively well described in fishes, much less is known about their structural patterns and evolutionary dynamics in Characiformes, one of the most diverse lineages of neotropical fishes. Part of this gap is due to the fact that most studies on U genes have focused solely on their coding regions, leaving their genomic organization and evolutionary history unexplored. Here, we characterized snRNAs in Characiformes using paired-end short-reads sequencing combined with assembly pipelines based on NOVOPlasty and Bowtie2. Our analyses revealed two distinct organizational patterns of U1: the U1-A class, consistently associated with U6 and found in all species, and the U1-B class, organized in tandem repeats and restricted to species of the families Characidae and Crenuchidae. We also detected a conserved syntenic arrangement of U2 and U5, while U4 was invariably found in isolation in all analyzed species. PSE elements are conserved across all snRNAs, whereas the TATA-box motif is unique to U6. Transcriptomic data confirmed active transcription of U1, U2, U4, U5 and U6, with U6 expressed at lower levels. Notably, all intergenic spacers, except those adjacent to U4, also transcribed. Whole-genome analyses of *Astyanax mexicanus* and *Hoplias malabaricus* supported the short-reads findings, uncovering large multigene arrays of snRNAs frequently associated with typical domains of transposable elements. Altogether, this study provides the first comprehensive overview of snRNA organization in Characiformes, uncovering lineage-specific genomic arrangements and highlighting the potential role of transposable elements in the dispersion of these genes.

Keywords: RNA sequence; genome; transposable elements; animal evolution.

Sumário

1. Introdução	8
1.1. Características gerais de pequenos RNAs nucleares	8
1.2. Evolução de pequenos RNAs nucleares	12
1.3. Ordem Characiformes e snRNAs em peixes	15
2. Objetivos	18
3. Material e Métodos	19
3.1. Dados de sequenciamento e montagem	19
3.2. Predição de estruturas secundárias	22
3.3. Estimativa de abundância e transcrição	22
3.4. Busca por <i>MOTIFs</i> conservados	22
3.5. Árvore de máxima verossimilhança	23
3.6. Análise de genomas completos	23
4. Resultados	24
4.1. Organização genômica do snRNA U1 em Characiformes: primeira descrição de duas classes distintas de U1 e associação consistente ao snRNA U6	25
4.2. Associação entre snRNAs U2 e U5: ocorrência frequente de sintenia entre os dois genes	28
4.3. Padrão genômico do snRNA U4: ausência de associação a outras famílias multigênicas	29
4.4. Pequenos RNAs nucleares no genoma de peixes Characiformes	30
4.5. Análise de genomas completos e associação de snRNAs a outros elementos repetitivos	36
5. Discussão	39
6. Referências	43
7. Material Suplementar	50

1. Introdução

1.1. Características gerais de pequenos RNAs nucleares

O genoma de organismos eucariotos varia em termos de tamanho e organização, devido principalmente à presença de elementos repetitivos, que representam mais da metade do DNA nuclear e podem ter funções biológicas, contribuindo para variações cariotípicas observadas em múltiplos grupos (Kidwell, 2002). Entre as sequências repetitivas, os elementos transponíveis (TEs) estão amplamente distribuídos nos genomas devido à sua capacidade de replicação e inserção em novas localizações. Esses elementos podem ser classificados em duas categorias principais: os retrotransposons (classe I), que utilizam RNA como intermediário no processo de “copiar e colar”, e os transposons de DNA (classe II), que atuam pelo mecanismo de “corte e colagem” (Wicker et al., 2007).

Sequências repetitivas também se organizam em tandem, como nos DNAs satélites e nas famílias multigênicas (Charlesworth et al., 1994; Garrido-Ramos, 2017). Os DNAs satélites (satDNAs) são longas cadeias de unidades repetitivas (monômeros), altamente similares entre si, localizadas principalmente em regiões heterocromáticas de centrômeros e extremidades cromossômicas (Garrido-Ramos, 2017). Já as famílias multigênicas são compostas por grupos de genes que surgiram por eventos de duplicação e apresentam similaridades estruturais e funcionais devido a uma origem comum (Nei & Rooney, 2005). Esse conjunto é representado por genes de histonas, DNAs ribossomais (rDNA) e pequenos RNAs nucleares (snRNAs U).

Os pequenos RNAs nucleares constituem um subconjunto de RNAs não codificantes localizados no nucleoplasma e estão classificados em duas categorias principais de acordo com os cofatores proteicos associados. A primeira categoria, chamada de classe Sm, abrange os snRNAs U1, U2, U4, U4atac, U5, U7, U11 e U12 e é marcada pela presença de um *cap* trimetilguanosina na extremidade 5', uma estrutura em *stem-loop* na extremidade 3' e um sítio de ligação específico para o conjunto de sete proteínas Sm que se organizam formando um anel. A segunda categoria, classe Lsm, está presente nos snRNAs U6 e U6atac e possui um *cap* monometilfosfato 5', uma estrutura em *stem-loop* 3' e uma região rica em uridinas que serve como sítio de ligação do anel formado por proteínas Lsm (Matera et al., 2007; Nilsen, 2003; Will & Lührmann, 2011).

Elementos regulatórios específicos de pequenos RNAs nucleares já foram descritos em detalhes para diversos vertebrados e são altamente conservados (Domitrovich & Kunkel, 2003; Hernandez, 2001). Por exemplo, o elemento de sequência proximal PSE (*proximal sequence element*) está localizado a aproximadamente 50–60 pares de bases antes do sítio de início da transcrição, sendo responsável pela determinação desse sítio (Hernandez, 2001; Jawdekar & Henry, 2008). Outro elemento regulatório importante é o elemento de sequência distal DSE (*distal sequence element*), localizado cerca de 220 pares de bases antes do sítio de início da transcrição, atuando como um *enhancer* e aumentando a eficiência transcricional (Hernandez, 2001; Jawdekar & Henry, 2008). Destaca-se que o snRNA U6 apresenta uma particularidade: além dos elementos PSE e DSE, sua região promotora contém um TATA-box altamente conservado, situado cerca de 30 pb a montante do ponto de início. A presença conjunta dos elementos PSE e TATA direciona o recrutamento do mecanismo específico da RNA polimerase III para o gene U6 e U6atac, enquanto a ausência do TATA-box especifica o recrutamento do aparato transcricional da RNA polimerase II (Hernandez, 2001; Jawdekar & Henry, 2008).

Os genes U participam de diversos processos celulares essenciais, incluindo o processamento de tRNAs, a maturação de rRNAs, a síntese de DNA telomérico e, sobretudo, são componentes fundamentais no *splicing* de pré-RNAs mensageiros, integrando o complexo do spliceossomo (Mattaj et al., 1993). O processo de *splicing* ocorre na maioria das linhagens eucarióticas, onde os íntrons são removidos dos mRNAs codificadores de proteínas pelo spliceossomo, um grande complexo ribonucleoproteico composto por cerca de 200 proteínas e cinco pequenos RNAs não codificadores (Nilsen, 2003). Nos eucariotos, existem dois tipos básicos de spliceossomos: i) spliceossomo dependente de U2 (spliceossomo maior), que catalisa íntrons do tipo U2, e ii) spliceossomo dependente de U12 (spliceossomo menor), menos abundante, que remove os raros íntrons do tipo U12 (Patel & Steitz, 2003). Várias proteínas e os snRNAs U1, U2, U4, U5 e U6 organizam o spliceossomo maior, enquanto o spliceossomo menor é formado pelos snRNAs U5 (comum aos dois tipos de spliceossomo), U11, U12, U4atac e U6atac, que possuem uma estrutura e função semelhantes aos snRNAs U1, U2, U4 e U6 (Patel & Steitz, 2003). O spliceossomo maior, formado pelos snRNAs U1, U2, U4, U5 e U6, é a variante predominante em metazoários, plantas e fungos (Marz et al., 2008; Valadkhan, 2005).

O cerne do spliceossomo é formado pela junção de cada molécula de snRNA U com pelo menos sete subunidades proteicas para formar uma snRNP (*small nuclear ribonucleoprotein*) (Will & Lührmann, 2011). Para que processo de *splicing* do pré-mRNA se inicie, o snRNP U1 reconhece e se liga ao sítio de *splicing* 5', enquanto o snRNP U2 se associa ao ponto de ramificação, promovendo o pareamento com a sequência adjacente, exceto pela adenosina conservada, que permanece desemparelhada devido ao seu papel catalítico crucial. A montagem do spliceossomo é finalizada com a entrada de um complexo tri-snRNP pré-formado, composto pelos snRNPs U4, U5 e U6. Em seguida, o spliceossomo sofre uma série de mudanças conformacionais que o torna cataliticamente ativo. Durante essa transição, os snRNPs U1 e U4 são dissociados do complexo. O snRNP U6 estabelece interação com o sítio de *splicing* 5', enquanto o snRNP U5 se alinha com as regiões exônicas adjacentes, promovendo a aproximação dos sítios de *splicing* 5' e 3' (Will & Lührmann, 2011) (Figura 1a).

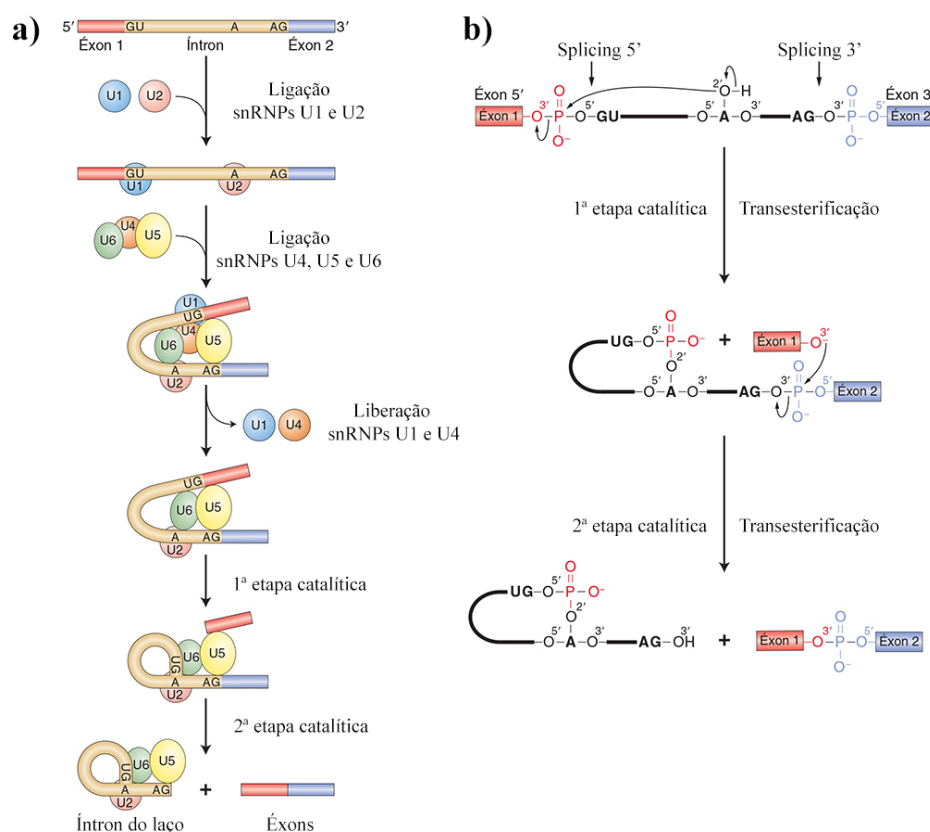


Figura 1. Mecanismo de *splicing*. **a)** Montagem e ativação do spliceossomo. **b)** Etapas catalíticas responsáveis pela remoção do íntron e ligação dos éxons. (Adaptado de Griffiths, Doebley e Peichel, 2022).

O *splicing* ocorre por meio de duas reações de transesterificação. Na primeira, o grupo 2'-hidroxila (2'-OH) da adenosina desemparelhada no ponto de ramificação realiza um ataque nucleofílico à ligação fosfodiéster no sítio de *splicing* 5', resultando na clivagem do pré-mRNA entre o éxon 5' e o íntron. Essa reação forma uma estrutura em laço (ou *lariat*), caracterizada por uma ligação 2'-5' entre a adenosina do ponto de ramificação e o início do íntron. Na segunda etapa, o grupo 3'-hidroxila (3'-OH) do éxon 5' realiza um novo ataque nucleofílico, agora sobre a ligação fosfodiéster no sítio de *splicing* 3'. Essa reação promove a união covalente entre os éxons 5' e 3', ao mesmo tempo em que libera o íntron em forma de laço. Após o término do *splicing*, os snRNPs U2, U5 e U6 são liberados do íntron excisado e podem ser reutilizados em novos ciclos de *splicing*, juntamente com os snRNPs U1 e U4, que haviam sido previamente removidos. Esse mecanismo é reiterado ao longo do pré-mRNA, garantindo a remoção precisa de todos os íntrons e a formação do transcrito maduro (Griffiths, Doebley e Peichel, 2022; Figura 1b)

A complexidade estrutural e funcional dos snRNAs e do spliceossomo reflete a importância central desses elementos para a regulação da expressão gênica nos eucariotos (Matera & Wang, 2014). Além de garantir a remoção precisa de íntrons e a formação de transcritos maduros funcionalmente competentes, o *splicing* também contribui para a geração de diversidade proteica por meio de eventos de *splicing* alternativo, permitindo que uma única unidade transcricional possa dar origem a múltiplas isoformas proteicas (Black, 2003).

1.2. Evolução de pequenos RNAs nucleares

Os pequenos RNAs nucleares (snRNAs) geralmente possuem menos de 300 nucleotídeos (Liu & Gall, 2007) e estão presentes em uma ampla diversidade de eucariotos. Eles apresentam elevado grau de conservação estrutural entre as espécies e compartilham diversas características comuns nas regiões flanqueadoras a montante do sítio 5' (Anjos et al., 2015). Apesar dos altos níveis de conservação de sequência entre espécies (Vierna et al., 2013), as unidades gênicas completas podem diferir quanto ao número de cópias, dinâmica cromossômica e organização genômica, muitas vezes com distribuição na forma de repetições em tandem (Anjos et al., 2015; Chairi & Gonzalez, 2015; Cross & Rebordinos, 2005; Liao, 1999; Manchado et al., 2006; Marzluff et al., 1983; Mattaj et al., 1983; Pelliccia et al., 2001). Parte dessa dinâmica cromossômica distinta pode ser explicada pelos dois principais modelos propostos para a evolução das famílias multigênicas U: o modelo de evolução em concerto e o modelo de *birth-and-death* (Nei & Rooney, 2005).

O modelo de evolução em concerto propõe que há uma homogeneização entre as cópias por meio de *crossing-over* desigual e conversão gênica, mecanismos centrais do processo de *molecular drive*, que promove a propagação e, eventualmente, a fixação de mutações em toda a família gênica (Eickbush & Eickbush, 2007; Walsh, 1987). Apesar de explicar determinados padrões observados em rDNAs, análises filogenéticas de famílias multigênicas revelaram inconsistências com o modelo de evolução em concerto, como agrupamentos intragenômicos incomuns e presença de pseudogenes (Rooney & Ward, 2005). O modelo de *birth-and-death* é proposto neste contexto e sugere que novas cópias surgem por duplicação, mas muitas são eliminadas ou convertidas em pseudogenes, enquanto outras poucas podem adquirir novas funções (Eirín-López et al., 2012).

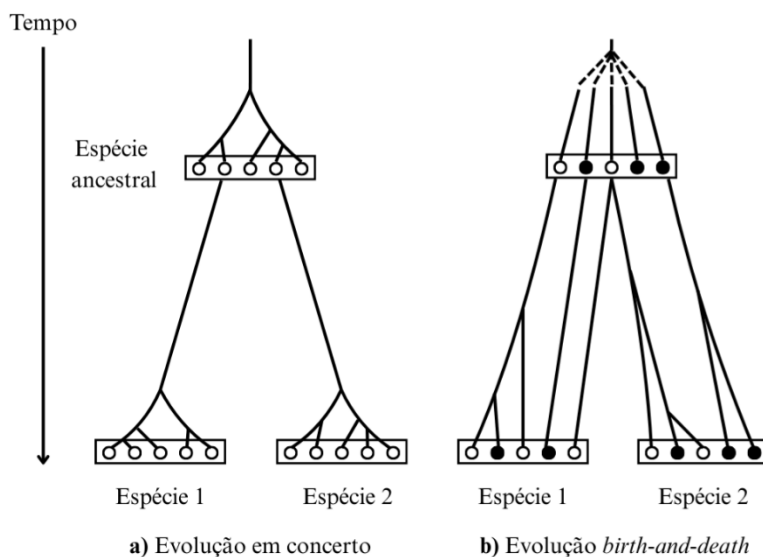


Figura 2. Diferentes modelos de evolução de famílias multigênicas. Círculos abertos representam genes funcionais e círculos fechados representam pseudogenes (Nei & Rooney, 2005).

Atualmente sabe-se que certas famílias multigênicas evoluem por meio de um modelo misto que integra aspectos da evolução em concerto e da evolução *birth-and-death* (Nei & Rooney, 2005). Nessa dinâmica, determinadas cópias podem se homogeneizar ao longo do tempo enquanto outras sofrem alterações tanto em número quanto em função, refletindo a complexidade desse processo evolutivo (Eirín-López et al., 2012; Merlo et al., 2012; Mount et al., 2007; Nei & Rooney, 2005; Pinhal et al., 2011).

Além de apresentarem uma dinâmica evolutiva complexa, relatos de invasões e atividade de elementos transponíveis em famílias multigênicas não são raros (Doucet et al., 2015; Glugoski et al., 2020; Malimpensa et al., 2020). Diferentes classes de retropseudogenes de snRNAs foram identificadas em vários genomas de mamíferos e sugerem que os caminhos de retrotransposição podem ser múltiplos e específicos para cada ambiente genômico, especialmente quando envolvem elementos nucleares longos e dispersos (*Long Interspersed Nuclear Element* - LINE). (Doucet et al., 2015). Além disso, Kojima (2015) descreveu uma nova classe de elementos nucleares curtos dispersos (*Short Interspersed Nuclear Elements* - SINE), o SINEU, cujas regiões 5' se originam dos snRNAs U1 e/ou U2 em genomas de crocodilianos. No mesmo sentido, elementos transponíveis da família Dada têm como principal característica a alta especificidade de

inserção em genes de RNA multicópia, especialmente tRNAs e snRNAs (Kojima, 2015). Os TEs que pertencem à essa família são capazes de reconhecer diretamente sequências-alvo conservadas, inserindo-se em locais específicos dentro dos genes, gerando duplicações do sítio-alvo (TSDs) (Kojima & Jurka, 2013). Outros elementos pertencentes à ordem de retrotransposons LINE, como R2, Mutsu, Keno e Dewa, também se inserem especificamente em genes de RNA, evidenciando um padrão de colonização de regiões altamente repetitivas e conservadas (Kojima & Jurka, 2013).

A evolução dos snRNAs em metazoários demonstra que os *loci* de U1 apresentam um comportamento dinâmico e móvel, caracterizado pela presença recorrente de pseudogenes, baixa conservação sintênica e ampla dispersão ao longo do genoma (Marz et al., 2008). Em contrapartida, análises conduzidas em *Oreochromis niloticus* demonstraram que, embora os genes U1 não atuem como TEs, eles são frequentemente encontrados flanqueados por esses elementos, os quais possivelmente contribuem para sua distribuição genômica (Cabral-De-Mello et al., 2012). A associação entre snRNAs e TEs — seja por mobilidade direta ou por atuarem como veículos de dispersão — reforça a hipótese de que a presença de elementos móveis nas proximidades dos *loci* U podem desempenhar um papel relevante no padrão de distribuição desses genes pelo genoma (Pucci et al., 2018; Yano et al., 2020).

1.3. Ordem Characiformes e snRNAs em peixes

A região Neotropical, que compreende os sistemas hidrográficos da América do Sul e América Central, representa uma das áreas biogeográficas mais ricas do planeta, abrigando a maior diversidade de peixes de água doce conhecida (Reis et al., 2016). Essa diversidade é dominada pela superordem Ostariophysi, que possui cerca de 95 famílias, 1.500 gêneros e 11.000 espécies, representando cerca de 70% das espécies de água doce. O grupo dos Ostariophysi inclui quatro ordens principais: Cypriniformes, Characiformes, Siluriformes e Gymnotiformes, com dados moleculares que apontam para a existência de um último ancestral comum há aproximadamente 150 milhões de anos atrás (Dai et al., 2018; Melo et al., 2022, 2024; Near et al., 2012). Estudos recentes sugerem parafília dos Characiformes, com os Siluriformes como grupo irmão de Characoidei, ambos formando um clado irmão de Citharinoidei, embora ainda não exista um consenso sobre as relações filogenéticas recalcitrantes nestes grupos (Arcila et al., 2017; Nakatani et al., 2011; Wheaton Dimmick & Larson, 1996).

No que se refere a número de espécies, a ordem Characiformes se destaca como um dos grupos de peixes mais ricos (Arcila et al., 2017), com 24 famílias e mais de 2.100 espécies reconhecidas, sendo que este número continua a crescer com novos estudos taxonômicos e descobertas contínuas. São peixes caracterizados por uma notável diversidade morfológica e ecológica, variando desde pequenos planctívoros e detritívoros até grandes predadores piscívoros (Reis et al., 2016). Além disso, a diversidade citogenética e genômica de Characiformes é amplamente documentada, incluindo fenômenos como polimorfismos cromossômicos, ocorrência de poliploidia, presença de cromossomos B e sistemas sexuais variados (De Souza E Sousa et al., 2017; Gazoni et al., 2021; Ponzio et al., 2018; Ricardo Vicari et al., 2006).

Os *clusters* repetidos em tandem são predominantemente organizados de forma individualizada para cada família gênica, embora haja ocorrência esporádica de *clusters* envolvendo mais de uma família, como observado no peixe europeu *Merluccius merluccius* (ordem Gadiformes), no qual os snRNAs U1 e U2 estão co-localizados (García-Souto et al., 2015). Nessa espécie, os snRNAs U1 estão amplamente distribuídos em quase todos os pares cromossômicos, refletindo alta plasticidade genômica, provavelmente impulsionada por transposição e duplicação gênica. Por outro lado, os snRNAs U2 formam um único *cluster* que coincide com o maior sítio de U1, sugerindo possível interação estrutural ou funcional entre essas famílias gênicas (García-Souto et

al., 2015). De forma semelhante, em espécies do gênero *Engraulis* (Clupeiformes), foram documentadas associações entre os snRNAs U2 e U5 (Chairi & Gonzalez, 2015). Em *Solea senegalensis*, detectou-se um *cluster* contendo o gene 5S rRNA ligado simultaneamente aos snRNAs U1, U2 e U5. Este arranjo representa a primeira evidência de ligação sintênica de múltiplos snRNAs a um gene 5S rRNA em peixes, indicando que eventos de duplicação, transposição e reorganização genômica podem favorecer a formação de *clusters* multigênicos, como observado em outros teleósteos (Manchado et al., 2006).

Na ordem Characiformes, os snRNAs U2 estão consistentemente localizados em dois pares de autossomos no gênero *Triportheus*. Entretanto, há variação quanto à localização de um desses *clusters*, que pode aparecer no par nº 4 ou em sintenia com *loci* de rDNA no par nº 3. Essa variação indica que, embora os *loci* de snRNA U2 sejam geralmente conservados, eles também podem sofrer rearranjos evolutivos, potencialmente influenciando seu contexto genômico e funcional (Yano et al., 2017). No mesmo gênero, dados de FISH revelaram que os genes U1 apresentam co-localização estável com *loci* de 5S no nono par cromossômico, característica considerada plesiomórfica (Yano et al., 2020). A ligação U1-5S também foi observada em espécies do gênero *Psalidodon*, indicando um possível padrão ancestral estabelecido na ordem (Silva et al., 2015). Em *A. jordani*, as unidades estão totalmente colocalizadas, enquanto em *P. paranae*, *P. bockmanni* e *P. fasciatus* os dois genes ocupam regiões diferentes do mesmo cromossomo. A ausência dessa ligação em outros grupos de Characiformes (Ferreira et al., 2007; Martins et al., 2006) sugere que a relação entre U1 e rDNA 5S pode ter seguido trajetórias evolutivas distintas, refletindo fusões, inversões, transposições e outros eventos ao longo da diversificação do grupo (Silva et al., 2015).

Diferentemente de espécies que apresentam U2 e U5 (Chairi & Gonzalez, 2015) ou U1 e U2 (Merlo et al., 2012), em *Astyanax* os snRNAs U2 são encontrados em um ou dois pares cromossômicos, sem associação com outras famílias multigênicas (Silva et al., 2015). Peixes do gênero *Characidium* apresentam um padrão ainda mais conservado para os snRNAs U2, restrito à região pericentromérica do primeiro par metacêntrico (Pucci et al., 2018). Estudos de sequência demonstraram, entretanto, que esses genes sofreram invasão por retrotransposons do tipo Tx1, indicando que, embora a posição cromossômica seja estável, a microestrutura interna dessas sequências pode ser remodelada por elementos móveis. Esses achados reforçam que, mesmo diante de um padrão conservado

de localização cromossômica, as famílias multigênicas de *Characidium* estão sujeitas a pressões evolutivas mediadas por TEs, capazes de promover variação microestrutural sem alterar significativamente a organização cariotípica (Pucci et al., 2018).

O conhecimento atual sobre os genes de snRNA em peixes baseia-se, em sua maior parte, em dados cromossômicos obtidos por FISH com sondas das regiões codificantes. Esses estudos indicaram que: i) sua organização genômica pode apresentar arranjos em tandem, dispersos ou mistos; ii) quando organizados em repetições em tandem, os *clusters* são homogeneizados de maneira semelhante aos sítios ribossomais, por meio de processos de evolução em concerto; iii) os *clusters* podem estar associados a outras sequências repetitivas, como genes ribossomais, outros genes de snRNA ou elementos transponíveis; e iv) a distribuição cromossômica é frequentemente preservada entre espécies próximas, embora eventos de dispersão possam resultar em significativa heterogeneidade entre espécies do mesmo gênero, e até mesmo entre populações da mesma espécie.

Investigações citogenéticas têm proporcionado avanços relevantes na compreensão da diversidade dos snRNAs e de seus padrões de distribuição genômica em eucariotos, embora o caminho evolutivo desses genes ainda não esteja completamente elucidado. Nesse sentido, a análise de unidades completas de snRNA, incluindo regiões codificantes e espaçadores intergênicos, utilizando *short-reads* de baixa cobertura, apresenta-se como uma alternativa promissora para investigar a origem, duplicação e diversificação dessas sequências em organismos não modelo. Nesse contexto, nosso objetivo foi caracterizar e comparar as unidades gênicas completas de snRNA U1, U2, U5 e U6 em espécies selecionadas de Characiformes, visando compreender melhor sua organização genômica e padrões evolutivos.

2. Objetivos

Caracterizar as estruturas e a dinâmica evolutiva de pequenos RNAs nucleares em espécies de peixes Characiformes utilizando abordagens genômicas.

- a) Desenvolver uma pipeline de caracterização para genes repetidos em tandem;
- b) Caracterizar pequenos RNAs nucleares completos (gene + espaçador) em 22 espécies de peixes Characiformes;
- c) Predizer estruturas secundárias e estabilidade termodinâmica de cada gene;
- d) Estimar abundância, divergência e níveis de transcrição de snRNAs caracterizados;
- e) Identificar *motifs* conservados associados a elementos regulatórios;
- f) Analisar genomas completos para validar a organização dos *clusters* multigênicos identificados através de *short-reads*.

3. Material e Métodos

3.1. Dados de sequenciamento e montagem

Foram utilizadas bibliotecas de sequenciamento massivo de abordagens genômicas de segunda geração de *short-reads* pareados em 22 espécies de peixes Characiformes (Tabela 1). Diferentes sequências de referência iniciais recuperadas do GenBank do NCBI foram utilizadas: U1 (*Psalidodon paranae*, KP411800.1), U2 (*Engraulis japonicus*, KP067950.1), U4 (*Maylandia zebra*, XR002721320), U5 (*Engraulis japonicus*, KP067950.1) e U6 (*Xenopus tropicalis*, NR033272.1). Para montar unidades completas dos genes de snRNAs U selecionamos pares de leituras com homologia aos diferentes genes U em cada biblioteca de DNA genômico separadamente através do *software* NOVOPlasty (Dierckxsens et al., 2017). As análises foram realizadas utilizando o servidor remoto localizado no Departamento de Biologia Estrutural e Funcional do Instituto de Biociências de Botucatu, sob a responsabilidade do prof. Dr. Claudio de Oliveira.

Tabela 1. Lista das espécies e famílias taxonômicas cujas bibliotecas genômicas foram utilizadas na caracterização de genes U. Estão indicadas as abordagens de sequenciamento e os números de acesso das bibliotecas.

Espécie	Família	Sequenciamento	Acesso
<i>Astyanax lacustris</i>	Characidae	HiSeq 2000 (2x150)	Biblioteca própria
<i>Astyanax mexicanus</i>	Characidae	HiSeq X Ten (2x150)	SRR6386652
<i>Bario sanctaefilomenae</i>	Characidae	HiSeq 2500 (2x100)	SRR5839692
<i>Brycon orbignyanus</i>	Bryconidae	HiSeq 2000 (2x90)	SRR6243207
<i>Characidium gomesi</i>	Crenuchidae	HiSeq 2500 (2x100)	SRR11679749
<i>Characidium zebra</i>	Crenuchidae	BGISEQ-500 (2x100)	SRR31478757
<i>Colossoma macropomum</i>	Serrasalminidae	BGISEQ-500 (2x100)	SRR22219024
<i>Cyphocharax modestus</i>	Curimatidae	BGISEQ-500 (2x150)	SRR26178311
<i>Erythrinus erythrinus</i>	Erythrinidae	HiSeq 1000 (2x150)	SRR10079924
<i>Hoplias intermedius</i>	Erythrinidae	HiSeq 1000 (2x150)	SRR10079921
<i>Hoplias malabaricus</i> KarD	Erythrinidae	HiSeq X Ten (2x150)	SRR11676684
<i>Megaleporinus elongatus</i>	Anostomidae	HiSeq 2000 (2x100)	SRR25308063
<i>Megaleporinus macrocephalus</i>	Anostomidae	MiSeq (2x250)	SRR7263033
<i>Piaractus mesopotamicus</i>	Serrasalminidae	MiSeq (2x100)	SRR11676685
<i>Prochilodus lineatus</i>	Prochilodontidae	BGISEQ-500 (2x100)	SRR11676686
<i>Prochilodus reticulatus</i>	Prochilodontidae	NovaSeq 6000 (2x150)	SRR31369841

<i>Psalidodon bockmanni</i>	Characidae	BGISEQ-500 (2x100)	Biblioteca própria
<i>Psalidodon paranae</i>	Characidae	HiSeq 2000 (2x150)	SRR5461470
<i>Psalidodon schubarti</i>	Characidae	BGISEQ-500 (2x100)	Biblioteca própria
<i>Pyrrhulina marylinae</i>	Lebiasinidae	BGISEQ-500 (2x150)	SRR25467276
<i>Pyrrhulina semifasciata</i>	Lebiasinidae	BGISEQ-500 (2x150)	SRR25476501
<i>Triportheus auritus</i>	Triportheidae	BGISEQ-500 (2x150)	SRR16764826

Usando o NOVOPlasty, o gene de referência inicial foi mapeado nas bibliotecas pareadas da espécie de interesse, utilizando os parâmetros: *k-mer* = 33; *insert size auto* = *yes*; *use quality scores* = *no*; *reduce ambiguous N's* = *yes*. O *software* organizou os *reads* similares ao gene-alvo em tabelas *hash*, permitindo a identificação de regiões sobrepostas e a extensão das sequências com base nos *reads* correspondentes detectados. Esse processo pode resultar em uma sequência consenso incompleta ou na montagem final do gene com seu espaçador. Nos casos em que a sequência permaneceu incompleta, foi realizada uma nova rodada de montagem, utilizando o *contig* incompleto da própria espécie como referência. Essa iteração foi repetida quantas vezes fosse necessário até que a sequência consenso estivesse completa, ou seja, apresentasse sinais de ser uma sequência repetida em tandem, seja pela repetição do gene dentro do *contig*, seja pela visualização em um *dot plot*. Após a obtenção de uma sequência consenso completa, os *reads* foram mapeados contra essa nova sequência consenso utilizando o Bowtie2 v2.5.1 (Langmead & Salzberg, 2012) com o objetivo de remover alinhamentos correspondentes ao *cluster* já obtido das bibliotecas da espécie, utilizando o SAMtools (Danecek et al., 2021). Em seguida, foi realizada uma nova rodada de montagem guiada pelo gene de referência, que poderia resultar em dois possíveis cenários: **a)** identificação de novas sequências consenso do gene de interesse, agora associadas a outros espaçadores e/ou genes; ou **b)** ausência de novas sequências consenso, indicando que não foram encontradas unidades repetidas em tandem adicionais. Como resultado, esse processo possibilitou a recuperação de unidades completas de todos os snRNAs, incluindo tanto os genes quanto seus espaçadores.

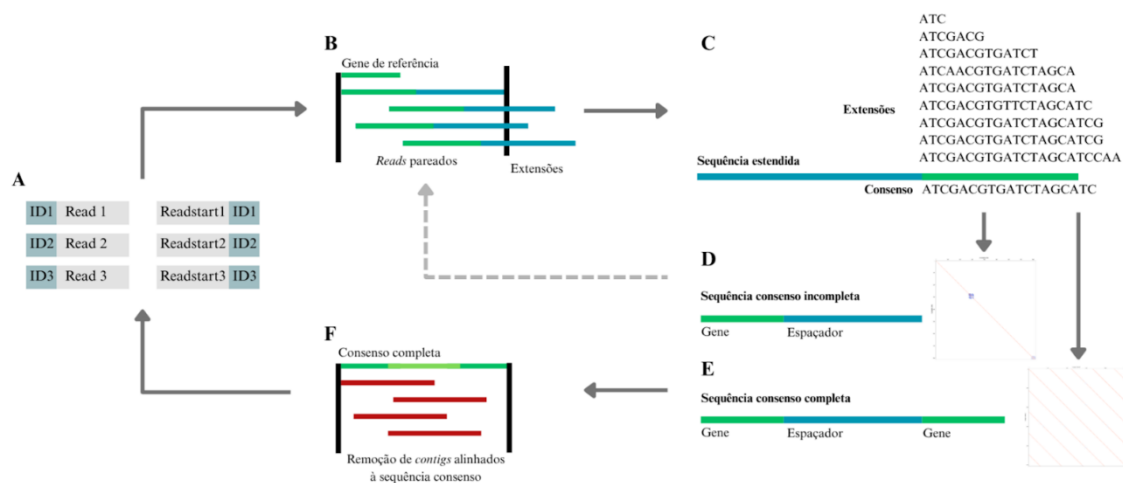


Figura 3. Fluxo de trabalho com o *software* NOVOPlasty associado ao Bowtie2. A) Os *reads* que apresentam similaridade com o gene de referência são organizados em tabelas *hash* que facilitam a identificação de regiões sobrepostas entre os *reads*. B) A biblioteca da espécie é mapeada e as regiões que correspondem ao início dos *reads* indicam os pontos de extensão da sequência. As posições dos *reads* pareados (*paired-end*) são verificadas, e cada *read* é alinhado à região montada previamente. C) A partir das extensões, é gerada uma sequência consenso representando a montagem preliminar do *contig*. D) Caso essa sequência ainda não esteja completa, novas rodadas de montagem (B) são realizadas com próprio *contig* incompleto da espécie sendo usado como referência, permitindo a extensão progressiva até a obtenção de uma montagem completa. E) Quando a montagem atinge a completude, a nova sequência consenso apresenta múltiplas repetições do gene e de seu espaçador, conforme esperado para repetições em tandem. F) A sequência consenso finalizada é mapeada nas bibliotecas e os *contigs* semelhantes a ela são removidos por meio do Bowtie2. O fluxo retorna à etapa A, onde novos *reads* semelhantes ao gene de referência são armazenados em tabelas *hash*. A partir desse ponto, o NOVOPlasty passa a buscar o gene de referência em uma biblioteca filtrada, livre dos *reads* já mapeados nas etapas anteriores, possibilitando a identificação de novas classes de famílias multigênicas agrupadas em tandem.

3.2. Predição de estruturas secundárias

As estruturas secundárias dos genes U foram determinadas após o alinhamento das sequências consenso obtidas para as espécies utilizando o MUSCLE (Edgar, 2004) implementado no Geneious (Kearse et al., 2012). A partir do alinhamento, a ferramenta RNAalifold (Bernhart et al., 2008) foi utilizada para modelar as estruturas secundárias e calcular a energia livre de Gibbs (ΔG) de cada gene.

3.3. Estimativa de abundância e transcrição

Parâmetros de abundância e divergência dos *reads* ao longo das unidades de snRNAs foram investigados com o objetivo de identificar a dispersão isolada das sequências ao longo do genoma. Para isso, foram utilizadas sequências de genes de cópia única como referência de normalização para ajustar os dados de abundância de elementos repetitivos como os *clusters* de snRNAs, expressando-os de maneira relativa em relação à quantidade de sequências esperada para um gene de cópia única. Em paralelo, as sequências brutas pareadas foram submetidas à filtragem de qualidade utilizando o *software* Trimmomatic v0.39 com os seguintes parâmetros: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10; LEADING:3; TRAILING:3; SLIDINGWINDOW:4:15. Após a filtragem foi realizada uma subamostragem aleatória de 5.000.000 leituras de cada arquivo filtrado (pares forward e reverse) utilizando o programa seqtk v1.3, com a opção -s100 para definir a aleatoriedade e reprodutibilidade. Finalmente, 10.000.000 de leituras *paired-end* de cada espécie foram selecionadas e mapeadas contra a unidade completa da própria espécie (gene + espaçadores), utilizando o Bowtie2 v2.5.1 (Langmead & Salzberg, 2012). O mesmo processo de mapeamento foi realizado utilizando bibliotecas de RNA-seq com depleção de RNA ribossomal da espécie *Psalidodon paranae*, a fim de inferir a transcrição nas unidades U recuperadas. Gráficos de cobertura foram gerados com base na contagem de bases mapeadas por posição, normalizada de acordo com o número total de nucleotídeos analisados na etapa de mapeamento.

3.4. Busca por MOTIFs conservados

A busca por *motifs* nos genes U foi conduzida através do *software* MEME (*Multiple Em for Motif Elicitation*) (Bailey et al., 2015), com o objetivo de identificar elementos conservados entre os diversos *clusters* já caracterizados. Foram utilizados segmentos de

até 150 pb a montante das sequências de snRNAs obtidas, com número de *motifs* por sequência = 0 ou 1; número máximo de *motifs* = 10; largura dos *motifs* = 5–25; número de sites em cada *motif* = 5–20. Após a identificação dos *motifs*, a busca por elementos conservados foi realizada nas sequências de snRNAs.

3.5. Árvore de máxima verossimilhança

Para compreender o processo evolutivo que levou à formação de duas classes bem definidas de U1 foi realizado um alinhamento das sequências U1-A e U1-B utilizando o MUSCLE (Edgar, 2004), seguido da reconstrução filogenética por meio de uma árvore de máxima verossimilhança. A árvore foi gerada utilizando o servidor web IQ-TREE (Trifinopoulos et al., 2016) e o modelo de substituição nucleotídica foi selecionado utilizando o ModelFinder (Kalyaanamoorthy et al., 2017), de acordo com o Critério de Informação Bayesiano (BIC), que identificou o modelo K2P+G4 como o de melhor ajuste. O suporte dos ramos foi calculado utilizando 1.000 réplicas de *bootstrap* (UFBoot) (Thi Hoang et al., 2017).

3.6. Análise de genomas completos

A identificação e extração de monômeros válidos no genoma completo de *Astyanax mexicanus* (GCF_023375975.1) e *Hoplias malabaricus* (GCA_029633875.2) foram realizadas através de um pipeline customizado utilizando a ferramenta BLAST+. O processo inicia com a entrada da biblioteca genômica e da sequência de referência. A biblioteca é preparada como um banco de dados para o BLAST e uma busca BLASTn é realizada, comparando a sequência de referência com a biblioteca genômica enquanto se aplicam critérios de filtragem como um E-value de 1×10^{-9} e uma cobertura mínima de 70%. Se o BLAST não encontrar correspondências, a execução é interrompida. Para os alinhamentos válidos, monômeros do mesmo cromossomo, com uma distância de até mil pares de bases, foram tratados como um único *array*. Essas informações foram então utilizadas para a extração final das sequências correspondentes, que foram analisadas utilizando o Geneious v7.1.9 (Kearse et al., 2012), com os genes de referência sendo utilizados como anotações.

4. Resultados

Após sucessivas etapas de montagem, foram recuperadas unidades completas dos snRNAs U1, U2, U4, U5 e U6 a partir de *short reads* das 22 espécies analisadas. Os *contigs* obtidos revelaram dois padrões distintos de distribuição para o snRNA U1, sendo um deles associado ao snRNA U6. Além disso, foi identificada uma associação recorrente entre os snRNAs U2 e U5, enquanto o snRNA U4 foi consistentemente detectado de forma isolada em todas as espécies.

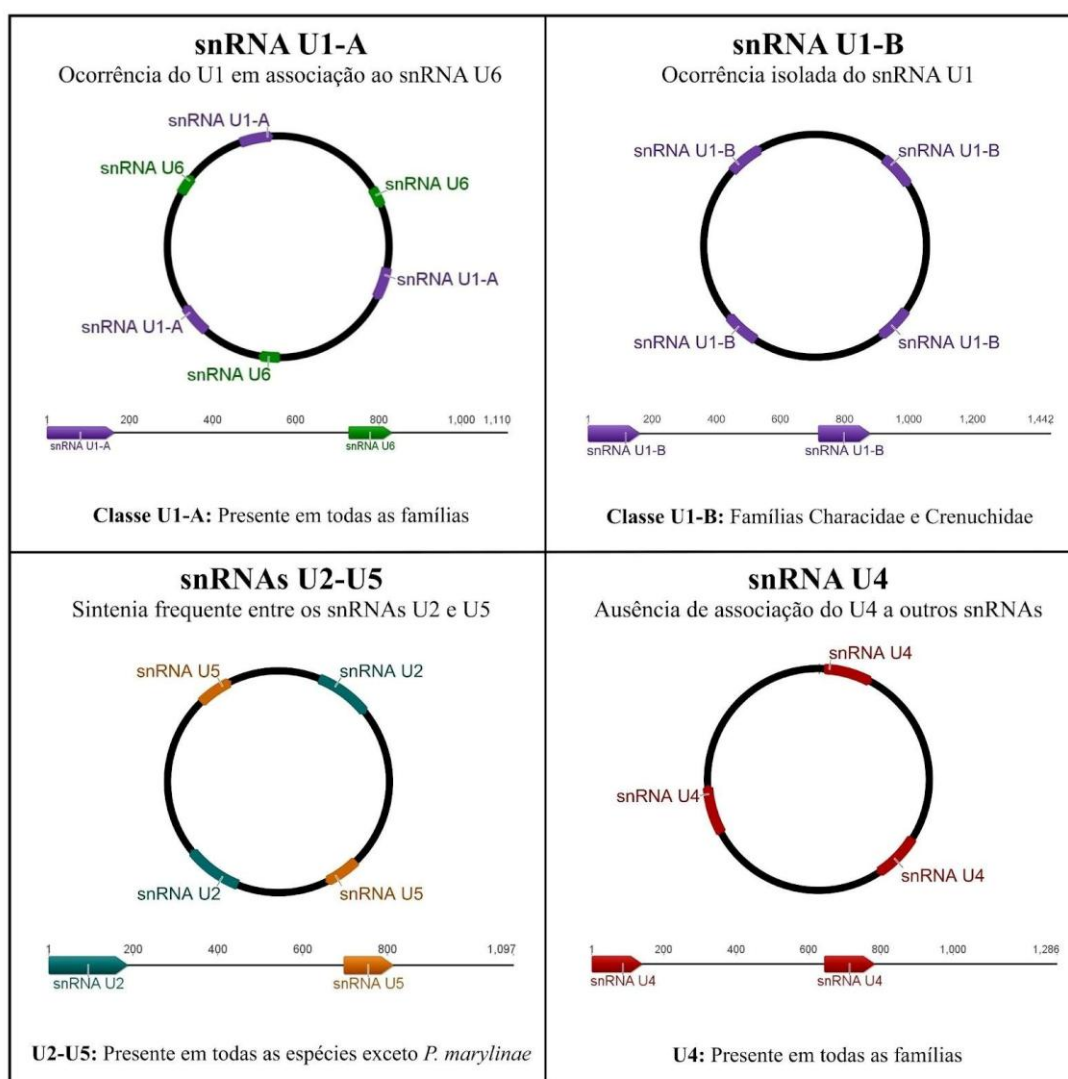


Figura 4. Padrões distintos de organização genômica dos pequenos RNAs nucleares identificados em espécies das famílias Characidae, Bryconidae, Crenuchidae, Serrasalminidae, Curimatidae, Erythrinidae, Anostomidae, Prochilodontidae, Lebiasinidae e Triportheidae.

4.1. Organização genômica do snRNA U1 em Characiformes: primeira descrição de duas classes distintas de U1 e associação consistente ao snRNA U6

Seis das 22 espécies apresentaram duas unidades independentes de U1 organizadas em tandem: uma unidade associada ao snRNA U6, denominada U1-A, e outra composta pelo gene U1 e um espaçador intergênico, denominada U1-B. A presença simultânea das duas unidades de U1 foi detectada apenas em peixes das famílias Characidae e Crenuchidae (Figura Suplementar 2), enquanto as demais famílias analisadas apresentaram exclusivamente a classe U1-A (Figura Suplementar 1). Para U1-A de *Triportheus auritus* e U1-B de *B. sanctaefilomenae* e *C. zebra* a região gênica do U1 apresentou menos de 60% de similaridade com o gene de referência. Em razão desse baixo percentual de similaridade, os *arrays* correspondentes foram excluídos das análises subsequentes. A ocorrência do snRNA U1 em associação ao DNA ribossomal 5S na espécie *Triportheus auritus* previamente descrita por Yano et al. (2020) também foi caracterizada no presente estudo.

As regiões codificadoras do snRNA U1 recuperadas apresentaram comprimento conservado de 164 nucleotídeos (nt) na maioria das espécies analisadas e, embora os espaçadores intergênicos correspondentes aos snRNAs U1-A e U1-B tenham apresentado tamanhos semelhantes dentro da mesma espécie (Tabela 2), suas sequências divergiram significativamente (Figura Suplementar 2). Por outro lado, observamos alta similaridade ao comparar os espaçadores U1-A ou U1-B separadamente entre espécies de uma mesma família. Os valores de similaridade ao gene de referência oscilaram entre 93,90% e 83,54% para a classe U1-A e 94,51% a 91,77% para a classe U1-B. As sequências de U1-B de *A. lacustris* e *P. schubarti* são idênticas. Ao analisarmos apenas o gene U6, o comprimento apresentou-se consistente (Tabela 2) e as similaridades com a referência variaram entre 73,5% e 70,7%. Todos os snRNAs U6 recuperados estavam em associação ao U1-A.

Tabela 2. Tamanho dos *contigs* recuperados para as diferentes classes de U1 em espécies de peixes Characiformes. O tamanho da região gênica e seus respectivos espaçadores nas espécies também estão descritos.

Espécie	Classe	Contig	U1	Esp.	U6	Esp.
<i>Astyanax lacustris</i>	U1-A	3599	164	563	105	278
	U1-B	3123	164	-	-	-
<i>Astyanax mexicanus</i>	U1-A	6061	164	560	105	278
	U1-B	3860	165	-	-	-
<i>Bario sanctaefilomenae</i>	U1-A	4256	164	576	105	264
<i>Brycon orbignyianus</i>	U1-A	3628	164	501	105	259
<i>Characidium gomesi</i>	U1-A	4951	164	815	105	246
	U1-B	3870	164	-	-	-
<i>Characidium zebra</i>	U1-A	11279	164	1343	105	226
<i>Colossoma macropomum</i>	U1-A	3998	164	446	105	610
<i>Cyphocharax modestus</i>	U1-A	3145	164	510	105	130
<i>Erythrinus erythrinus</i>	U1-A	4026	164	401	105	251
<i>Hoplias intermedius</i>	U1-A	1736	164	402	105	195
<i>Hoplias malabaricus</i> KarD	U1-A	4233	164	411	105	263
<i>Megaleporinus elongatus</i>	U1-A	4830	164	841	105	179
<i>Megaleporinus macrocephalus</i>	U1-A	2592	164	848	105	179
<i>Piaractus mesopotamicus</i>	U1-A	5592	164	511	105	454
<i>Prochilodus lineatus</i>	U1-A	4161	164	500	105	152
<i>Prochilodus reticulatus</i>	U1-A	3246	164	527	105	152
<i>Psalidodon bockmanni</i>	U1-A	3338	164	567	105	278
	U1-B	2170	164	-	-	-
<i>Psalidodon paranae</i>	U1-A	5246	164	568	105	278
	U1-B	936	164	-	-	-
<i>Psalidodon schubarti</i>	U1-A	2682	164	569	107	279
	U1-B	1449	164	-	-	-
<i>Pyrrhulina marylinae</i>	U1-A	3970	164	527	105	241
<i>Pyrrhulina semifasciata</i>	U1-A	5224	164	526	105	240
<i>Triportheus auritus</i>	U1 + 5S	6397	164	741	-	-

A árvore de máxima verossimilhança das unidades de snRNA U1 confirmou a existência de dois grupos bem definidos (U1-A e U1-B), com notável dissimilaridade entre eles dentro dos genomas, mas elevada similaridade entre espécies para ambos os tipos. Curiosamente, embora os representantes da família Lebiasinidae, *P. marylinae* e *P. semifasciata*, apresentem sequências inicialmente classificadas como U1-A devido à sua associação ao gene U6, os resultados indicam que para esta família os U1 exibem maior

semelhança com aqueles classificados como U1-B. Através da árvore de máxima verossimilhança também foi possível inferir que em *Characidium gomesi* o gene U1-B não parece de fato de agrupar aos genes U1-B das demais espécies.

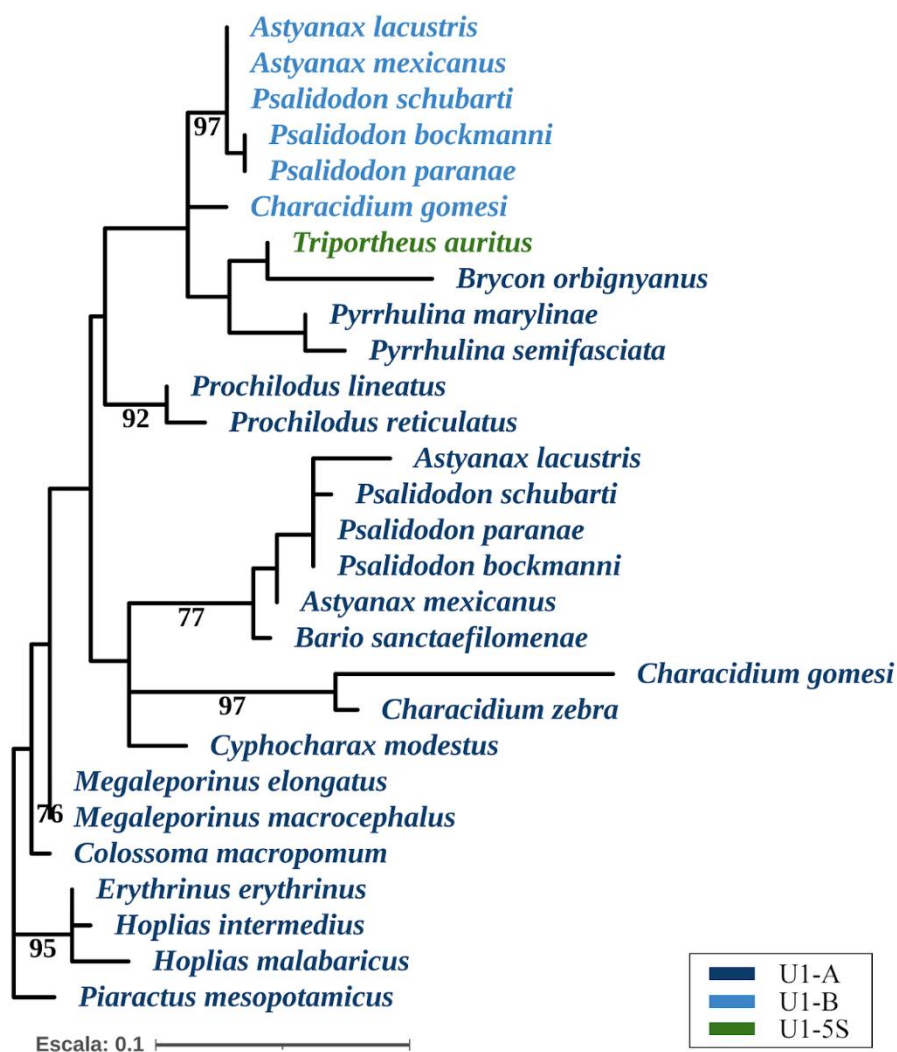


Figura 5. Árvore de máxima verossimilhança baseada nas sequências de snRNAs U1-A, U1-B e U1+5S em 22 espécies de peixes. Os valores nos ramos representam os suportes *bootstrap* (1.000 réplicas).

4.2. Associação entre snRNAs U2 e U5: ocorrência frequente de sintenia entre os dois genes

A investigação dos *arrays* correspondentes ao gene U2 evidenciou uma associação consistente entre esse gene e o snRNA U5. Entre as 22 espécies analisadas, 21 exibiram a organização U2-U5. A espécie *Pyrrhulina marylinae*, da família Lebiasinidae, foi a única a apresentar o snRNA U2 em tandem sem associação ao U5 (Figura Suplementar 3). A região gênica do snRNA U2 teve padrão de 188 nt, com poucas exceções (Tabela 3). Em relação à similaridade ao gene de referência as espécies variaram entre 99% e 85% de semelhança. Para o snRNA U5, a região codificadora demonstrou maior instabilidade no comprimento, com variações de 116 a 121 nt. Essa instabilidade também fica refletida nos valores mais baixos de similaridade à referência (entre 78,8% e 66%).

Tabela 3. Tamanho dos *contigs*, das regiões gênicas e dos espaçadores de U2 e U5.

Espécie	Contig	U2	Esp.	U5	Esp.
<i>Astyanax lacustris</i>	943	188	537	119	90
<i>Astyanax mexicanus</i>	2976	188	540	119	305
<i>Bario sanctaefilomenae</i>	1134	188	541	119	286
<i>Brycon orbignyanus</i>	2962	188	436-463	119	237
<i>Characidium gomesi</i>	848	188	587	116	312
<i>Characidium zebra</i>	1239	188	587	118	241
<i>Colossoma macropomum</i>	2790	188	508	118	283
<i>Cyphocharax modestus</i>	1218	189	527	120	381
<i>Erythrinus erythrinus</i>	930	188	373	119	250
<i>Hoplias intermedius</i>	3758	188	298-332	119	222-226
<i>Hoplias malabaricus</i> KarD	868	188	306	119	258
<i>Megaleporinus elongatus</i>	1139	188	539	118	294
<i>Megaleporinus macrocephalus</i>	921	188	288	118	327
<i>Piaractus mesopotamicus</i>	2700	188	561-566	116	286
<i>Prochilodus lineatus</i>	4505	188	786	116	684-686
<i>Prochilodus reticulatus</i>	1813	188	440	118	658
<i>Psalidodon bockmanni</i>	1225	188	551	118	207
<i>Psalidodon paranae</i>	3935	188	538	119	227-237
<i>Psalidodon schubarti</i>	1140	188	536	119	298
<i>Pyrrhulina marylinae</i>	4503	188	843-875	-	-
<i>Pyrrhulina semifasciata</i>	2628	190	180-183	119	299
<i>Triportheus auritus</i>	1408	190	804	121	295

4.3. Padrão genômico do snRNA U4: ausência de associação a outras famílias multigênicas

As iterações realizadas para a obtenção de *contigs* correspondentes ao gene U4 recuperaram *contigs* para as 22 espécies analisadas e não foram encontradas associações com outras famílias multigênicas. O U4 se mostrou o gene mais conservado do presente estudo, com o tamanho de 141 nucleotídeos (Tabela 4) em 21 espécies e altos valores de similaridade à referência (100% a 90%).

Tabela 4. Tamanho dos *contigs*, das regiões gênicas e dos espaçadores do snRNA U4.

Espécie	Contig	U4	Esp.
<i>Astyanax lacustris</i>	8410	141	1332-474
<i>Astyanax mexicanus</i>	8540	141	512-676
<i>Bario sanctaefilomenae</i>	5516	141	1170
<i>Brycon orbignyanus</i>	1397	141	560
<i>Characidium gomesi</i>	1523	141	742
<i>Characidium zebra</i>	1177	141	843
<i>Colossoma macropomum</i>	1145	141	609
<i>Cyphocharax modestus</i>	4230	141	975-744
<i>Erythrinus erythrinus</i>	768	141	300
<i>Hoplias intermedius</i>	2287	141	1006
<i>Hoplias malabaricus</i> KarD	2173	141	502
<i>Megaleporinus elongatus</i>	6804	141	656
<i>Megaleporinus macrocephalus</i>	2246	141	661
<i>Piaractus mesopotamicus</i>	4000	141	596
<i>Prochilodus lineatus</i>	2136	141	567
<i>Prochilodus reticulatus</i>	3717	141	564
<i>Psalidodon bockmanni</i>	5401	141	649
<i>Psalidodon paranae</i>	3392	141	652
<i>Psalidodon schubarti</i>	3247	141	873
<i>Pyrrhulina marylinae</i>	2533	141	280
<i>Pyrrhulina semifasciata</i>	2007	141	280
<i>Triportheus auritus</i>	3370	141	547

4.4. Pequenos RNAs nucleares no genoma de peixes Characiformes

A partir do alinhamento de todas as sequências consenso, foi possível prever a estrutura secundária dos pequenos RNAs nucleares. O snRNA U1 apresentou o mesmo padrão de quatro *stem-loops* previamente descrito por Azambuja et al. (2022), com energia livre de Gibbs de $-93,07$ kcal/mol. A estrutura secundária do snRNA U2 também apresentou quatro *stem-loops*, com energia livre de Gibbs de $-116,98$ kcal/mol, enquanto a do snRNA U4 foi caracterizada por um longo *stem-loop*, além de algumas alças menores em sua estrutura, apresentando energia livre de Gibbs de $-63,37$ kcal/mol. Para o snRNA U5, foi predita uma estrutura secundária composta por dois *stem-loops* e uma haste simples, com energia livre de Gibbs de $-56,68$ kcal/mol. Por fim, a estrutura secundária predita para o snRNA U6 consiste em um *stem-loop*, uma haste simples e algumas alças, com energia livre de Gibbs de $-36,89$ kcal/mol.

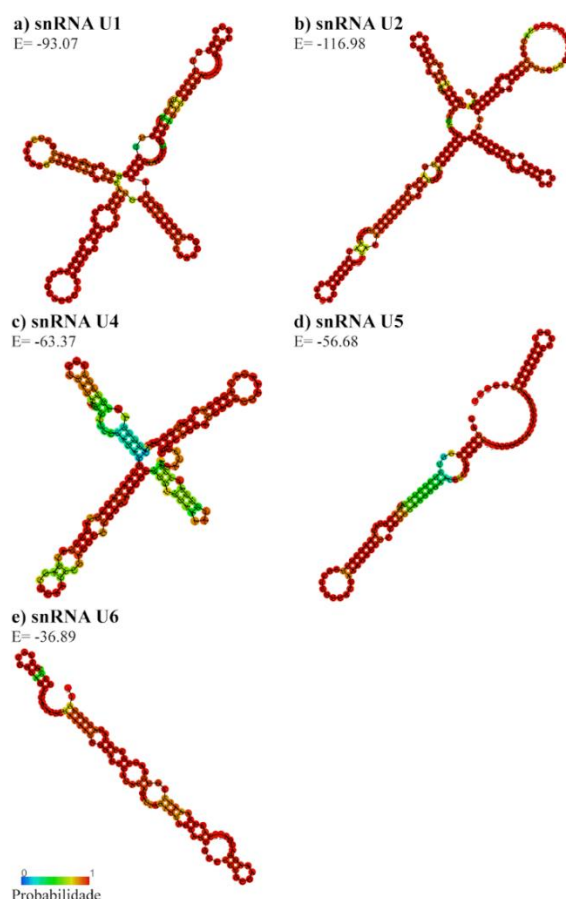


Figura 6. Estruturas secundárias preditas a partir do alinhamento das sequências dos pequenos RNAs nucleares U1, U2, U4, U5 e U6 recuperadas de 22 espécies de peixes Characiformes, com seus respectivos valores de energia livre.

Em todas as espécies de Characiformes analisadas os genes de snRNA identificados nos genomas apresentaram *motifs* conservados característicos de regiões promotoras. O elemento PSE foi detectado de forma recorrente em todos os snRNAs, enquanto a TATA-box foi observada unicamente nos genes U6. Para o snRNA U1, o PSE apresentou 21 nucleotídeos (5'-CCCHTCACCCTCCTCCTGCAT-3'), presente em 26 das 28 sequências analisadas e localizado entre as posições -64 e -48 em relação ao início do gene. Para o U2, um consenso de PSE com 25 nucleotídeos (5'-ACCGTGCAGTAGCTTCGGACTTGCG-3') foi identificado em 17 das 22 sequências, situado entre as posições -61 e -36. O consenso de PSE do snRNA U5 consistiu em uma sequência de 25 nucleotídeos (5'-CTACCCTCCGCCAGCTTCCTAGGCC-3'), presente em 19 das 21 sequências, localizado entre as posições -62 e -37. Um consenso de PSE (5'-RGTAACCGTTCCTGTGATGTC-3') entre as posições -63 e -43 foi observado em todas as sequências de snRNA U4. No caso do snRNA U6, o consenso PSE (5'-CTCTGAGACCTGCCCTCTTGCAGAA-3') foi identificado em 19 das 22 sequências analisadas, localizado entre as posições -77 e -52. Além disso, foi identificado um possível TATA-box (5'-TATATGTG-3') nas posições -32 a -24, presente em todas as sequências de U6.

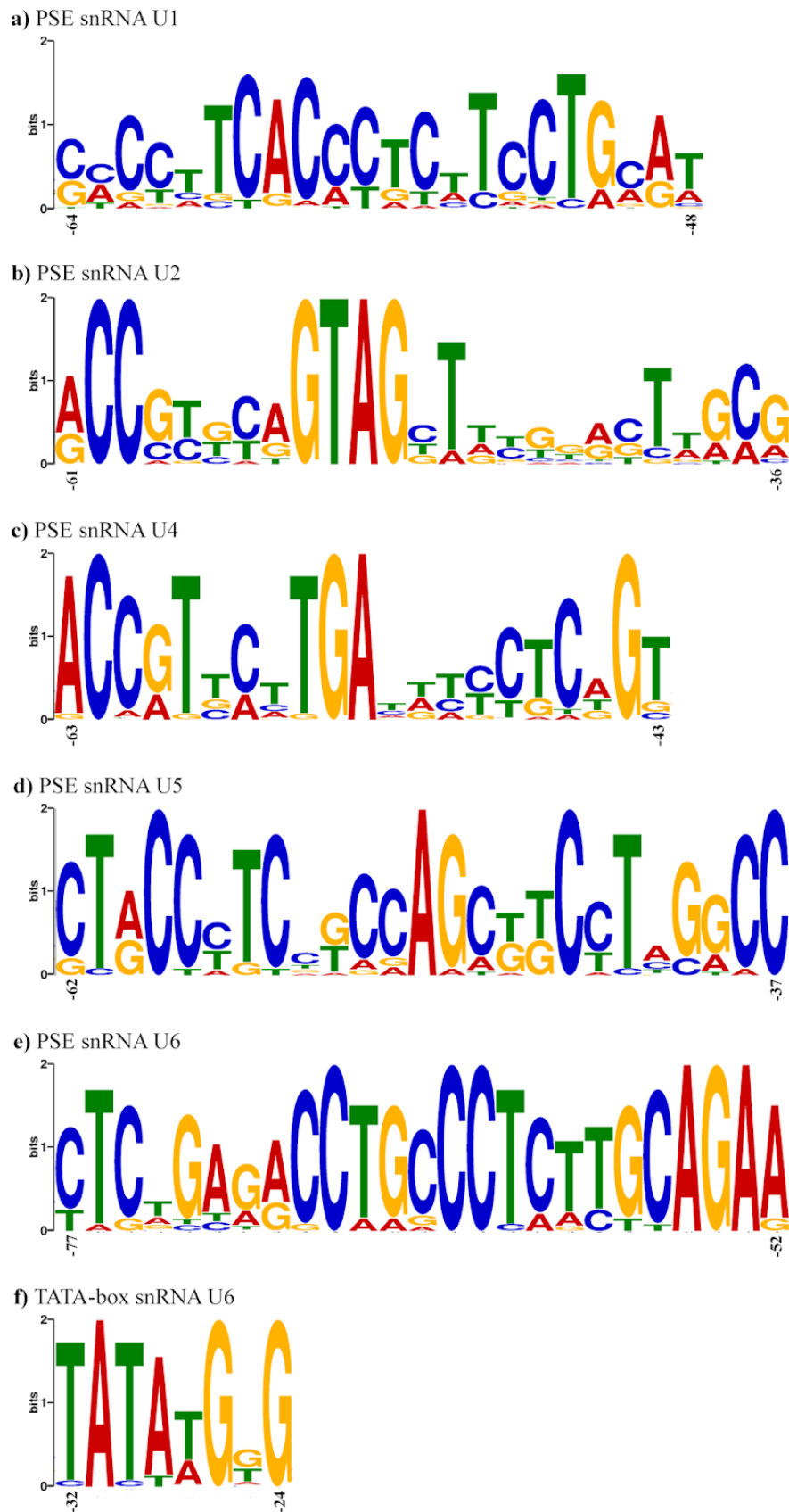


Figura 7. Elementos conservados identificados a montante das regiões gênicas de snRNAs nos genomas de peixes Characiformes.

Perfis de cobertura foram gerados a partir dos *contigs* U1-A e U6 e indicaram que, para a maioria das espécies analisadas, há predominância de maior cobertura nas regiões codificantes em relação às regiões espaçadoras (Figura Suplementar 5). Em casos excepcionais, picos de cobertura foram detectados especificamente em zonas intergênicas. Utilizando o *software* CENSOR foi possível estabelecer que alguns desses picos provavelmente correspondem a sequências repetitivas semelhantes a elementos transponíveis já previamente descritos em outros organismos. Em *Characidium zebra*, por exemplo, picos de cobertura no espaçador coincidem com retrotransposons do tipo Rex-Babar. No mesmo sentido, em espécies da família Anostomidae foi possível comprovar a presença de elementos que correspondem a classe de transposons de DNA piggyBac no espaçador U1-A e, apesar de compartilharem esse elemento, cada uma dessas espécies também apresenta hits com TEs distintos: *M. elongatus* exibe um elemento da família LINE-1 enquanto *M. macrocephalus* possui um elemento da família Bel-Pao.

De maneira geral, foi observado que sequências repetitivas que se assemelham a elementos transponíveis estão distribuídas entre os genes U1 e U6 de forma heterogênea, abrangendo diferentes classes que podem variar entre as espécies. Destaca-se o transposon Kolobok, na família Characidae, identificado nos *clusters* de *Psalidodon bockmanni*, *P. paranae* e *P. schubarti* e ausente nas demais espécies. *Astyanax mexicanus* apresentou associação a um transposon de DNA não especificado, enquanto em *Bario sanctaefilomenae* e *Astyanax lacustris* a busca não retornou resultados. Para a classe U1-B, os perfis seguiram o mesmo padrão de mapeamento da classe anterior, com predominância de cobertura nas regiões gênicas e casos específicos de mapeamento nos espaçadores que correspondem a elementos transponíveis (Figura Suplementar 6). Apesar de determinadas classes de TEs como LINE1, Rex-Babar e Mariner serem encontradas tanto em U1-A como em U1-B, esses elementos não são compartilhados entre espécies que possuem simultaneamente os dois *clusters* de U1. Destaca-se, ainda, a espécie *Characidium gomesi*, por ser a única a apresentar espaçadores de U1-B contendo regiões homólogas ao elemento UnaSINE1.

Além de elucidar as regiões espaçadoras dos diferentes *clusters* de U1, as análises com o CENSOR também revelaram que todas as unidades de U6 mostram ~88,91% de similaridade com o transposon de DNA DADA-U6, elemento móvel que tem como alvo de transposição o gene U6. TSDs característicos do Dada-U6 (5'GCGAAA3') foram

identificados nos *loci* snRNA U6 analisados aqui, assim como as repetições invertidas subterminais (5'CAGAGAAGA3') que estão separadas dos TSDs por 14 nucleotídeos. Apesar do CENSOR não ter retornado resultados de similaridade ao transposon de DNA Dada-U1, TSDs típicos desse TE foram detectados nas sequências de snRNA U1 (5'GCGAAT3'), juntamente com repetições invertidas subterminais (5'ATTGCAC3'), localizadas a 19 nucleotídeos dos TSDs. A repetição invertida subterminal de Dada-U1 coincide com o que já foi descrito anteriormente como Box C do gene U1 (Vierna et al., 2013; Figura 8).

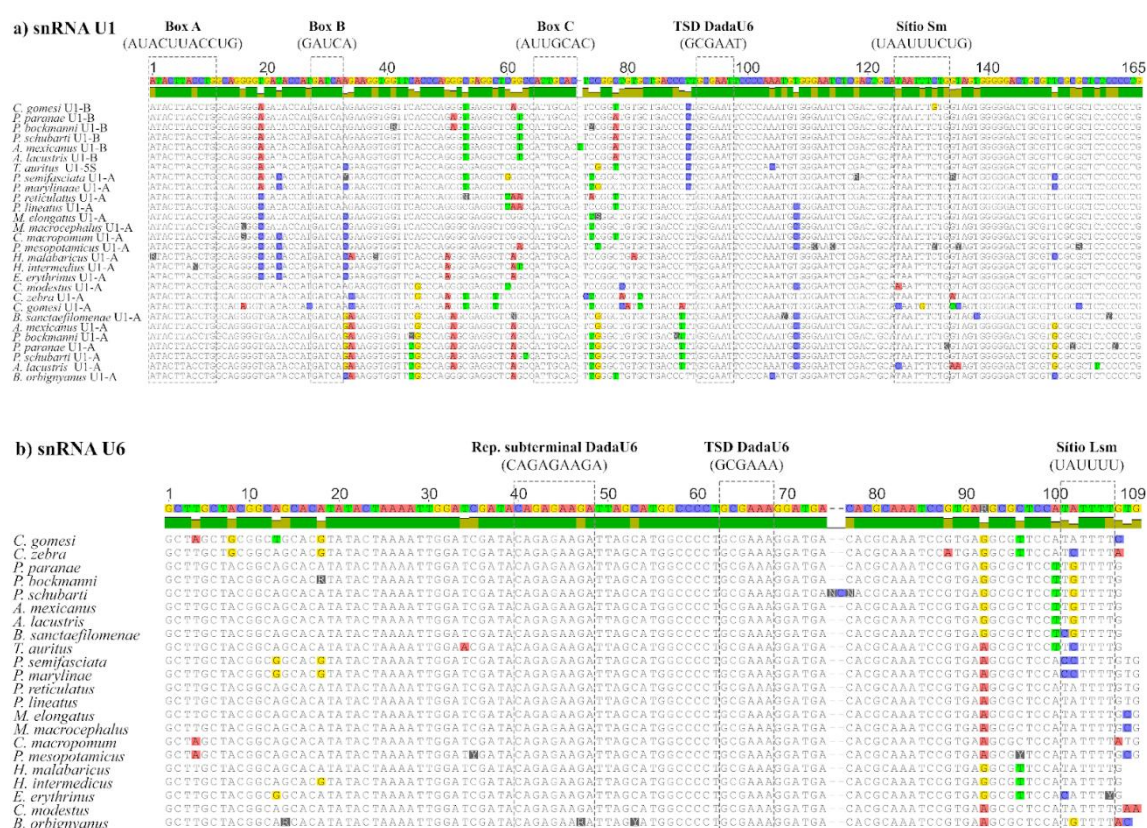


Figura 8. Regiões gênicas dos snRNAs U1 e U6 nas espécies analisadas evidenciando TSDs, repetições invertidas subterminais e sítios de ligação a proteínas Sm/Lsm. Em a) estão destacadas as boxes conservadas do gene U1.

As análises de cobertura para os genes U2, U5 e regiões adjacentes revelaram padrões semelhantes aos descritos anteriormente, onde o mapeamento ocorre predominantemente nas regiões gênicas dos *clusters* com sequências repetitivas

altamente mapeadas em determinados espaçadores (Figura Suplementar 7). Os elementos de maior ocorrência neste *cluster* multigênico correspondem a retrotransposons do tipo Gypsy e Copia. Adicionalmente, para todos os *loci* de U2, o *software* CENSOR mostrou ~85,86% de similaridade com os retrotransposons do tipo DIRS da espécie de raia *Amblyraja radiata*. Os *loci* de snRNA U5 não mostraram similaridade com sequências do banco de dados para a maioria das espécies. Entretanto, em casos específicos foram encontradas sobreposições dessa região gênica com elementos repetitivos como Tx1 de *Corbicula fluminea* (~85,33% em *Characidium gomesi*, *Characidium zebra* e *Triportheus auritus*), Bel-Pao de *Anthonomus grandis* (~80% em *Psalidodon paranae*) e P-element de *Solenopsis invicta* (~88% em *Astyanax lacustris*).

Os *clusters* de U4 foram os mais homogêneos em relação ao mapeamento observado na região gênica e intergênica (Figura Suplementar 8). A espécie *Hoplias intermedius* apresentou o padrão mais distinto para este gene, com um espaçador altamente mapeado que possui semelhança com três tipos distintos de retrotransposons. Uma série de elementos foram identificados nos espaçadores, além de haver sobreposição de TEs na região gênica U4 para as espécies *Erythrinus erythrinus* (transposon hAT-5 de *Ilyophis*), *Megaleporinus elongatus* (LTR Copia de *Picea abies*) e *Astyanax mexicanus* (retrotransposon Gypsy de *Drosophila suzukii*).

O mapeamento de RNA-seq em bibliotecas com depleção de RNA ribossomal da espécie *Psalidodon paranae* indicou que ambas as classes de U1 estão potencialmente sendo transcritas no genoma da espécie, assim como as unidades de snRNAs U2, U4 e U5. Para o snRNA U6, o mapeamento revelou transcrição gênica em um nível inferior em comparação aos demais. Com exceção do U4, os genes de snRNA possuem espaçadores intergênicos que são transcritos em algum nível.

4.5. Análise de genomas completos e associação de snRNAs a outros elementos repetitivos

A partir do genoma completo de *Astyanax mexicanus* foi possível identificar um *array* da classe U1-A localizado no cromossomo 18, com aproximadamente 260 mil pares de bases, contendo 224 cópias do gene U1 e 225 cópias de U6. Neste mesmo cromossomo se encontra o *cluster* do snRNA U4, identificado com 150 mil pares de bases e 96 cópias do gene. A classe U1-B foi recuperada em um único *array* no cromossomo 8, com cerca de 40 mil pares de bases e 39 cópias de U1. Além disso, foram recuperados dois *arrays* de U2-U5 em cromossomos distintos de *A. mexicanus*, sendo o maior deles composto por 60 mil pares de bases e 40 cópias de U2 e 40 de cópias de U5 (cromossomo 6) e o menor apresenta 34 mil pares de bases, 20 cópias de U2 e 20 cópias de U5 (cromossomo 27).

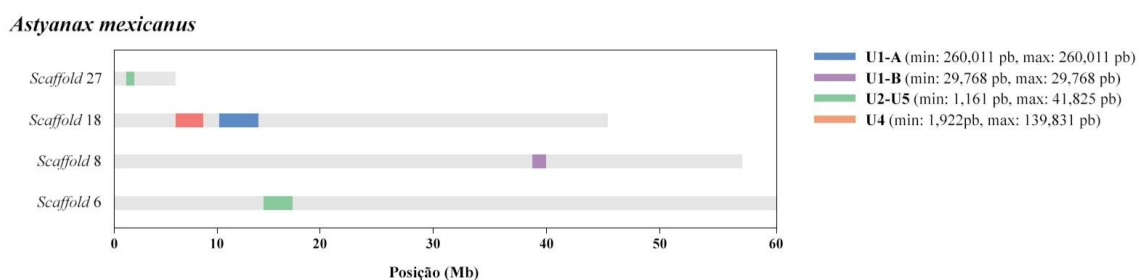


Figura 9. Distribuição e organização dos *arrays* de snRNAs no genoma de *Astyanax mexicanus*.

Em *Hoplias malabaricus* os resultados confirmam o padrão anteriormente sugerido por análises com *short-reads*, demonstrando a presença exclusiva da classe U1-A, representada por um *array* de 142 mil pares de bases, com 196 cópias de U1 e 212 de U6, no cromossomo 2 da espécie. Neste mesmo cromossomo também foi identificado o *array* de U2 (622 cópias) e U5 (617 cópias), com 193 mil pares de bases. O *array* do gene U4, por sua vez, está localizado no cromossomo 4 da espécie e apresentou 92 mil pares de bases e 176 cópias do gene.

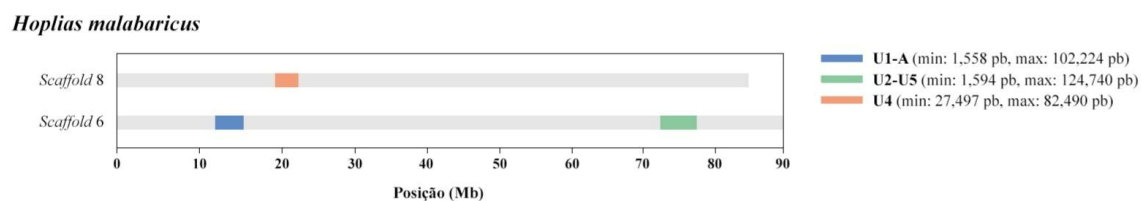


Figura 10. Distribuição e organização dos *arrays* de snRNAs no genoma de *Hoplias malabaricus*.

A busca por domínios conservados revelou, no *array* U1-A de *Astyanax mexicanus*, a presença de um domínio Exonuclease-Endonuclease-Phosphatase (EEP) e dois domínios relacionados à transcriptase reversa (RT) de retrotransposons e retrovírus Non-LTR (*long terminal repeats*). Um domínio de função desconhecida (DUF6729) também foi localizado ao final do *array*. Em contrapartida, a busca utilizando o *array* U1-B de *Astyanax mexicanus* não encontrou nenhum domínio. Para U1-A de *Hoplias malabaricus* um único domínio conservado (DUF6729) foi localizado. Em ambas as espécies os *arrays* U2-U5 apresentaram hits com múltiplos elementos como Non-LTR, endonucleases L1-EN e domínios de ligação a zinco em RT, distribuídos de maneira distinta do início ao fim dos *arrays*. No *array* U4 de *Hoplias malabaricus*, a análise de domínios conservados revelou a presença de um domínio SCAN próximo ao início do *array*, seguido de domínios de ligação a zinco de integrase e domínio central de integrase. Também foram detectadas transcriptases reversas e o domínio de RNase H da família Ty3/Gypsy. No *array* U4 de *Astyanax mexicanus*, os domínios identificados são relacionados a domínios de transposases (Figura 11). Os intervalos e E-value correspondentes a cada domínio estão listados na Tabela Suplementar 1 e 2.

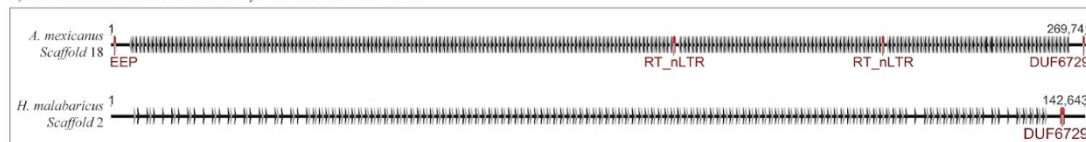
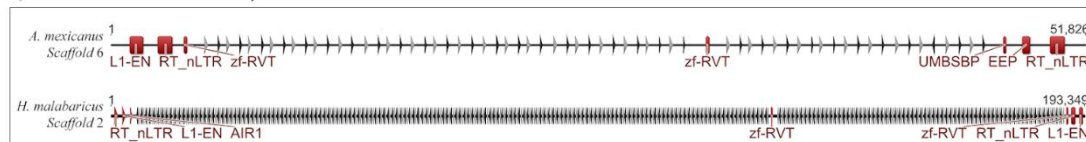
a) Domínios identificados em *arrays* de snRNA U1-A e U6b) Domínios identificados em *arrays* de snRNA U2 e U5c) Domínios identificados em *arrays* de snRNA U4

Figura 11. Domínios conservados identificados em *arrays* de pequenos RNAs nucleares obtidos através do genoma completo das espécies *Astyanax mexicanus* e *Hoplias malabaricus*.

5. Discussão

Nossas análises mostraram a existência de duas classes de genes U1 (U1-A e U1-B), cada um com sequências consenso distintas. Uma forte relação entre a região codificadora tida como U1-A e o gene U6 foi estabelecida, enquanto o U1-B foi majoritariamente encontrado formando *clusters* isolados. As análises filogenéticas comprovaram o padrão de agrupamento gênico interespecífico, com unidades homogeneizadas de U1-A e U1-B. Tal homogeneização das repetições é o padrão esperado para a evolução em concerto de famílias multigênicas em *arrays* clusterizados, por meio de crossing-over desigual e/ou conversão gênica (Mount et al., 2007), mas a existência e persistência de longo prazo de diferentes classes do U1 sugere fortemente que o modelo de *birth-and-death* (Nei & Rooney, 2005) também atua durante a evolução dessa família repetitiva em peixes, indicando uma combinação complexa de deleções, inserções e principalmente duplicações ocorridas ao longo da evolução desses genes nos peixes.

Como esperado em um sistema submetido à evolução em concerto, os genes que codificam U1, U2, U4, U5 e U6 apresentaram pouca ou nenhuma variabilidade de sequência intragenômica, embora haja ampla variação nos espaçadores intergênicos. Dessa forma, é provável que a recombinação genética esteja atuando em diferentes frequências ao longo dos *arrays*, sugerindo uma homogeneização incompleta das unidades, como demonstrado em diferentes modelos (Galián et al., 2013; Harpke & Peterson, 2006; Lunerová et al., 2017).

A análise de sequências de snRNAs e suas regiões flangeadoras em peixes mostrou associação com diferentes domínios associados a elementos transponíveis (Cabral-De-Mello et al., 2012; Dulz et al., 2020; Malimpensa et al., 2020; Pucci et al., 2018; Yano et al., 2020), mostrando que esses elementos podem estar envolvidos na dispersão das sequências de snRNA. A identificação de TSDs típicos de transposons tipo Dada nos genes U1 e U6 corroboram com a hipótese de que elementos transponíveis se inserem em lugares específicos de pequenos RNAs nucleares para realizar transposição (Kojima & Jurka, 2013). Os snRNAs são ótimos alvos para elementos transponíveis já que existem em grande número no genoma de peixes, são altamente conservadas no genoma das espécies e, apesar de serem essenciais para o organismo, a redundância faz com que um possível erro na sequência não prejudique o organismo (Kojima & Jurka, 2013). A coincidência da Box C do gene U1 com a repetição subinvertida terminal do TE DadaU1 pode ser um indício de que essa classe de elemento móvel tem papel na amplificação e

dispersão dos *loci* de U1 e U6 nos peixes Characiformes, utilizando sequências conservadas presentes nas regiões gênicas de pequenos RNAs nucleares para realizar a transposição. A espécie *Characidium gomesi*, por exemplo, apresentou um U1-B filogeneticamente distante em relação aos demais representantes dessa classe gênica. O fato de seus espaçadores estarem associados a elementos transponíveis distintos e exclusivos reforça a hipótese de que TEs podem atuar na diferenciação e dispersão dessas unidades (Kojima & Jurka, 2013; Pucci et al., 2018). Casos semelhantes já foram descritos para Us em outros vertebrados, nos quais a associação a retrotransposons contribui para a mobilidade e a diversificação da família gênica (Doucet et al., 2015; Malimpensa et al., 2020).

A sintenia consistente entre U2 e U5 destaca a forte conservação evolutiva dessa associação, sugerindo que a proximidade física entre esses genes pode estar relacionada a requisitos co-transcricionais (Cross & Rebordinos, 2005). A ausência de U5 em associação com U2 em *Pyrrhulina marylinae* pode representar uma exceção que reflete eventos específicos de rearranjo cromossômico ou perda local, alinhando-se ao conceito de evolução por *birth-and-death* em famílias multigênicas (Eirín-López et al., 2012). Os *clusters* de pequenos RNAs U4 mostraram-se altamente conservados tanto em sequência quanto em organização, sem associações com outras famílias multigênicas. Porém, embora não formem *clusters* sintênicos, snRNAs U4 não estão imunes à influência de TEs, com descrição de inserções sobrepostas ao gene.

Ambas as unidades de U1 foram encontradas no genoma montado de *A. mexicanus*, ao passo que apenas a unidade U1-A está presente no genoma de *H. malabaricus*, confirmando os padrões inferidos a partir de *short-reads*, com *arrays* multigênicos de grande extensão e associados a diferentes domínios específicos de elementos transponíveis. Apesar da ligação entre distintas unidades de famílias multigênicas ser relativamente comum e conservada, sugerimos que sua manutenção evolutiva não é necessariamente estável. A ligação entre U1 e U6 foi observada em todas as famílias analisadas no presente estudo, mas parece ser mais restrita a estes grupos uma vez que não foi descrita anteriormente. Nesse sentido, a organização genômica dos genes U parece ser lábil, como refletido, por exemplo, na ligação entre U1 e U2, restrita a uma única família da ordem Pleuronectiformes (Manchado et al., 2006); e na mesma ligação U1-U2 em *Argyrosomus regius* (Perciformes), que provavelmente surgiu de forma independente (Merlo et al., 2012) e não está presente em nenhuma das famílias analisadas aqui. Da

mesma forma, nem todos os padrões descritos por Silva et al. (2015) estão necessariamente refletidos nos representantes da família Characidae analisados neste estudo. A espécie *Astyanax jordani*, considerada a mais próxima filogeneticamente de *A. mexicanus*, compartilha com esta última a presença de um par cromossômico contendo U2. Entretanto, *A. jordani* apresenta três pares cromossômicos de U1, sendo um par submetacêntrico co-localizado com rDNA 5S, enquanto *A. mexicanus* não apresenta essa associação sintênica (Silva et al., 2015).

As predições de estruturas secundárias para os snRNAs de peixes Characiformes mostraram padrões compatíveis com aqueles descritos por Azambuja et al. (2022) e que são recorrentes em diversos peixes e outros vertebrados. A preservação desses arranjos estruturais reforça a ideia de que a maioria dos snRNAs desempenha funções essenciais nas células desses organismos. Além da conservação estrutural, também foram observados elementos regulatórios típicos associados à transcrição desses genes. Em Characiformes, também conseguimos identificar a presença dos elementos PSE direcionando o recrutamento do mecanismo específico da RNA polimerase II em snRNAs U1-A, U1-B, U2, U4 e U5 enquanto os snRNAs U6 apresentaram os dois principais elementos para o recrutamento da RNA polimerase III (PSE e TATA-box) (Hernandez, 2001).

Os gráficos de cobertura e o mapeamento de RNA-seq indicaram que os pequenos RNAs nucleares estão ativos em termos transcricionais, ainda que em diferentes níveis de expressão. Os espaçadores intergênicos, tradicionalmente considerados regiões não funcionais, também apresentaram sinais de transcrição, sugerindo que podem abrigar sequências regulatórias ou resíduos de inserções móveis. Esses achados corroboram a ideia de que o espaçador não é neutro, mas sim um componente dinâmico sujeito à influência de TEs e possivelmente envolvido em mecanismos de evolução (Wicker et al., 2007).

Por fim, nosso trabalho demonstrou que a montagem de famílias multigênicas a partir de *short-reads* facilita a análise dessas sequências em espécies de peixes. Também possibilita a realização de estudos evolutivos amplos considerando diversos táxons. Como já é conhecido, o conteúdo dos genes U tende a ser bastante conservado entre táxons relacionados. No entanto, essa condição estável não se reflete na organização desses genes já que múltiplas e independentes transições na composição e ordem gênica de cada *cluster* parecem ter ocorrido diversas vezes durante a evolução dos peixes. Por

exemplo, fornecemos evidências da ocorrência de duas classes de U1 em peixes da família Characidae e Crenuchidae que se agrupam separadamente nos genomas; e ao menos uma linhagem de U1 também está presente nos demais Characiformes, seguindo seus próprios caminhos evolutivos, como previsto pela teoria do modelo misto de evolução de famílias multigênicas.

6. Referências

- Anjos, A., Ruiz-Ruano, F. J., Camacho, J. P. M., Loreto, V., Cabrero, J., De Souza, M. J., & Cabral-De-Mello, D. C. (2015). U1 snDNA clusters in grasshoppers: Chromosomal dynamics and genomic organization. *Heredity*, *114*(2), 207–219. <https://doi.org/10.1038/hdy.2014.87>
- Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., & Betancur, R. R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology and Evolution*, *1*(2). <https://doi.org/10.1038/s41559-016-0020>
- Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), W39–W49. <https://doi.org/10.1093/nar/gkv416>
- Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., & Stadler, P. F. (2008). RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, *9*. <https://doi.org/10.1186/1471-2105-9-474>
- Black, D. L. (2003). *Mechanisms of Alternative Pre-Messenger RNA Splicing*. <http://www.mimc.ucla.edu/faculty/black/index.html>
- Cabral-De-Mello, D. C., Valente, G. T., Nakajima, R. T., & Martins, C. (2012). Genomic organization and comparative chromosome mapping of the U1 snRNA gene in cichlid fish, with an emphasis in *Oreochromis niloticus*. *Chromosome Research*, *20*(2), 279–292. <https://doi.org/10.1007/s10577-011-9271-y>
- Chairi, H., & Gonzalez, L. R. (2015). Structure and Organization of the Engraulidae Family U2 snRNA: An Evolutionary Model Gene? *Journal of Molecular Evolution*, *80*(3–4), 209–218. <https://doi.org/10.1007/s00239-015-9674-z>
- Charlesworth, B., Be, P. S., & Stephan, W. (1994). *The evolutionary dynamics of repetitive DNA in eukaryotes*.
- Cross, I., & Rebordinos, L. (2005). 5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas* oysters: Does the (CT)_n·(GA)_n microsatellite stabilize this novel linkage of large tandem arrays? *Genome*, *48*(6), 1116–1119. <https://doi.org/10.1139/g05-075>
- Dai, W., Zou, M., Yang, L., Du, K., Chen, W., Shen, Y., Mayden, R. L., & He, S. (2018). Phylogenomic Perspective on the Relationships and Evolutionary History of the Major Otocephalan Lineages. *Scientific Reports*, *8*(1). <https://doi.org/10.1038/s41598-017-18432-5>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). <https://doi.org/10.1093/gigascience/giab008>

- De Souza E Sousa, J. F., Viana, P. F., Bertollo, L. A. C., Cioffi, M. B., & Feldberg, E. (2017). Evolutionary Relationships among Boulengerella Species (Ctenoluciidae, Characiformes): Genomic Organization of Repetitive DNAs and Highly Conserved Karyotypes. *Cytogenetic and Genome Research*, 152(4), 194–203. <https://doi.org/10.1159/000480141>
- Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45(4). <https://doi.org/10.1093/nar/gkw955>
- Domitrovich, A. M., & Kunkel, G. R. (2003). Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. *Nucleic Acids Research*, 31(9), 2344–2352. <https://doi.org/10.1093/nar/gkg331>
- Doucet, A. J., Droc, G., Siol, O., Audoux, J., & Gilbert, N. (2015). U6 snRNA pseudogenes: Markers of retrotransposition dynamics in mammals. *Molecular Biology and Evolution*, 32(7), 1815–1832. <https://doi.org/10.1093/molbev/msv062>
- Dulz, T. A., Azambuja, M., Nascimento, V. D., Lorscheider, C. A., Noletto, R. B., Moreira-Filho, O., Nogaroto, V., Diniz, D., de Mello Affonso, P. R. A., & Vicari, M. R. (2020). Karyotypic diversification in two megaleporinus species (Characiformes, Anostomidae) inferred from in situ localization of repetitive DNA sequences. *Zebrafish*, 17(5), 333–341. <https://doi.org/10.1089/zeb.2020.1918>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Eickbush, T. H., & Eickbush, D. G. (2007). Finely orchestrated movements: Evolution of the ribosomal RNA genes. In *Genetics* (Vol. 175, Number 2, pp. 477–485). <https://doi.org/10.1534/genetics.107.071399>
- Eirín-López, J. M., Rebordinos, L., Rooney, A. P., & Rozas, J. (2012). The Birth-and-Death Evolution of Multigene Families Revisited. In *Genome Dyn. Basel, Karger* (Vol. 7).
- Ferreira, I. A., Oliveira, C., Venere, P. C., Galetti, P. M., & Martins, C. (2007). 5S rDNA variation and its phylogenetic inference in the genus Leporinus (Characiformes: Anostomidae). *Genetica*, 129(3), 253–257. <https://doi.org/10.1007/s10709-006-0005-6>
- Galián, J. A., Rosato, M., & Rosselló, J. A. (2013). *Partial Sequence Homogenization in the 5S Multigene Families May Generate Sequence Chimeras and Spurious Results in Phylogenetic Reconstructions Downloaded from.* <http://sysbio.oxfordjournals.org/>
- García-Souto, D., Troncoso, T., Pérez, M., & Pasantes, J. J. (2015). Molecular cytogenetic analysis of the European hake *Merluccius merluccius* (Merlucciidae,

- Gadiformes): U1 and U2 snRNA gene clusters map to the same location. *PLoS ONE*, 10(12). <https://doi.org/10.1371/journal.pone.0146150>
- Garrido-Ramos, M. A. (2017). Satellite DNA: An evolving topic. In *Genes* (Vol. 8, Number 9). MDPI AG. <https://doi.org/10.3390/genes8090230>
- Gazoni, T., Dorigon, N. S., Da Silva, M. J., Cholak, L. R., Haddad, C. F. B., & Parise-Maltempi, P. P. (2021). Chromosome Mapping of U2 snDNA in Species of *Leptodactylus* (Anura, Leptodactylidae). *Cytogenetic and Genome Research*, 161(1–2), 63–69. <https://doi.org/10.1159/000515047>
- Glugoski, L., Deon, G., Schott, S., Vicari, M. R., Nogaroto, V., & Moreira-Filho, O. (2020). Comparative cytogenetic analyses in ancistrus species (Siluriformes: Loricariidae). *Neotropical Ichthyology*, 18(2), 1–16. <https://doi.org/10.1590/1982-0224-2020-0013>
- Harpke, D., & Peterson, A. (2006). Non-concerted ITS evolution in Mammillaria (Cactaceae). *Molecular Phylogenetics and Evolution*, 41(3), 579–593. <https://doi.org/10.1016/j.ympev.2006.05.036>
- Hernandez, N. (2001). Small Nuclear RNA Genes: A Model System to Study Fundamental Mechanisms of Transcription. In *Journal of Biological Chemistry* (Vol. 276, Number 29, pp. 26733–26736). <https://doi.org/10.1074/jbc.R100032200>
- Jawdekar, G. W., & Henry, R. W. (2008). Transcriptional regulation of human small nuclear RNA genes. In *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (Vol. 1779, Number 5, pp. 295–305). <https://doi.org/10.1016/j.bbagr.2008.04.001>
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6), 587–589. <https://doi.org/10.1038/nmeth.4285>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. In *Genetica* (Vol. 115).
- Kojima, K. K. (2015). A new class of SINEs with snRNA gene-derived heads. *Genome Biology and Evolution*, 7(6), 1702–1712. <https://doi.org/10.1093/gbe/evv100>
- Kojima, K. K., & Jurka, J. (2013). A Superfamily of DNA Transposons Targeting Multicopy Small RNA Genes. *PLoS ONE*, 8(7). <https://doi.org/10.1371/journal.pone.0068260>

- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Liao, D. (1999). Concerted evolution: Molecular mechanism and biological implications. *American Journal of Human Genetics*, 64(1), 24–30. <https://doi.org/10.1086/302221>
- Liu, J.-L., & Gall, J. G. (2007). *U bodies are cytoplasmic structures that contain uridine-rich small nuclear ribonucleoproteins and associate with P bodies*. www.pnas.org/cgi/content/full/
- Lunerová, J., Renny-Byfield, S., Matyášek, R., Leitch, A., & Kovařík, A. (2017). Concerted evolution rapidly eliminates sequence variation in rDNA coding regions but not in intergenic spacers in *Nicotiana tabacum* allotetraploid. *Plant Systematics and Evolution*, 303(8), 1043–1060. <https://doi.org/10.1007/s00606-017-1442-7>
- Malimpensa, G. D. C., Traldi, J. B., Martinez, J. D. F., Deon, G., Azambuja, M., Nogaroto, V., Vicari, M. R., & Moreira-Filho, O. (2020). Chromosomal diversification in two species of pimelodus (Siluriformes: Pimelodidae): Comparative cytogenetic mapping of multigene families. *Zebrafish*, 17(4), 278–286. <https://doi.org/10.1089/zeb.2020.1892>
- Manchado, M., Zuasti, E., Cross, I., Merlo, A., Infante, C., & Rebordinos, L. (2006). Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: A new linkage to the U1, U2, and U5 small nuclear RNA genes. *Genome*, 49(1), 79–86. <https://doi.org/10.1139/g05-068>
- Martins, C., Ferreira, I. A., Oliveira, C., Foresti, F., & Galetti, P. M. (2006). A tandemly repetitive centromeric DNA sequence of the fish *Hoplias malabaricus* (Characiformes: Erythrinidae) is derived from 5S rDNA. *Genetica*, 127(1–3), 133–141. <https://doi.org/10.1007/s10709-005-2674-y>
- Marz, M., Kirsten, T., & Stadler, P. F. (2008). Evolution of spliceosomal snRNA genes in metazoan animals. *Journal of Molecular Evolution*, 67(6), 594–607. <https://doi.org/10.1007/s00239-008-9149-6>
- Marzluff, W. F., Brown, D. T., Lobo, S., & Wang, S.-S. (1983). *Isolation and characterization of two linked mouse Ulb small nuclear RNA genes* (Vol. 1).
- Matera, A. G., Terns, R. M., & Terns, M. P. (2007). Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs. In *Nature Reviews Molecular Cell Biology* (Vol. 8, Number 3, pp. 209–220). <https://doi.org/10.1038/nrm2124>
- Matera, A. G., & Wang, Z. (2014). A day in the life of the spliceosome. In *Nature Reviews Molecular Cell Biology* (Vol. 15, Number 2, pp. 108–121). <https://doi.org/10.1038/nrm3742>

- Mattaj, I. W., Tollervey, D., & Séraphin, B. (1993). Small nuclear RNAs in messenger RNA and ribosomal RNA processing. *The FASEB Journal*, 7(1), 47–53.
<https://doi.org/10.1096/fasebj.7.1.8422974>
- Mattaj, I. W., Zeller, R., & De Robertis, E. (1983). *Xenopus laevis* U2 snRNA genes: tandemly repeated transcription units sharing 5' and 3' flanking homology with other RNA polymerase II transcribed genes Switzerland Communicated by. In *The EMBO Journal* (Vol. 2, Number 11).
- Melo, B. F., Ota, R. P., Benine, R. C., Carvalho, F. R., Lima, F. C. T., Mattox, G. M. T., Souza, C. S., Faria, T. C., Reia, L., Roxo, F. F., Valdez-Moreno, M., Near, T. J., & Oliveira, C. (2024). Phylogenomics of Characidae, a hyper-diverse Neotropical freshwater fish lineage, with a phylogenetic classification including four families (Teleostei: Characiformes). *Zoological Journal of the Linnean Society*, 202(1).
<https://doi.org/10.1093/zoolinlean/zlae101>
- Melo, B. F., Sidlauskas, B. L., Near, T. J., Roxo, F. F., Ghezelayagh, A., Ochoa, L. E., Stiassny, M. L. J., Arroyave, J., Chang, J., Faircloth, B. C., MacGuigan, D. J., Harrington, R. C., Benine, R. C., Burns, M. D., Hoekzema, K., Sanches, N. C., Maldonado-Ocampo, J. A., Castro, R. M. C., Foresti, F., ... Oliveira, C. (2022). Accelerated Diversification Explains the Exceptional Species Richness of Tropical Characoid Fishes. *Systematic Biology*, 71(1), 78–92.
<https://doi.org/10.1093/sysbio/syab040>
- Merlo, M. A., Cross, I., Palazón, J. L., Úbeda-Manzanaro, M., Sarasquete, C., & Rebordinos, L. (2012). Evidence for 5S rDNA Horizontal Transfer in the toadfish *Halobatrachus didactylus* (Schneider, 1801) based on the analysis of three multigene families. *BMC Evolutionary Biology*, 12(1).
<https://doi.org/10.1186/1471-2148-12-201>
- Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K., & Makalowski, W. (2007). Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA*, 13(1), 5–14.
<https://doi.org/10.1261/rna.259207>
- Nakatani, M., Miya, M., Mabuchi, K., Saitoh, K., & Nishida, M. (2011). Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. *BMC Evolutionary Biology*, 11(1).
<https://doi.org/10.1186/1471-2148-11-177>
- Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Wainwright, P. C., Friedman, M., & Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, 109(34), 13698–13703.
<https://doi.org/10.1073/pnas.1206625109>
- Nei, M., & Rooney, A. P. (2005). *Concerted and Birth-and-Death Evolution of Multigene Families* *.

- Nilsen, T. W. (2003). The spliceosome: The most complex macromolecular machine in the cell? In *BioEssays* (Vol. 25, Number 12, pp. 1147–1149). <https://doi.org/10.1002/bies.10394>
- Patel, A. A., & Steitz, J. A. (2003). Splicing double: Insights from the second spliceosome. In *Nature Reviews Molecular Cell Biology* (Vol. 4, Number 12, pp. 960–970). <https://doi.org/10.1038/nrm1259>
- Pelliccia, F., Barzotti, R., Bucciarelli, E., & Rocchi, A. (2001). 5S ribosomal and U1 small nuclear RNA genes: A new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. *Genome*, 44(3), 331–335. <https://doi.org/10.1139/gen-44-3-331>
- Pinhal, D., Yoshimura, T. S., Araki, C. S., & Martins, C. (2011). The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: An example from freshwater stingrays. *BMC Evolutionary Biology*, 11(1). <https://doi.org/10.1186/1471-2148-11-151>
- Ponzio, J. C., Piscor, D., & Parise-Maltempi, P. P. (2018). Chromosomal locations of U2 snDNA clusters in Megaleporinus, Leporinus and Schizodon (Characiformes: Anostomidae). *Biologia (Poland)*, 73(3), 295–298. <https://doi.org/10.2478/s11756-018-0031-8>
- Pucci, M. B., Nogaroto, V., Moreira-Filho, O., & Vicari, M. R. (2018). Dispersion of transposable elements and multigene families: Microstructural variation in Characidium (Characiformes: Crenuchidae) genomes. *Genetics and Molecular Biology*, 41(3), 585–592. <https://doi.org/10.1590/1678-4685-gmb-2017-0121>
- Reis, R. E., Albert, J. S., Di Dario, F., Mincarone, M. M., Petry, P., & Rocha, L. A. (2016). Fish biodiversity and conservation in South America. *Journal of Fish Biology*, 89(1), 12–47. <https://doi.org/10.1111/jfb.13016>
- Ricardo Vicari, M., Cristina de Almeida, M., Antonio Carlos Bertollo, L., Moreira-Filho, O., & Ferreira Artoni, R. (2006). *Cytogenetic analysis and chromosomal characteristics of the polymorphic 18S rDNA in the fish Prochilodus lineatus (Characiformes, Prochilodontidae)*. www.sbg.org.br
- Rooney, A. P., & Ward, T. J. (2005). *Evolution of a large ribosomal RNA multigene family in filamentous fungi: Birth and death of a concerted evolution paradigm*. www.ncbi.nlm.nih.gov/genomes/FUNGI/funtab.html
- Silva, D. M. Z. A., Utsunomia, R., Pansonato-Alves, J. C., Oliveira, C., & Foresti, F. (2015). Chromosomal mapping of repetitive DNA sequences in five species of astyanax (Characiformes, Characidae) reveals independent location of U1 and U2 snRNA sites and association of U1 snRNA and 5S rDNA. *Cytogenetic and Genome Research*, 146(2), 144–152. <https://doi.org/10.1159/000438813>

- Thi Hoang, D., Chernomor, O., von Haeseler, A., Quang Minh, B., Sy Vinh, L., & Rosenberg, M. S. (2017). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*, 35(2), 518–522. <https://doi.org/10.5281/zenodo.854445>
- Trifinopoulos, J., Nguyen, L. T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44(W1), W232–W235. <https://doi.org/10.1093/NAR/GKW256>
- Valadkhan, S. (2005). snRNAs as the catalysts of pre-mRNA splicing. In *Current Opinion in Chemical Biology* (Vol. 9, Number 6, pp. 603–608). Elsevier Ltd. <https://doi.org/10.1016/j.cbpa.2005.10.008>
- Vierna, J., Wehner, S., Höner Zu Siederdissen, C., Martínez-Lage, A., & Marz, M. (2013). Systematic analysis and evolution of 5S ribosomal DNA in metazoans. *Heredity*, 111(5), 410–421. <https://doi.org/10.1038/hdy.2013.63>
- Walsh, J. B. (1987). *Sequence-Dependent Gene Conversion: Can Duplicated Genes Diverge Fast Enough to Escape Conversion?*
- Wheaton Dimmick, W., & Larson, A. (1996). A Molecular and Morphological Perspective on the Phylogenetic Relationships of the Otophysan Fishes. In *MOLECULAR PHYLOGENETICS AND EVOLUTION* (Vol. 6, Number 1).
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., Sanmiguel, P., & Schulman, A. H. (2007). *A unified classification system for eukaryotic transposable elements*. www.nature.com/reviews/genetics
- Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, 3(7), 1–2. <https://doi.org/10.1101/cshperspect.a003707>
- Yano, C. F., Bertollo, L. A. C., Rebordinos, L., Merlo, M. A., Liehr, T., Portela-Bens, S., & Cioffi, M. de B. (2017). Evolutionary Dynamics of rDNAs and U2 Small Nuclear DNAs in *Triportheus* (Characiformes, Triportheidae): High Variability and Particular Syntenic Organization. *Zebrafish*, 14(2), 146–154. <https://doi.org/10.1089/zeb.2016.1351>
- Yano, C. F., Merlo, M. A., Portela-Bens, S., Cioffi, M. de B., Bertollo, L. A. C., Santos-Júnior, C. D., & Rebordinos, L. (2020). Evolutionary Dynamics of Multigene Families in *Triportheus* (Characiformes, Triportheidae): A Transposon Mediated Mechanism? *Frontiers in Marine Science*, 7. <https://doi.org/10.3389/fmars.2020.00006>

7. Material Suplementar

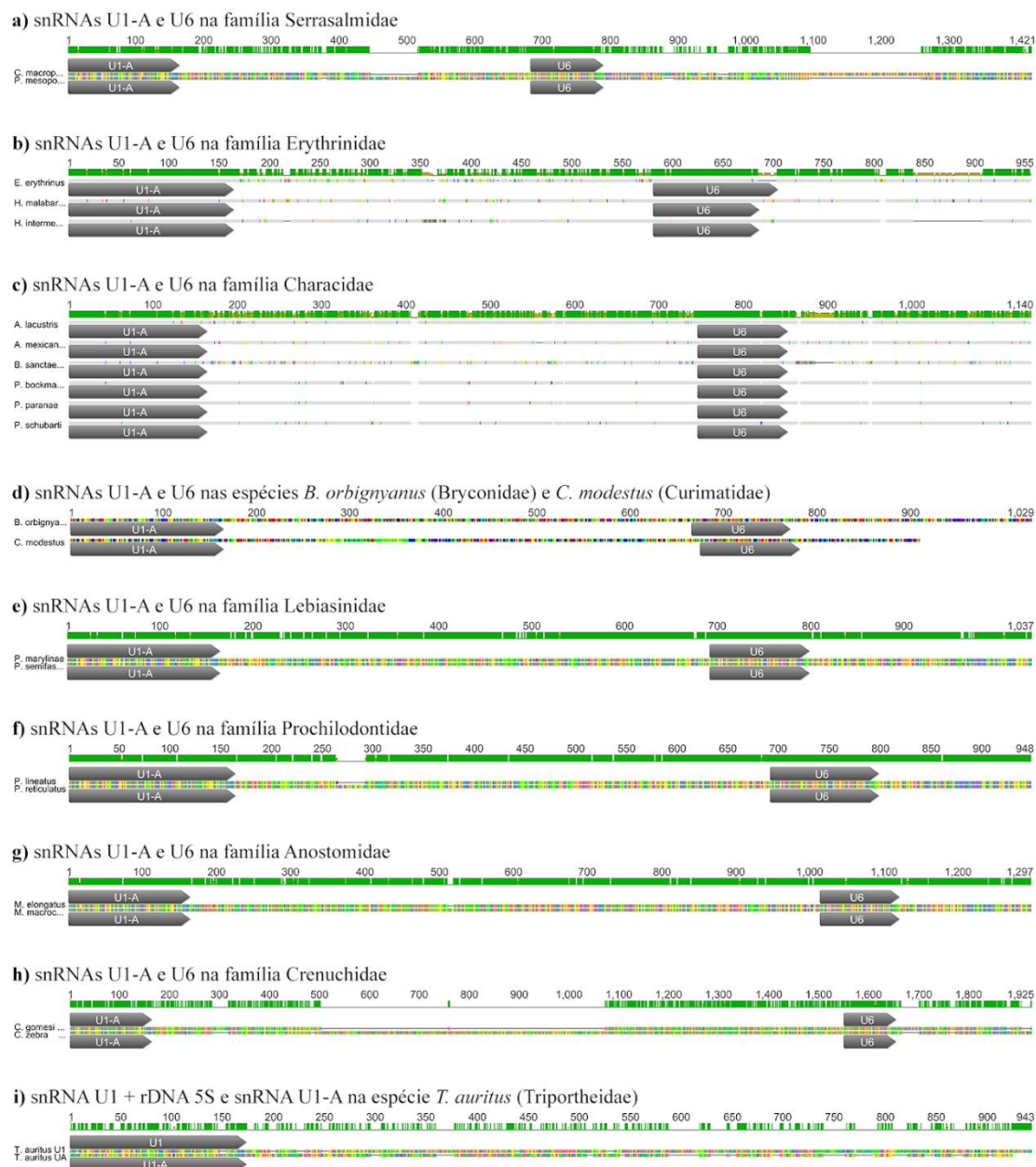


Figura Suplementar 1. Alinhamentos dos *contigs* de U1-A entre espécies de uma mesma família. Em **d)** estão representados os *contigs* não alinhados das famílias Bryconidae e Curimatidae. Em **i)** estão alinhados os *contigs* de U1-A e U1+5S na espécie *Triportheus auritus*.

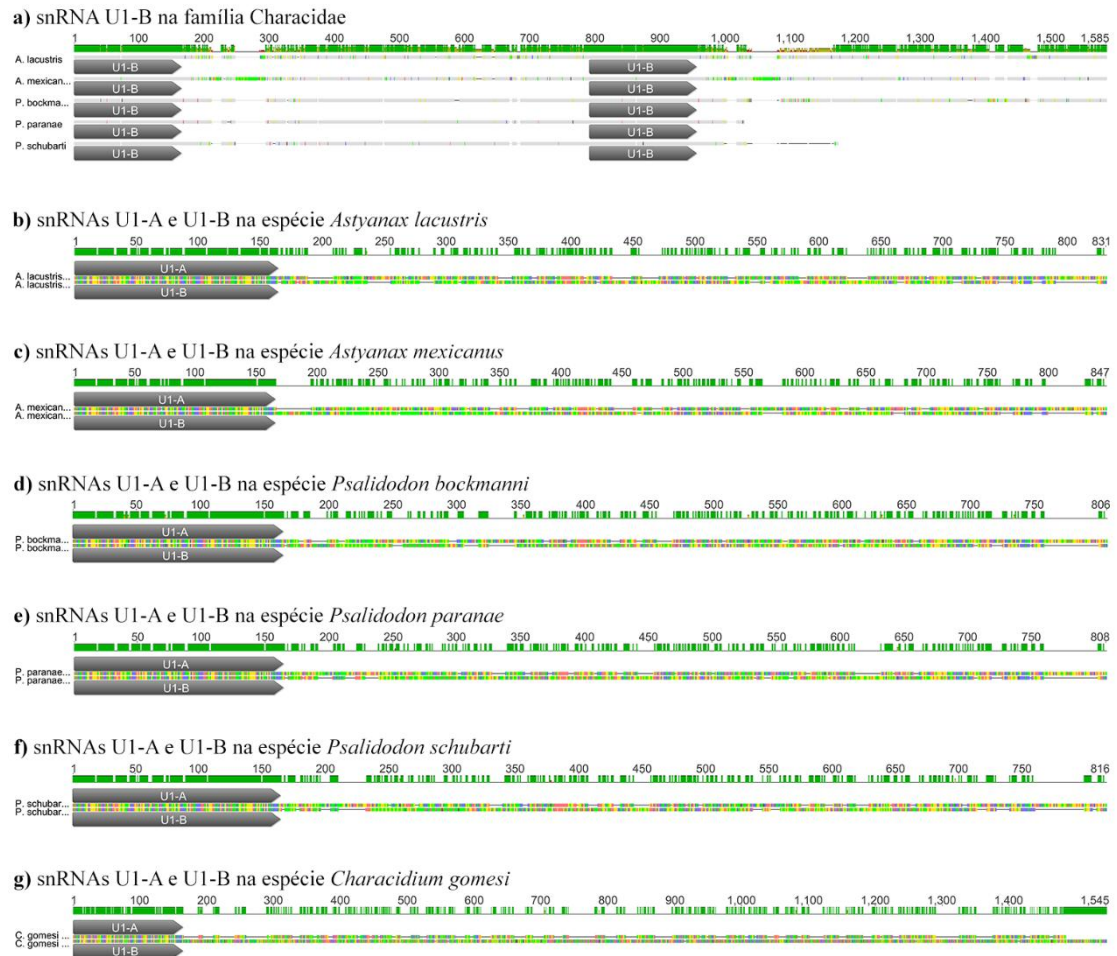


Figura Suplementar 2. Alinhamentos dos *contigs* de U1-A e U1-B em uma mesma espécie. Em **a)** estão representados os *contigs* alinhados de U1-B na família Characidae.

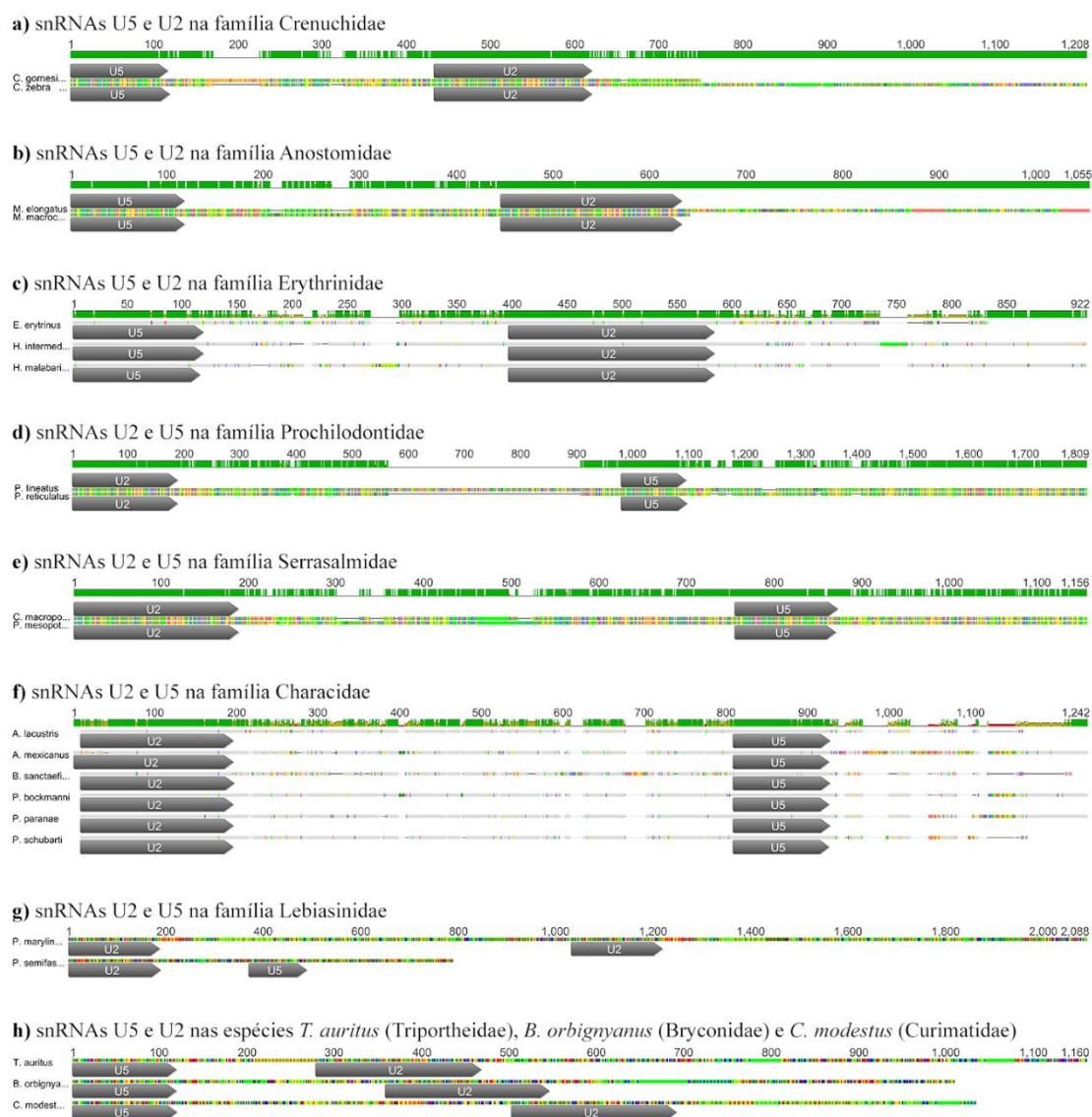


Figura Suplementar 3. Alinhamentos dos *contigs* de U2-U5 entre espécies de uma mesma família. Em **g)** estão representados os *contigs* não alinhados das espécies *P. marylinae* (U2) e *P. semifasciata* (U2-U5), representantes da família Lebiasinidae, e em **h)**, os *contigs* não alinhados das espécies das famílias Triportheidae, Bryconidae e Curimatidae.

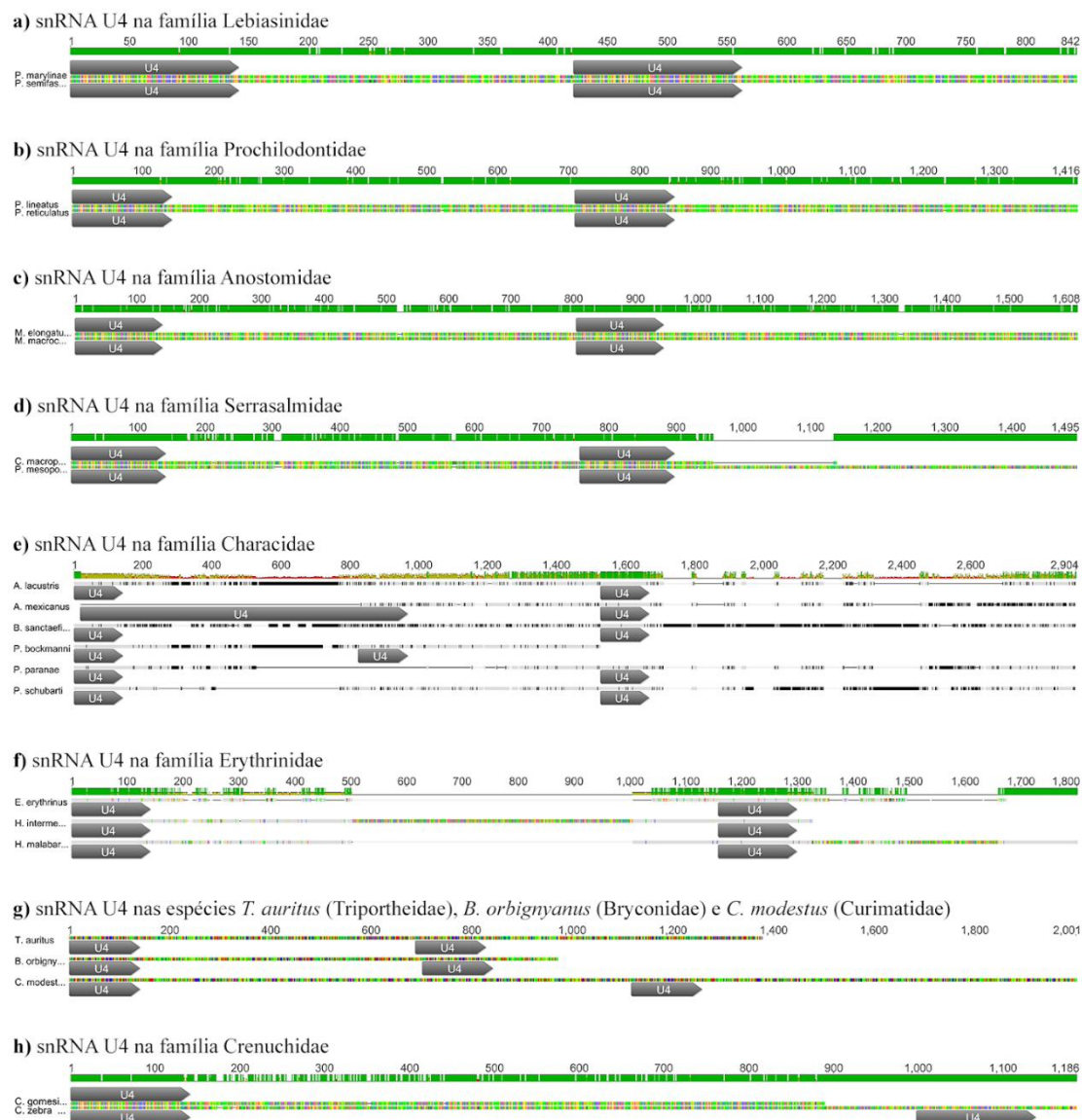


Figura Suplementar 4. Alinhamentos dos *contigs* de U4 entre espécies de uma mesma família. Em **g)** estão representados os *contigs* não alinhados das espécies das famílias Triportheidae, Bryconidae e Curimatidae.

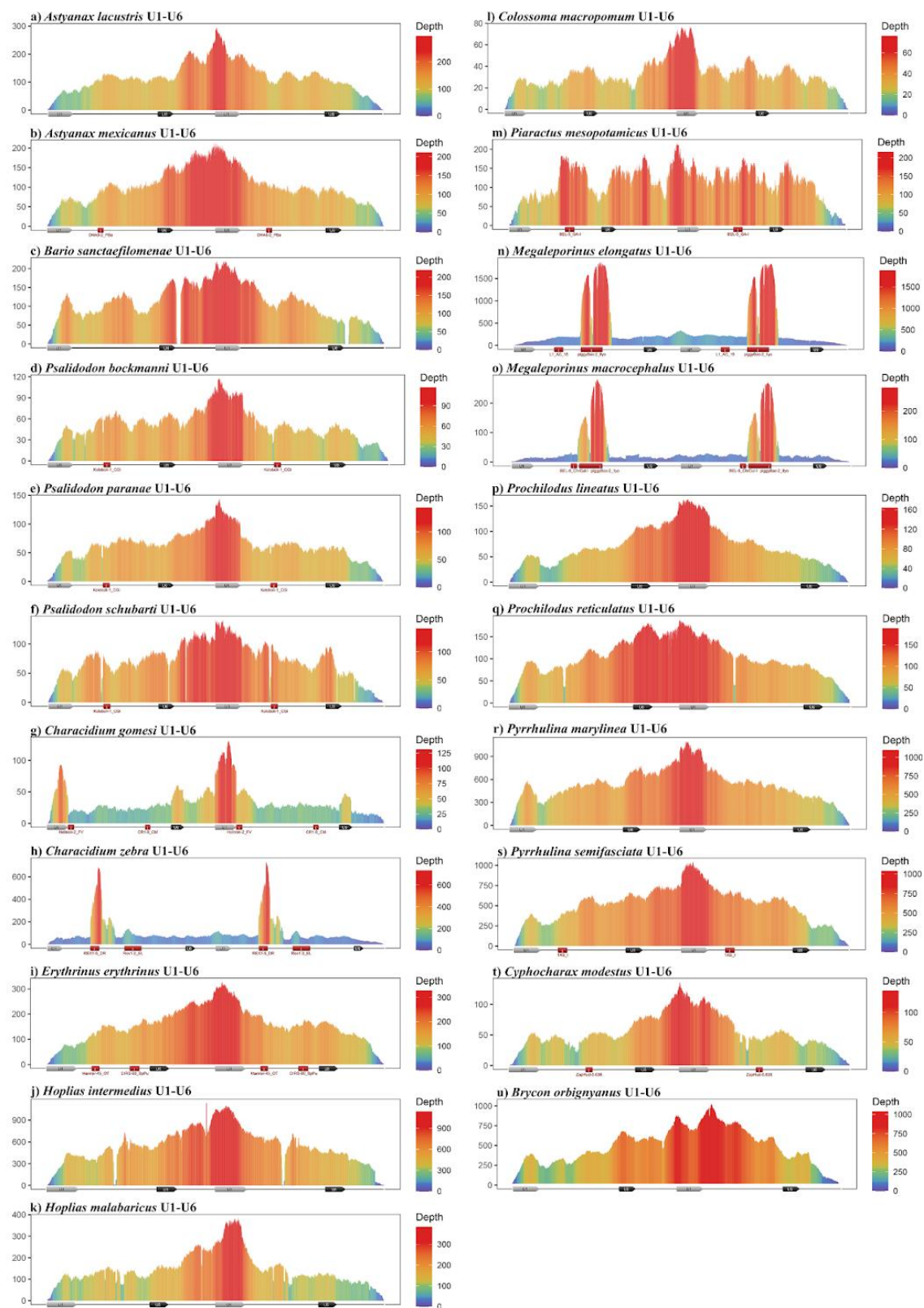


Figura Supplementar 5. Perfis de cobertura gerados a partir de *clusters* de snRNAs U1-U6 e U6. Elementos móveis identificados nos *clusters* também estão evidenciados.

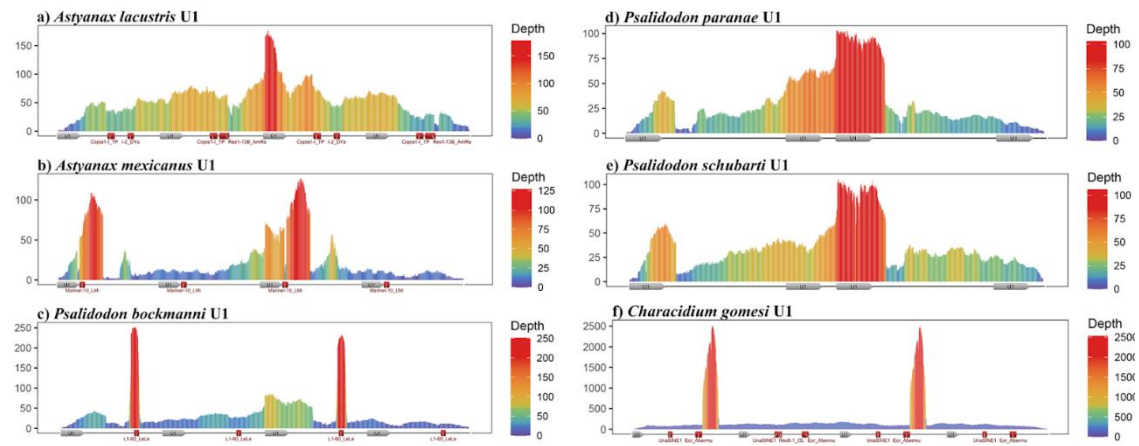


Figura Suplementar 6. Perfis de cobertura gerados a partir de *clusters* de snRNA U1-B. Elementos móveis identificados nos *clusters* também estão evidenciados.



Figura Suplementar 7. Perfis de cobertura gerados a partir de *clusters* de snRNAs U2 e U5. Elementos móveis identificados nos *clusters* também estão evidenciados.

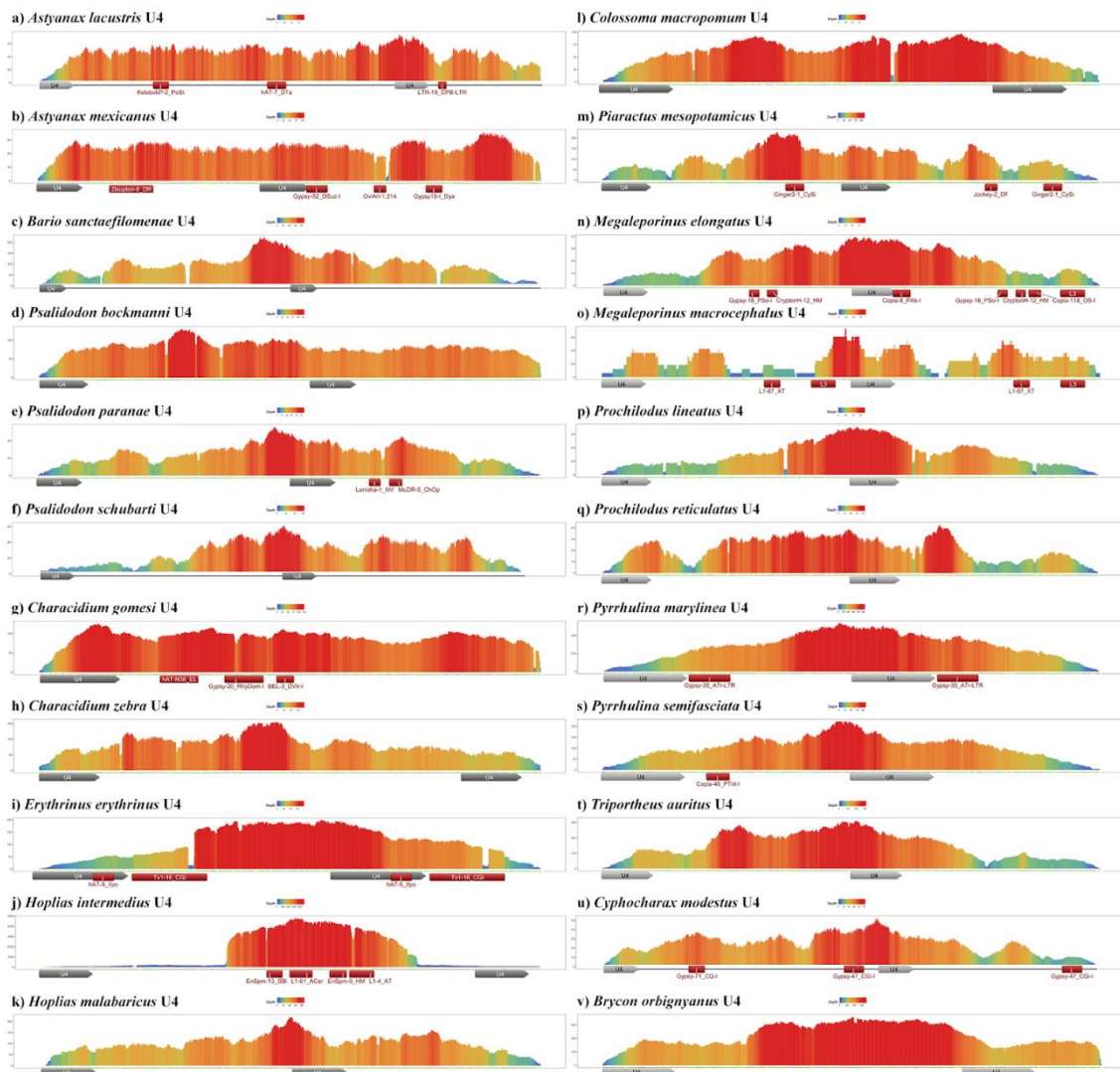


Figura Suplementar 8. Perfis de cobertura gerados a partir de *clusters* de snRNAs U4. Elementos móveis identificados nos *clusters* também estão evidenciados.

Tabela Suplementar 1. Domínios conservados identificados nos *arrays* de pequenos RNAs nucleares de *Hoplias malabaricus* com respectivos códigos de identificação, intervalos genômicos e valores de significância.

U1A-U6			
Lista de hits	Identificação	Intervalo	E-value
Domain of unknown function	DUF6729	139.052-139.750	6.80e-111
U2-U5			
Lista de hits	Identificação	Intervalo	E-value
Non-LTR retrotransposon and non-LTR reverse transcriptase	RT_nLTR_like	721-1497	7.67e-55
Endonuclease domain	L1-EN	2.293-2.940	2.98e-28
Arginine methyltransferase-interacting protein	AIR1	3.988-4.269	5.74e-06
Zinc-binding in reverse transcriptase	zf-RVT	131.158-131.295	6.39e-04
Zinc-binding in reverse transcriptase	zf-RVT	189.898-190.107	8.19e-08
Non-LTR retrotransposon and non-LTR reverse transcriptase	RT_nLTR_like	190.717-191.484	1.60e-55
Endonuclease domain	L1-EN	192.280-192.969	4.19e-44
U4			
Lista de hits	Identificação	Intervalo	E-value
SCAN domain	SCAN	151-417	1.45e-09
Integrase zinc binding domain	Integrase_H2C2	1.447-1.617	1.39e-17
Integrase core domain	rve	1.723-1.965	6.92e-11
Reverse transcriptases	RT_LTR	2.902-3.426	1.19e-65
Ty3/Gypsy family of RNase HI	RNase_HI_RT_Ty3	3.709-4.068	2.79e-48

Tabela Suplementar 2. Domínios conservados identificados nos *arrays* de pequenos RNAs nucleares de *Astyanax mexicanus* com respectivos códigos de identificação, intervalos genômicos e valores de significância.

U1A-U6			
Lista de hits	Identificação	Intervalo	E-value
Exonuclease-Endonuclease-Phosphatase	EEP	819-1.361	9.68e-05
Non-LTR retrotransposon and non-LTR reverse transcriptase	RT_nLTR	155.646-156.329	1.51e-30
Non-LTR retrotransposon and non-LTR reverse transcriptase	RT_nLTR	213.379-214.062	1.42e-30
Domain of unknown function	DUF6729	269.312-269.656	4.03e-47
U2-U5			
Lista de hits	Identificação	Intervalo	E-value
Non-LTR retrotransposon and non-LTR reverse transcriptase	RT_nLTR_like	1.112-1.885	4.82e-48
Exonuclease-Endonuclease-Phosphatase	EEP super family	2.928-3.344	1.44e-22
Universal minicircle sequence binding protein	UMSBP	4.183-4.362	1.19e-08
Zinc-binding in reverse transcriptase	zf-RVT	19.978-20.187	1.26e-08
Zinc-binding in reverse transcriptase	zf-RVT	47.750-47.959	6.99e-09
Non-LTR retrotransposon and non-LTR reverse transcriptase	RT_nLTR_like	48.568-49.335	1.26e-54
Endonuclease domain	L1-EN	50.131-50.820	1.42e-44
U4			
Lista de hits	Identificação	Intervalo	E-value
Transposase	HTH_Tnp_Tc3	148.787-148.972	5.12e-10
IS630 family transposase	transpos IS630	148.817-149.152	1.18e-06