UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

LORENA MARIA RUDNIK

ANÁLISE EM LARGA ESCALA DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS DE PEIXES

DISSERTAÇÃO

CORNÉLIO PROCÓPIO
2023

LORENA MARIA RUDNIK

ANÁLISE EM LARGA ESCALA DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS DE PEIXES

Large-scale analysis of Elements Transposables in fish genomes.

Dissertação apresentada como requisito à obtenção do grau de Mestre em Bioinformática PPGBIOINFO do Programa da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Alexandre R. Paschoal
Coorientador: Laurival A. Vilas-Boas
Colaborador: Roberto Ferreira Artoni
Colaboradora: Liliane Santana OliveiraColaborador: Pedro Gabriel Nachtigall

CORNÉLIO PROCÓPIO

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

LORENA MARIA RUDNIK

ANÁLISE EM LARGA ESCALA DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS DE PEIXES

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 21 de Agosto de 2023

Dr. Alexandre Rossi Paschoal, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Danillo Pinhal, Doutorado - Universidade Estadual Paulista Júlio de Mesquita Filho (Unesp) Dr.

Laurival Antonio Vilas Boas, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Renata Da Rosa, Doutorado - Universidade Estadual de Londrina (Uel)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 21/08/2023.

Dedico este trabalho a
todos aqueles a quem esta
pesquisa possa ajudar de
alguma forma

AGRADECIMENTOS

A minha família, meus pais João e Cláudia, e aos meus irmãos, Luis, Larissa e Lucas, que sempre estiveram ao meu lado dando todo o incentivo e suporte, sendo meus modelos de coragem e perseverança.

Agradeço ao meu orientador Prof. Dr. Alexandre Rossi Paschoal e ao meu coorientador Prof. Dr. Laurival Vilas-Boas, pelo incentivo e paciência, por toda a ajuda e confiança nessa trajetória.

Agradeço aos colaboradores Prof. Dr. Roberto Ferreira Artoni, Liliane Santana Oliveira e Pedro Gabriel Nachtigall, que me auxiliaram e contribuíram no desenvolvimento deste trabalho.

A NAPI de Bioinformática e a Fundação Araucária, órgão de fomento à pesquisa científica, pelo apoio financeiro.

Ao PPGBioinfo, pelo auxílio financeiro direcionados para participações em eventos de bioinformática.

A UTFPR campus Cornélio Procópio, pela estrutura e a realização deste trabalho, por todo o conhecimento dos docentes que fizeram parte da minha formação.

A todos os meus colegas e amigos da Bioinfo, em especial a Jéssica, Murilo, Nayane, Sara e Vitor, pelo companheirismo, força e apoio, vocês foram essenciais nesses anos.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo”

Albert Einstein

LISTA DE FIGURA

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BR	Brasil
cDNA	DNA complementar
CNGB	*China National GeneBank (Banco Genético Nacional da China)*
DIRs	Sequência de Repetição Intermediária de Dictyostelium (*Dictyostelium Intermediate Repeat Sequence*)
DB	Banco de dados (*Data base*)
DNA	Ácido desoxirribonucleico
EDTA	*Extensive de-novo TE Annotator*
ERV	Endogenous retroviruses
GB	*Gigabytes*
GFF	*Gene-Finding Format* (Formato de descoberta de genes)
LINE	Elementos nucleares longos intercalados (*Long Interpersed Nuclear Element*)
LTR	Sequências terminais repetidas (*Long Terminal Repeat*)
NCBI	*National Center for Biotechnology Information* (Centro Nacional de Informação sobre Biotecnologia)
NGDC	*National Genomics Data Center* (Centro Nacional de Dados Genômicos)
RNA	Ácido ribonucleico
RNAm	RNA mensageiro
SINE	Elementos nucleares curtos intercalados (*Short Interspersed Nuclear Element*)
TERL	Aprendiz de representação de elementos transponíveis (*Transposable Elements Representation Learner*)
TEs	Elementos genéticos móveis (*Transposable elements*)
TIR	Repetição terminal invertida (*Terminal Inverted Repeat*)
tRNA	RNA transportador
UEL	Universidade Estadual de Londrina
UEPG	Universidade Estadual de Ponta Grossa
UNESP	Universidade Estadual Paulista
UTFPR	Universidade Tecnológica Federal do Paraná

SUMÁRIO

RESUMO

Os elementos transponíveis (TE) têm sido amplamente estudados devido à sua diversidade e abundância nos genomas de várias espécies, e seu impacto na variabilidade genômica e papel funcional. A presença dos TEs em peixes também pode ter implicações para a conservação e manejo de espécies ameaçadas, uma vez que esses elementos podem afetar a estabilidade genômica e a saúde geral dos animais. Ainda que estejamos na era da grande massa de dados, a informação de qualidade e em larga-escala de TE em peixes é pobre, seja pela quantidade ou curadoria dos dados. Nesse sentido, de modo a preencher esta lacuna, esta pesquisa tem como objetivo a análise da anotação em larga escala dos elementos transponíveis em 168 genomas completos de peixes. Para atingir este objetivo, foi criado um pipeline com abordagens distintas de modo a cobrir estratégias diferentes para anotação dos elementos. Os resultados obtidos neste estudo fornecem informações importantes para estudos funcionais e evolutivos do papel dos TEs nos peixes. Com a revisão bibliográfica obtivemos respostas de ferramentas e banco de dados utilizados nas pesquisas com a mesma temática. Com o fluxo de trabalho criado, utilizando três ferramentas, e tendo comparação com banco de dados FISHTEDB e a pesquisa de Hilsdorf, encontramos ao total 107.866 elementos em 168 espécies de peixes. Importante ressaltar que o trabalho contou com o apoio da NAPI Bioinformática via Fundação Araucária por meio do convênio 66/2021.

Palavras-chave: Peixes. Elementos transponíveis. Bioinformática. Anotação automática.

Capítulo 1.

1.1 Introdução

Aquicultura é o cultivo e criação de animais aquáticos, sendo a atividade agropecuária que mais cresce no Brasil. O aumento do consumo mundial de peixes é motivado pelo crescimento da população mundial e os benefícios nutricionais que proporciona, além da alta quantidade dos estoques de peixes nativos (BORGUETTI et al., 2003).

De acordo com o Anuário Peixe BR de 2023 da Piscicultura (Associação Brasileira de Piscicultura), o Brasil produziu 860.355 toneladas de peixes em 2022, tendo um crescimento de 2,3% (Anuário Peixes BR, 2023). Os peixes nativos do Brasil, sendo a maior produção da espécie *Colossoma macropomum* (Tambaqui), representaram 31,04% da produção nacional no ano de 2022, referente à 267.060 toneladas. Além disso, a maior concentração da produção do Tambaqui se encontra na região Norte do país (Associação Brasileira da Piscicultura, 2023). No geral, o estado do Paraná é o maior produtor de peixes de cultivo do país, principalmente com a espécie Tilápia, correspondendo a 166 mil toneladas em 2021 segundo o CNA (Confederação da Agricultura e Pecuária do Brasil, 2021).

Para as importações do Brasil, o peixe que obteve maior adesão foi o Salmão, sendo 89% da importação, seguido por Pangasius e Curimatá (Associação Brasileira da Piscicultura, 2023). Já em relação às exportações, no ano de 2022, os maiores destinos foram, Estados Unidos com 81%, Canadá 5% e México 2%, tendo um crescimento de 15%, sendo a Tilápia a espécie mais exportada (Associação Brasileira da Piscicultura, 2023).

Além da grande importância econômica, os peixes são excelentes indicadores do estado do ambiente, pois podem mostrar perturbações de diferentes escalas devido à sua mobilidade, sua proximidade ao topo da cadeia alimentar e seu estilo de vida (FREITAS et al, 2009). Algumas espécies de peixes são sensíveis a mudanças nas propriedades químicas e físicas da água, que podem ser causadas por perturbações ambientais, como agitação do sedimento de fundo de lagos, contaminação por vários tipos de poluentes orgânicos ou inorgânicos, variações naturais e outros fatores (SOUZA, F. et al, 2008).

Considerando o cenário apresentado, a pesquisa genética em peixes pode fornecer uma vantagem tecnológica na produtividade e na proteção contra patógenos.

Em particular, os elementos transponíveis (TEs) ou elementos móveis são um dos campos de estudo do genoma que influenciam toda a sua organização. Esses elementos, que podem se mover ou ser copiados dentro do genoma, podem contribuir para a compreensão da organização e manutenção da estrutura genômica e também servir como ferramentas auxiliares nos estudos da evolução genética dos organismos (BIÉMONT et al, 2006).

No entanto, exceto pelos genomas modelos, como o *Danio rerio*, as informações sobre TEs em peixes são pouco exploradas. Por exemplo, o FISHTEDB (SHAO, F. et al, 2018), o banco de dados referência de TEs em peixes, contém apenas 63 genomas disponíveis, enquanto existem 168 genomas completamente sequenciados nos bancos de dados primários de sequenciamento genômico.

Considerando a deficiência na anotação de elementos transponíveis em peixes, este projeto realizou a anotação em 168 genomas completamente sequenciados e disponíveis em bancos públicos. Ademais, contribuiu para a definição de um workflow com diferentes estratégias de anotação de TEs, possibilitando maior evidência computacional da informação anotada como sendo um TE. Por fim, acredita-se que a disponibilização das informações de TEs nos 168 genomas de peixes irá ampliar a discussão e os estudos acerca do papel funcional desses elementos e, consequentemente, contribuir para o aprimoramento das pesquisas genéticas em aquicultura.

1.2 Elementos transponíveis

Os elementos transponíveis (TEs), também chamados de DNA saltantes ou elementos parasitas (MARTINS, 2007), foram descobertos graças à pesquisa de Barbara McClintock. Em seu trabalho, analisando a diferenciação de cores nos grãos da espécie *Zea mays* (milho indiano), observou-se que essa diferenciação ocorria por conta de elementos controladores que se moviam dentro do cromossomo, alterando a atividade dos genes. Essa transposição se dava por conta de dois fatores, sendo: Ds - dissociador, presente no local da quebra, e Ac- ativador, onde era necessário para ocorrer essa quebra e a transposição dos elementos. A suspeita que seria elementos transponíveis, se deu pois Bárbara não conseguia mapear no cromossomo o fator Ac, em algumas plantas foi mapeado em uma localização, e em plantas da mesma linhagem era mapeado em outra posição do cromossomo (GRIFFITHS et al, 2013). Na época sua pesquisa e descoberta não foi aceita, pois já havia estudos onde se comprovaram que os genes eram dispostos de maneira fixa, sendo impossível

haver essa locomoção. Apenas quarenta anos mais tarde, com uma pesquisa de *E. coli*, onde encontraram um elemento de inserção em que se movia dentro do genoma, foi que a pesquisa de Barbara McClintock foi aceita. Em 1983 ela ganhou o prêmio Nobel de Fisiologia ou Medicina por sua descoberta (GRIFFITHS et al, 2013; MLA, 1983).

Os elementos transponíveis são sequências genômicas capazes de mobilização e replicação, sendo componentes abundantes e antigos dos genomas (CARARETO et al, 2015). Eles são classificados em elementos de classe I e classe II de acordo com a presença ou ausência de um intermediário de transposição. Dentro de cada classe, os TEs são ainda subdivididos em ordens, com base em seu mecanismo de inserção, estrutura e proteínas codificadas; em superfamílias, com base em sua estratégia de replicação; e em famílias, com base na conservação da sequência, como mostra a figura 1 (WICKER et al . 2007 ; KAPITONOV & JURKA 2008). Sendo classificados ainda em elementos autônomos e não autônomos. Os autônomos codificam todos os elementos necessários para sua transposição. Os não autônomos necessitam dos elementos autônomos para poder se transpor no genoma, pois não codificam as enzimas necessárias para se mover sozinhos (WICKER et al . 2007).

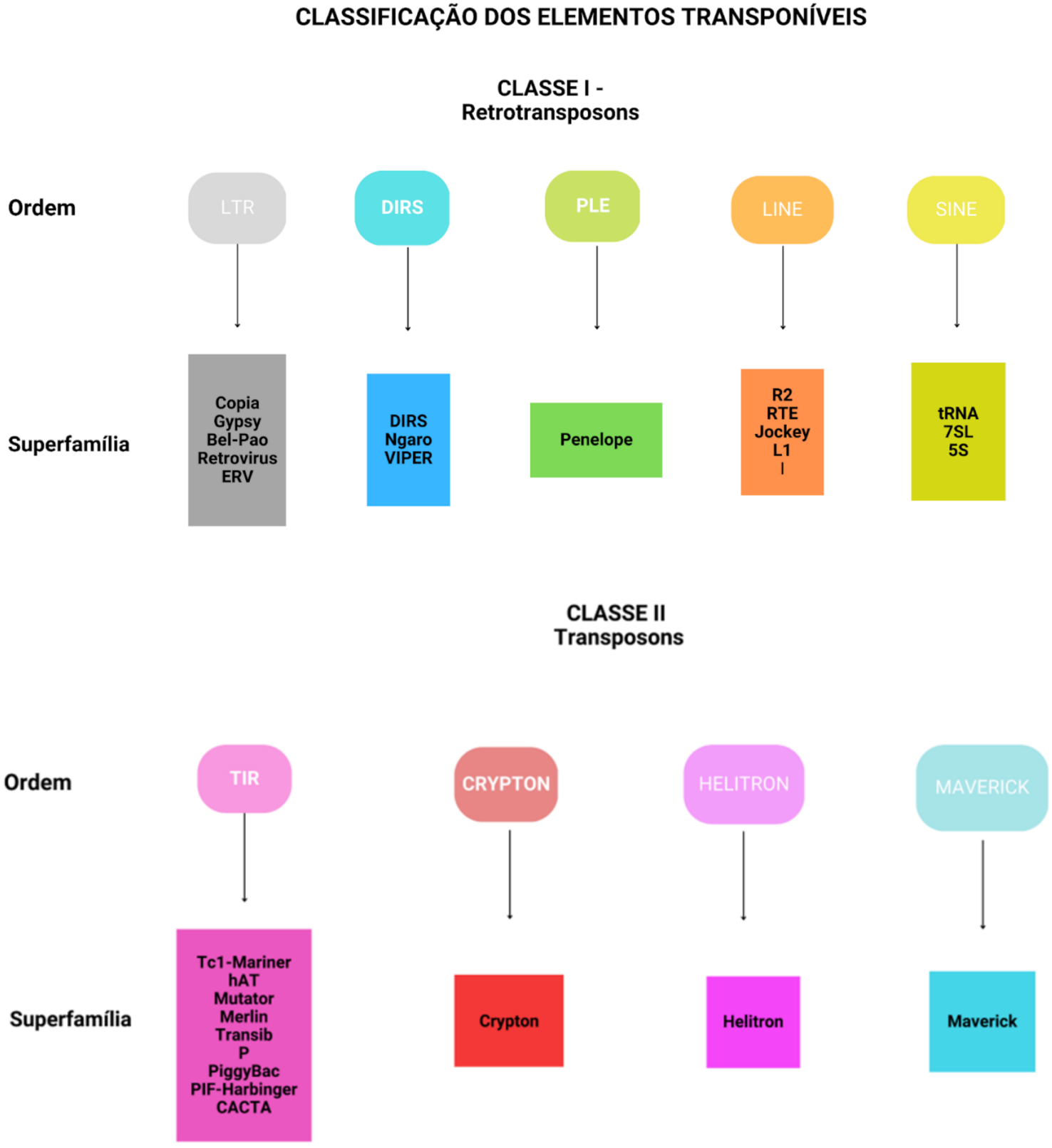


Figura 1: Classificação dos elementos transponíveis propostos por Wicker (WICKER, T. et al,2007), mostrando sua classificação de ordem e superfamília.

Fonte: Autoria própria.

Os elementos transponíveis pertencentes a Classe I são conhecidos como retrotransposons. A transposição desses elementos ocorre através da síntese de uma molécula de RNA mensageiro (mRNA), que é direcionada para o citoplasma, onde é produzido enzimas relacionadas à transposição desses elementos. Uma das enzimas produzidas é a transcriptase reversa, responsável pela síntese de uma nova cópia em DNA do retroelemento. A inserção do retroelemento no sítio receptor é realizada pelas enzimas do próprio retroelemento, e esse mecanismo é conhecido como “copia e cola” (CARARETO et al, 2015). A Classe I é dividida pelas ordens: retrotransposons com LTR, sem LTR (LINE e SINE) , DIRs e Penelope-like (PLE).

Os elementos da Classe II, conhecidos como transposons ou transposons de DNA, realizam a sua mobilização por meio de um intermediário de DNA. O processo de transposição nessa classe consiste na síntese de enzimas necessárias para a mobilização do transposon. A Classe II é dividida em duas subclasses I e II, diferenciando o número de fitas da molécula de DNA que são cortadas durante a transposição. A Subclasse I realiza a transposição via mecanismo de corta-e-cola, mediado pela DNA transposase, cortando ambas as fitas na extremidade. Já a Subclasse II realiza a transposição pelo mecanismo copia-e-cola, transpondo por replicação (CRAIG, 2002; WICKER et al.,2007).

1.3 Abordagens computacionais para anotação de TEs

Com o surgimento do Projeto do Genoma Humano, em 1980, para o sequenciamento do genoma, se viu um aumento do volume de dados biológicos. Devido a isso, se viu a necessidade de desenvolver ferramentas para a análise e interpretação dessa grande quantidade e complexidade de dados. O termo bioinformática foi usado inicialmente no começo dos anos 1970, pelos cientistas Paulien Hogeweg e Ben Hesper, definindo como estudos de processos informacionais dos sistemas biológicos, sendo uma ciência interdisciplinar, abrangendo matemática, computação e biologia (COSTA, 2004; HESPER et al,1970).

Por meio dos estudos dos elementos transponíveis, vem aumentando a demanda de análises cada vez mais detalhadas para obter informações sobre sua função e influência nos genomas. Na área computacional, a identificação automatizada de repetições dispersas é complexa devido à replicação, inserção e excisão dos elementos repetidos móveis nos genomas. Isso resulta em cópias fragmentadas e alterações complexas, dificultando a identificação e a definição de famílias repetidas (PRICE et al, 2005).

Dessa forma, há um número crescente de algoritmos desenvolvidos para estudar

e obter mais informações sobre esses elementos transponíveis. As ferramentas disponíveis, como o RepeatMasker (SMIT et al, 2022), o RepeatExplorer (NOVAK et al, 2010) e o RepeatModeler (HUBLEY et al, 2019), foram desenvolvidas para auxiliar na identificação e caracterização dos elementos transponíveis, permitindo assim a descoberta de novas superfamílias e fornecendo dados para estudos sobre a evolução dos genes (LERAT, 2010). Estas ferramentas são o estado da arte e são utilizadas em anotações de TEs em qualquer genoma. Entretanto, o RepeatModeler (HUBLEY et al,2019) e o EDTA (OU et al, 2019), por exemplo, tem um alto custo computacional para ser executado com grande quantidade de genomas. Nesse sentido, abordagens ab initio ou deep learning podem ser uma solução, como o TERL (CRUZ et al,2020). Na seção 1.5 será feita uma revisão dos estudos e metodologias disponíveis em peixes utilizados pela literatura.

1.4 Justificativa

Os TEs são elementos importantes para se entender a organização e manutenção da estrutura do genoma, além de auxiliar nos estudos da evolução genética dos organismos (BIÉMONT et al, 2006). O conhecimento da composição e arquitetura do genoma é fundamental na compreensão dos processos evolutivos responsáveis nas espécies de peixes. O surgimento de tecnologias de sequenciamento de alto rendimento, fornecendo uma grande quantidade de dados genômicos, o que tem sido eficaz para obter informações sobre a evolução dos genomas de diferentes organismos, incluindo os peixes (CARDUCCI et al,2020).

A presença de elementos transponíveis em genomas de peixes contribui para definir a estrutura de seus genomas, contribuindo para o processo de adaptação em diferentes ambientes, por meio de rearranjos genômicos além de interferir no controle da atividade gênica (KOGA et al, 2016; PEREIRA et al, 2015; CANAPA et al, 2015; CARDUCCI et al, 2019).

Dessa forma, para buscar novas informações sobre a presença dos elementos transponíveis nos genomas das espécies estudadas, e contribuir com o conhecimento acerca desses elementos, o presente trabalho propôs a identificação e análise das sequências repetitivas em espécies de peixes, assim desenvolvendo um *workflow* para anotação de TEs, com abordagens diferentes, e a construção de um banco de dados curado, assim contribuindo para os estudos de elementos transponíveis.

1.5 Revisão de literatura

A revisão de literatura foi feita na base de dados PubMed. Com a restrição

desses termos no Título e Resumo, os termos para a busca foram usados ““**transposable elements”[Title/Abstract] OR "mobile elements”[Title/Abstract] and fish**”. Em seguida, foram aplicados filtros de seleção nos resultados obtidos, tendo como critérios: papers em inglês; pesquisas relacionadas com peixes, referentes a elementos transponíveis.

Com os artigos selecionados, utilizamos o pacote *Bibliometrix R* (ARIA, M. & CUCCURULLO, C., 2017), por meio do programa *RStudio* (Posit team, 2022. v.2022.12.0), para investigar informações relevantes nos artigos filtrados.

Como resultado da análise de literatura obtivemos 423 artigos da base Pubmed que foram filtrados, totalizando ao final 93 artigos para a revisão de literatura (Figura 2).

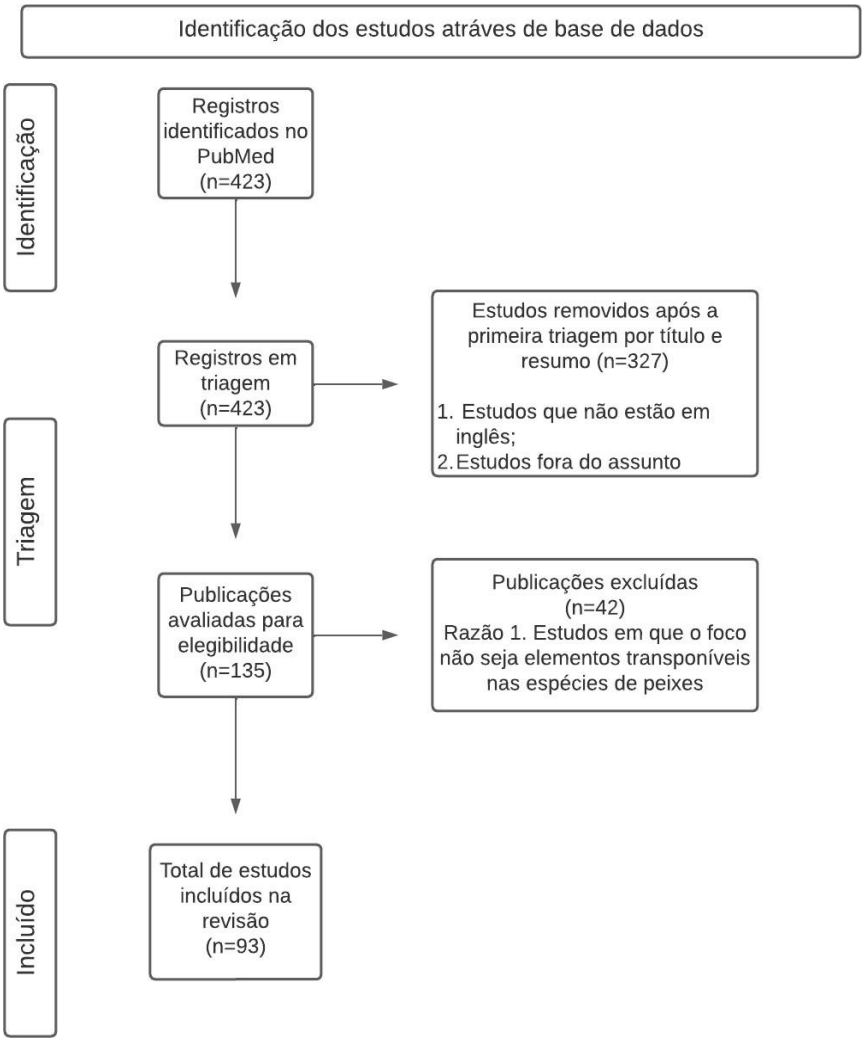


Figura 2: Fluxograma da estratégia para a revisão bibliográfica sistemática. Primeiramente identificamos 423 artigos na base de dados do Pubmed. Na triagem, removemos 327 artigos, onde estavam fora do tema procurado, e não publicados em inglês. Logo após removemos 42 artigos onde o foco não era elementos transponíveis nas espécies de peixes. Totalizando ao final 93 artigos presentes nesta revisão.

A partir dos dados geramos a tabela S1, disponível no material suplementar, que apresenta os objetivos, resultados e resumo de cada artigo selecionado. Para uma melhor observação das pesquisas relacionadas a esse elementos em espécies de peixes, montamos uma linha do tempo, (Figura 3, 4 e 5) contendo as pesquisas e suas respectivas espécies estudadas neste tema.

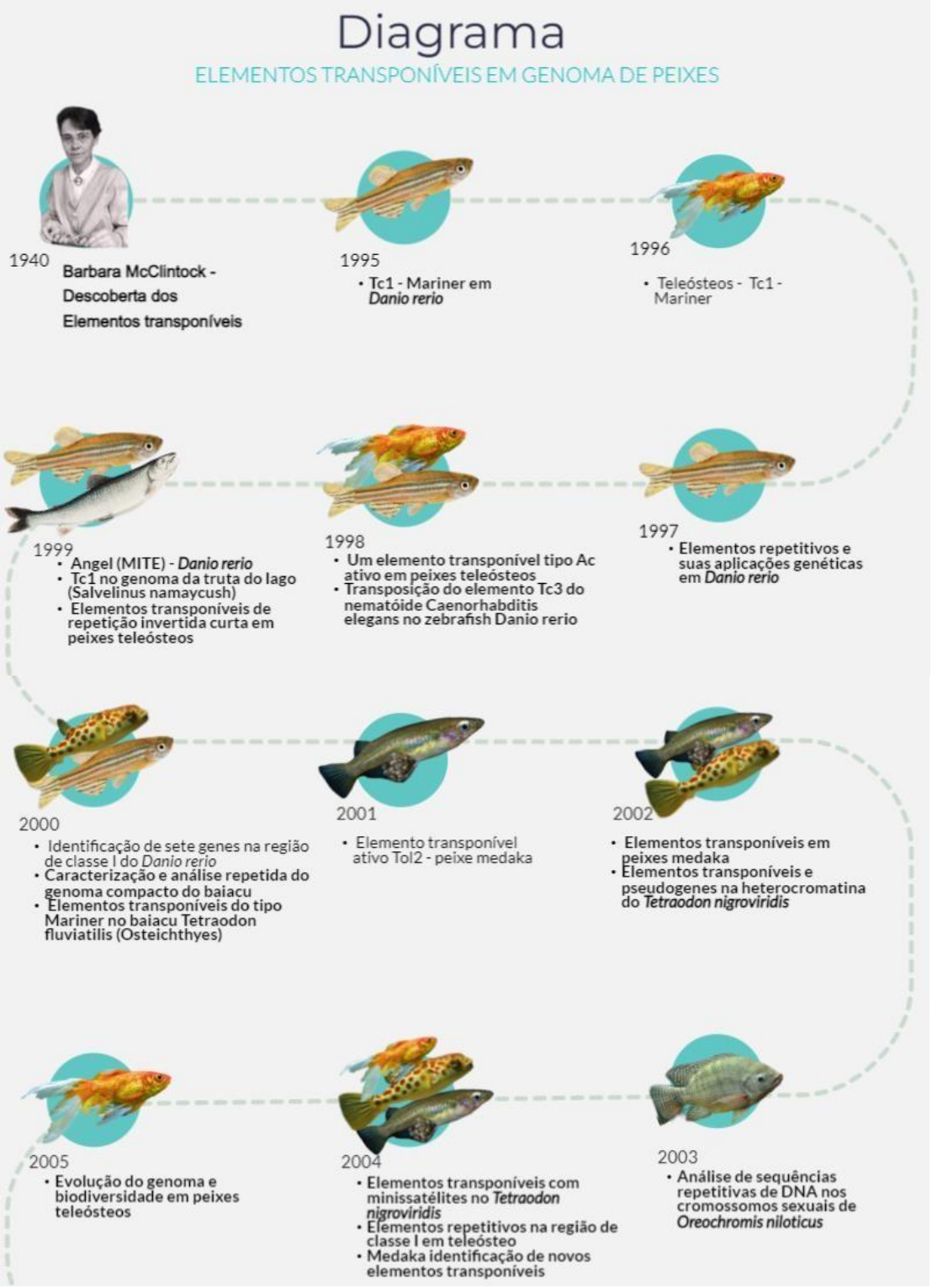


Figura 3: Diagrama temporal dos principais trabalhos de TEs em peixes. Diagrama linha do tempo com os papers sobre elementos transponíveis em espécies de peixes. Começando em 1940 com a descoberta desses elementos pela Bárbara McClintock até as pesquisas de 2003.

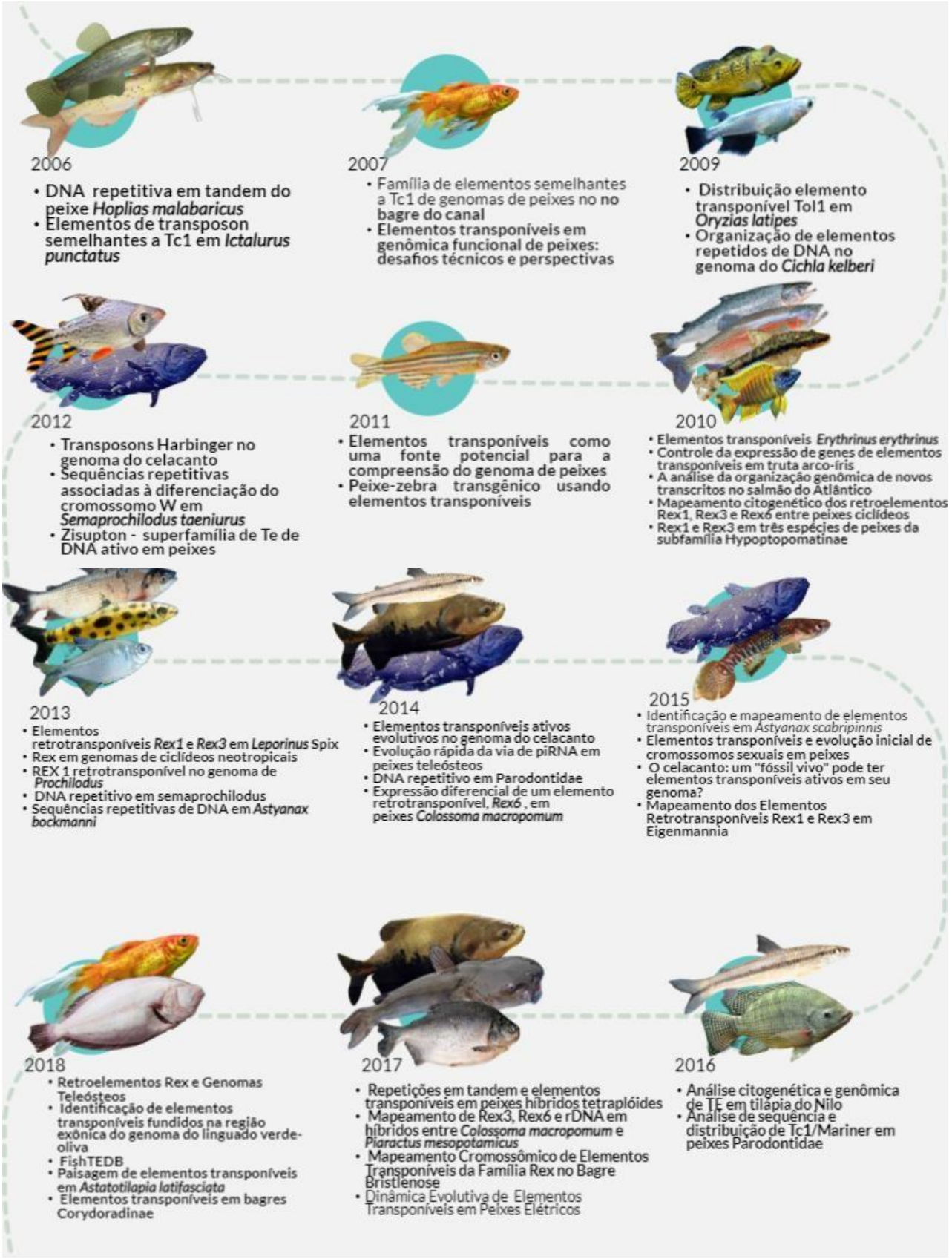


Figura 4: Diagrama linha do tempo com os papers sobre elementos transponíveis em espécies de peixes. Principais trabalhos entre os anos de 2006 - 2016.

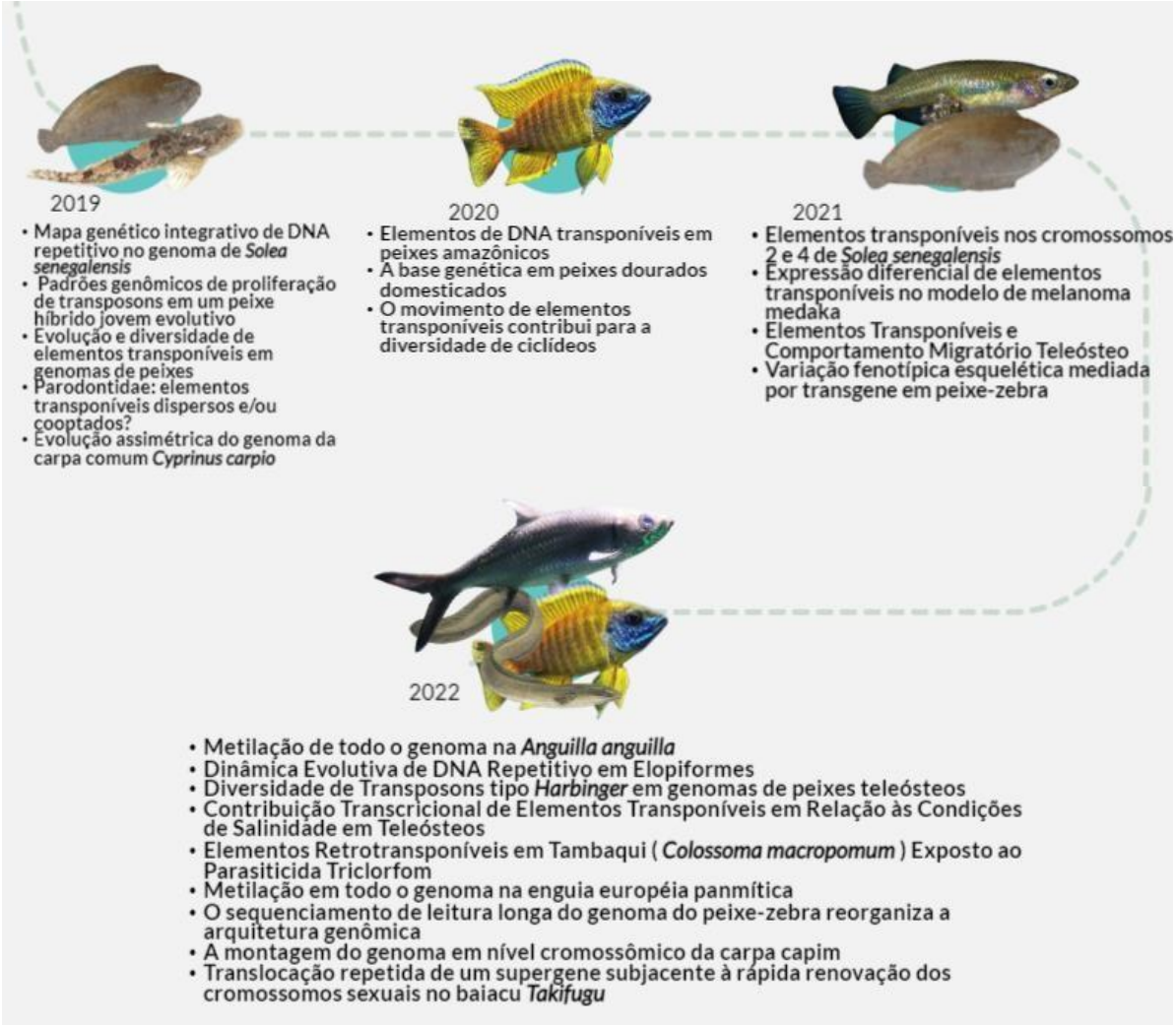


Figura 5: Diagrama linha do tempo com os papers sobre elementos transponíveis em espécies de peixes. Principais trabalhos entre os anos de 2019 - 2022.

A partir dos 93 artigos filtrados, utilizamos o programa *Bibliometrix* para buscar termos relevantes. Utilizamos a opção “Produção científica” para verificar o número de artigos publicados anualmente (Figura 6). Como esperado, com o passar dos anos, o número de publicações de artigos com o tema pesquisado foi aumentando. Mesmo havendo uma queda drástica no ano de 2016, em 2017 houve um aumento considerável. Há também uma queda razoável nos anos de 2019 e 2020, podendo ser causada pela pandemia.

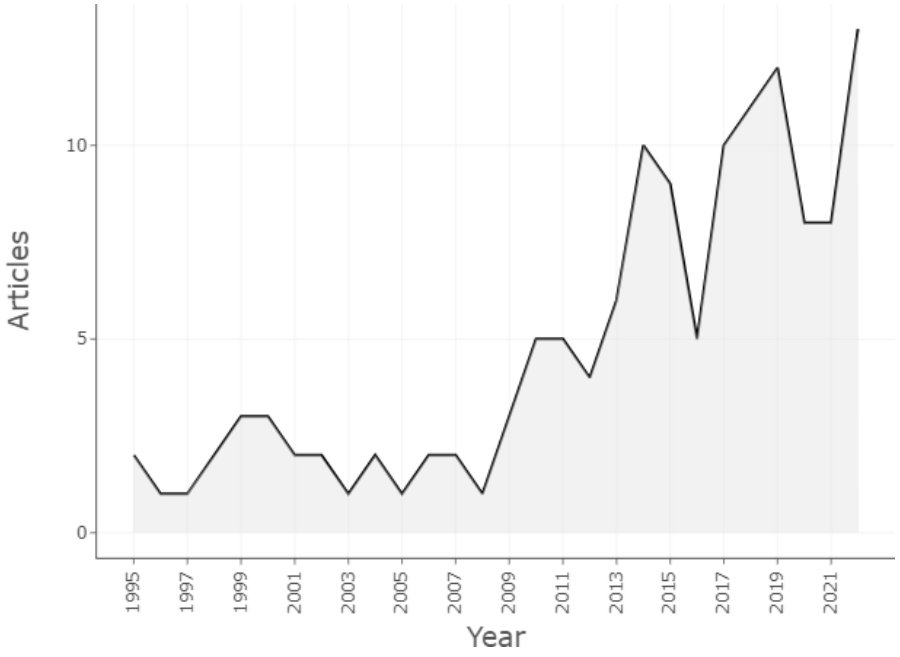


Figura 6: Gráfico dos estudos relativos às TEs em peixes nos períodos de 1995-2022. Apresentado pelo eixo x os anos de publicações dos artigos, e o eixo y a quantidade desses trabalhos.

E por fim, usamos a opção “Autores mais relevantes”, que apresenta os autores que tiveram mais artigos publicados (Figura 7). Apresentado como o autor que obteve mais publicação, Volff J. N., é um pesquisador francês que estuda principalmente as áreas de genética, genoma, gene, vertebrados e retrotransposon. Vale a pena ressaltar que os autores Fausto Foresti (3° lugar), César Martins (4° lugar), Roberto Ferreira Artoni (5° lugar) e Marcelo Ricardo Vicari (6° lugar) são pesquisadores brasileiros, os dois primeiros trabalhando na UNESP e os dois últimos na UEPG.

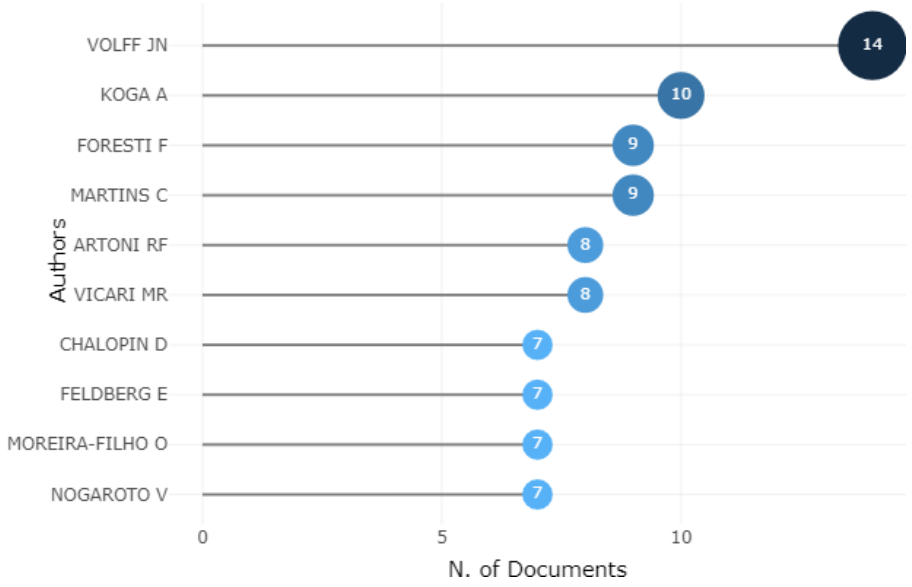


Figura 7: Relação dos autores mais relevantes sobre os estudos envolvendo TEs e peixes. O gráfico apresenta os 10 autores com maiores números de publicações sobre o assunto. No eixo x, a quantidade de artigos, e no eixo y o nome dos autores.

Com o texto de busca dos artigos, encontramos o banco de dados FISHTEDB (SHAO, F. et al, 2018), um banco de dados de elementos transponíveis, que conta com 38 espécies de peixes. Porém, este banco de dados, classificou apenas 60% das sequências nas superfamílias destas espécies, ainda existindo muitos TEs não classificados. Tendo em vista o FISHTEDB, buscamos a criação de um novo banco de dados mais completo através dessa nova pesquisa, com isso, em nosso trabalho, acrescentamos as 38 espécies como aperfeiçoamento do mesmo. Identificamos o artigo HILSDORF, A. W. S. et al,2021, no qual o autor fez a montagem do genoma do *Colossoma macropomum*, Tambaqui, uma espécie nativa do Brasil. Com o resultado do trabalho sendo, uma cobertura de 52,49% por elementos transponíveis, das quais 49,78% são repetições intercaladas.

1.6 Objetivos

Anotação em larga escala de elementos transponíveis nos genomas públicos de 168 peixes por meio da utilização de métodos computacionais e apresentar um estudo sobre os TEs em espécies de peixes.

Objetivos específicos

- Revisão do estado da arte, dos dados e bancos de TEs em peixes.
- Coletar e pré-processar as sequências genômicas das espécies estudadas.
- Criação de um *workflow* para anotação dos elementos transponíveis.
- Anotar os elementos transponíveis em 168 genomas completos e disponíveis em bancos públicos.
- Analisar e interpretar os resultados da anotação em larga escala.

Capítulo 2

ANÁLISE EM LARGA ESCALA DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS DE PEIXES

Lorena Maria Rudnik [1], Laurival Vilas-Boas [2] , Roberto Ferreira Artoni [3], Pedro Gabriel Nachtigall [4], Liliane Santana Oliveira[1*], Alexandre Rossi Paschoal [1*]

[1] Departamento de Computação (DACOM), Programa de Pós-Graduação em Bioinformática (PPGABIOINFO), Universidade Tecnológica Federal do Paraná (UTFPR) - Câmpus Cornélio Procópio, Brasil.

[2] Departamento de Biologia Geral - CCB Universidade Estadual de Londrina, Programa de Pós-Graduação em Bioinformática (PPGABIOINFO), Universidade Tecnológica Federal do Paraná (UTFPR) - Câmpus Cornélio Procópio, Brasil.

[3] Universidade Estadual de Ponta Grossa (UEPG), Ponta Grossa, Paraná, Brasil

[4] Universidade Estadual Paulista (UNESP), São Paulo, Brasil

* Autor correspondente: paschoal@utfpr.edu.br; liliane.sntn@gmail.com

RESUMO

Sequências repetitivas, como os elementos transponíveis (TEs), têm sido associadas a variabilidade genética ou em alterações fenotípicas e vão desde o papel na estrutura dos cromossomos até o mecanismo de manutenção dos telômeros e centrômeros. Ainda é incipiente a pesquisa nas espécies de peixes, por exemplo, o FishTEDB é o único repositório central de TEs em peixes e contém apenas 38 das 168 espécies com genoma completo disponíveis em banco de dados primários. Assim, apenas 14,6% dessas espécies de peixes têm dados disponíveis para estudos dos TEs. Diante disso, o objetivo deste trabalho foi realizar uma análise de larga escala dos elementos transponíveis em 168 genomas completos de peixes disponíveis em bancos públicos. Para tanto, foi criado um fluxo de trabalho para a anotação que envolveu quatro etapas com cinco estratégias de anotação. Neste artigo em particular, serão apresentados os resultados da ferramenta TERL, que foi usada para classificação dos elementos transponíveis nos dados de transcritos de 142 espécies de peixes. E as ferramentas RepeatModeler2 e EDTA, classificando TEs em 37 peixes. Realizamos a comparação dos resultados com bancos de dados já existentes, observando um resultado diferente com mais evidência de TEs pela nossa estratégia.

Palavras-chave: Bioinformática. Elementos transponíveis. Bancos de dados. Modelagem. Análise em Larga-escala.

1 INTRODUÇÃO

Os peixes desempenham papéis estruturais e funcionais vitais no ecossistema aquático, com ampla variação temporal e escala geográfica, apresentam grandes níveis de biodiversidade em termos de morfologia, comportamental, ecologia, entre outros (SIMONETTI et al, 2014). As espécies de peixes têm sido utilizadas como excelentes modelos em estudos de desenvolvimento, fisiologia, comportamento, toxicologia, evolução genética, entre outras (SIMONETTI et al, 2014). Atualmente, são conhecidas mais de 35.000 espécies de peixes segundo a base de dados Fishbase (versão 08/2022; FROESE, R. et al,2022).

Domitille e colaboradores (2013) observaram a composição e atividade de TEs, em particular, os genomas de peixes teleósteos são dominados por transposons de DNA e contêm poucas cópias antigas de TEs, sugerindo que esses elementos tiveram um papel eficaz na evolução da espécie e contribuído para a diversidade genética. Alguns estudos, como Volff et al 2003, sugerem a existência de significativa diferença no conteúdo de TEs entre sub linhagens de vertebrados, e que os genomas de peixes teleósteos apresentam uma maior diversidade de TEs quando comparado com genomas de outros vertebrados. Os peixes baiacu (*Tetraodon nigroviridis*) exibem uma diversidade de TE muito maior e com mais famílias de TE ativas do que os humanos, apesar de terem genomas com apenas um décimo do tamanho. (VOLFF et al, 2003).

Feng (SHAO F. et al, 2019) notaram que TEs de peixes estão distribuídos regularmente, espalhados de forma relativamente uniforme por todo o genoma dos peixes estudados. Tal fato indica que os TEs desempenham um papel vital na evolução dos peixes. Por meio de análises de correlação adicionais, confirmou-se também que o efeito dos retrotransposons no tamanho do genoma foi maior do que o dos transposons de DNA, sendo os LTRs estarem significativamente correlacionados com o tamanho do genomas.

Mesmo com toda essa importância para estudos de TEs em espécies de peixes, ainda são poucos os dados disponíveis publicamente. Como é o caso de bancos de dados, hoje existe apenas o FISHTEDB, que se delimita a 63 espécies de peixes. Ou o Dfam, que há apenas informações desses elementos presentes na espécie *Danio rerio*. Portanto, este estudo teve como objetivo a anotação em larga escala dos TEs no genoma de 168 espécies de peixes disponíveis em banco de dados públicos. Desenvolvemos um workflow com estratégias diferentes da anotação do TE,

com a proposta de dar confiabilidade nesta anotação *in silico* destes TEs.

2 METODOLOGIA

A metodologia aplicada neste trabalho envolveu duas etapas. Na primeira, um levantamento da literatura foi feito para atualização sobre o estado da arte, bem como a pesquisa dos dados públicos sobre elementos transponíveis disponíveis na literatura científica. A segunda, foi criado e implementado um *workflow* para anotação de TEs em genomas completos de peixes públicos disponíveis.

2.2 Anotação de elementos transponíveis em larga escala em peixes

Desenvolvemos um fluxo de trabalho para a análise em larga escala dos elementos transponíveis por meio de várias ferramentas explicadas a seguir. Para a classificação de TEs, selecionamos as sequências de genoma e transcriptoma das espécies de peixes disponíveis em bancos de dados públicos (Ensembl, NCBI e CNGB) e aplicamos três estratégias, todas demonstradas na Figura 8: (A) anotação em dados de transcritos via a ferramenta TERL (CRUZ, M. H. et al, 2020, v.1); (B) anotação via genomas de peixes representantes de toda a população analisada via RepeatModeler2 (HUBLEY, R. et al, 2019, v.2.0.3 e v.2.0.4) e EDTA (OU, S. et al, 2019, v. 1.9.6); e (C) via os dados públicos da literatura. Em seguida, nos passos (D) e (E) esses dados serviram para construção de uma biblioteca não redundante de TEs obtidos em genomas de peixes, seguido de análise em larga escala via programa RepeatMasker (SMIT, A., et al, 2022, v. 4.1.4). Por fim, como resultado será apresentado um banco de dados de elementos transponíveis de ocorrência em genomas de espécies de peixes.

Para a explicação mais detalhada do fluxo de trabalho, separamos as fases por seções, apresentadas abaixo.

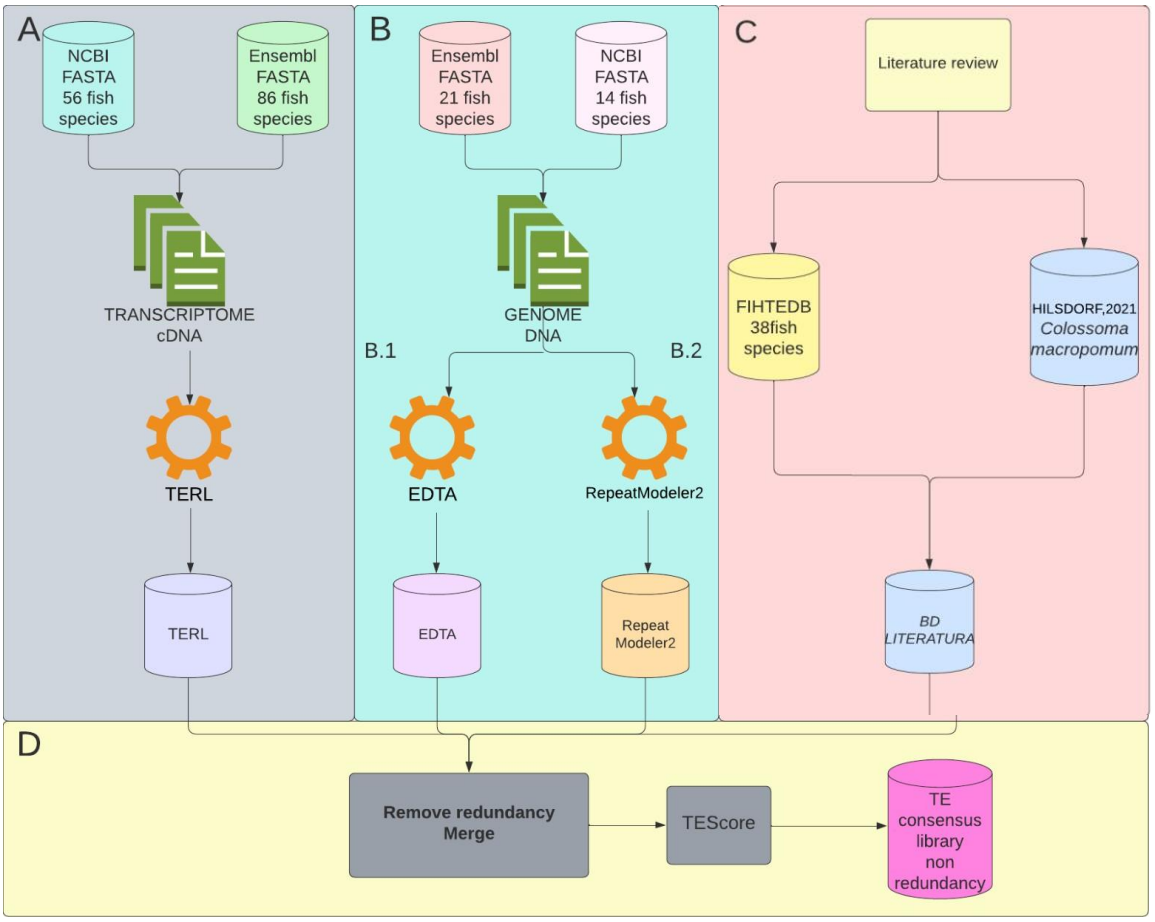


Figura 8: Visão geral da análise em larga escala dos elementos transponíveis. A) Utilização da ferramenta TERL, para a identificação de TEs nos transcriptomas. B) Utilização das ferramentas B.1 - EDTA e B.2 - RepeatModeler2 para classificar sequências do genoma de 35 peixes escolhidos, gerando um banco de dados para cada ferramenta. C) Banco de dados disponíveis publicamente em foco nas espécies de peixes presentes na literatura. D) União das sequências obtidas nas etapas anteriores e remoção das sequências redundantes, após realizada a implementação do TEScore e assim formando um banco de dados de elementos transponíveis consenso e sem redundância.

2.2.1 Conjunto de dados coletados e utilizados

Utilizou-se dados de genoma (sequência de DNA), de transcritos (sequência de cDNA). Foram também analisados dados de transcriptoma obtidos via RNA-Seq de apenas o Tambacu por não haver genoma disponível.

Os dados das sequências de DNA das 37 espécies de peixes utilizadas, foram obtidas em bancos de dados públicos, sendo: NCBI - *National Center for Biotechnology Information* (https://www.ncbi.nlm.nih.gov/. PEPER, C., 1988) de onde se obteve 14 sequências genômicas das espécies dos peixes; Ensembl

(http://www.ensembl.org/. *European Bioinformatics Institute* e *Wellcome Trust Sanger Institute*, 1999) de onde foram obtidas 23 sequências genômicas. Já no caso dos transcriptomas, foram coletadas sequências de 142 espécies, sendo 56 sequências a partir do NCBI e 86 conjuntos de sequências a partir do Ensembl. As informações dos peixes selecionadas, totalizando 168 espécies, estão descritas na tabela S2, presente no material suplementar, sendo apresentadas a ordem, família e os genomas.

2.2.2 Fase A: Criação de biblioteca de dados de TEs em transcritos

Para análise dos transcriptomas, utilizamos a ferramenta TERL (*Transposable Elements Representation Learner*. CRUZ, M. H. P. et al, 2020). Esta abordagem utiliza redes neurais convolucionais profundas para a representação dos dados de entrada e assim classificar da melhor forma os elementos transponíveis presentes na sequência. Os elementos transponíveis foram classificados de acordo com sua superfamília e ordem, utilizando como entrada o conjunto de sequências do transcriptoma de cada uma das espécies avaliadas. Como parâmetro foi selecionado o modelo de classificação DS1, treinado e disponível pela própria ferramenta, classificando em superfamílias: Copia; Gypsy; Bel Pao; ERV; L1; Mariner e hAT. Como saída da ferramenta, foram obtidos arquivos em formato de fasta.

2.2.3 Fase B: Construção de biblioteca representativa "core" de 35 espécies dos 168 peixes

Para a construção do núcleo representativo de sequências, selecionamos as espécies de acordo com critério que incluiu importância econômica, distribuição geográfica, representatividade de diferentes grupos taxonômicos e importância em estudos, totalizando 37 espécies, sendo descritas separadamente na tabela 1. Foi feita a filtragem de espécies, por conta do tempo de processamento das ferramentas utilizadas. Os dados obtidos foram usados para criação de uma biblioteca onde pudéssemos determinar as distribuições dos TEs encontrados nestas espécies, investigando assim, o panorama desses elementos.

ESPÉCIES	MOTIVO DA ESCOLHA
Amblyraja radiata	Modelo para estudos em genética de conservação
Amphiprion percula	Peixes ornamentais, têm características sexuais hermafroditas protândricas
Astyanax mexicanus	Espécie amplamente cultivada em aquários e na aquicultura, apresenta ausência parcial ou total de olhos, sendo peixes cavernícolas
Callorhinchus milii	Representante de peixes cartilaginosos, com grande importância na pesca e nos estudos de seu genoma
Carassius auratus	Espécie ornamental, um dos peixes de aquário mais comuns no mundo
Carcharodon carcharias	Importante para genética de conservação e estudos evolutivos
Colossoma macropomum	Espécie economicamente importante, utilizada na culinária, possui carne de alta qualidade, espécie nativa brasileira
Cyprinus carpio	Importância econômica, possui carne de qualidade, procurada na pesca esportiva
Danio rerio	Espécie de grande valor para pesquisa, organismo modelo para estudos genéticos e de desenvolvimento de vertebrados
Electrophorus electricus	Produz altas descargas elétricas, utilizado como peixe ornamental, consumido por algumas comunidades ribeirinhas, espécie importante para pesquisas científicas
Erpetoichthys calabaricus	O único peixe pulmonado disponível no Ensembl, similar ao spotted gar, não sofreu o round de duplicação genômica específico de teleósteos, importante modelo para estudos evolutivos
Gadus morhua	Economicamente importante, utilizado na culinária, em perigo de extinção devido à pesca e ao uso de arrastões de fundo
Ictalurus punctatus	O peixe-gato tem grande importância econômica, sendo muito utilizado na culinária
Leucoraja erinacea	Importante em estudos evolutivos
Latimeria chalumnae	Espécie de peixe de água doce, raro representante atual do grupo de peixes fósseis.
Larimichthys crocea	Espécie com muito valor econômico, sendo amplamente exportada, utilizada na culinária
Lepisosteus oculatus	Espécie de peixe teleósteo que faz parte de um clado que não passou por fenômeno de duplicação do genoma ocorrido no grupo, importante para estudos em genética.

Lophius litulon	Possui característica incomum de adaptação fenotípica para camuflagem ambiental
Lota lota	Possui importância comercial por alta qualidade nutricional de sua carne
Megalops cyprinoides	Espécies comercializadas, cultivadas em tanques e populares na pesca
Oncorhynchus mykiss	Importância comercial e parte da família Salmonidae, que passou por fenômeno específica no clado de duplicação genômica
Oreochromis niloticus	Importante na piscicultura no Brasil de no mundo, muito usado na culinária
Oryzias Javanicus	Usado para reprodução em aquários, comum em plantações de arroz
Paralichthys olivaceus	Alta qualidade nutricional de sua carne
Piaractus mesopotamicus	Amplamente utilizado na culinária, possui grande quantidade de carne de excelente qualidade, cultivada em tanques, na aquicultura, importante economicamente no Brasil
Pristis pectinata	Ilegal a sua captura, possessão ou lesão devido a estar ameaçado de extinção
Psettodes erumei	Alta qualidade nutricional de sua carne
Pygocentrus nattereri	Importante fonte de renda para criação em aquicultura, carne de excelente qualidade, muito utilizada na culinária
Rhincodon typus	Importante para estudos em genética de conservação e estudos evolutivos
Salarias fasciatus	Utilizado na criação como peixe ornamental, além de ser uma importante espécie para estudos
Salmo salar	Economicamente importante, usado na culinária
Scyliorhinus canicula	Usado como modelo em estudos evolutivos em embriologia e desenvolvimento
Tetraodon nigroviridis	Peixes ornamentais, usado em aquários
Thunnus maccoyii	Usado na culinária, economicamente importante
Toxotes chatareus	Alta qualidade nutricional de sua carne

Tabela 1 - Espécies de peixes selecionadas do conjunto inicial e algumas características gerais que contribuíram nos critérios de inclusão no grupo de análise.

A análise desses genomas foi realizada com as ferramentas RepeatModeler2 (HUBLEY, R. et al, 2019, v2.0.3 e v.2.0.4) e EDTA (OU, S. et al, 2019, v1.9.6). O RepeatModeler2 (disponível em ((http://www.repeatmasker.org/RepeatModeler/)), é uma ferramenta desenvolvida para anotação *de novo* de TE. Essa ferramenta utiliza outros dois programas, o RepeatScout (PRICE, A. L. et al, 2005) e o RECON (BAO, Z. et al, 2002), que identificam as famílias TE em uma sequência genômica. Utilizamos os parâmetros padrão, para cada espécie separadamente. Já para a classificação utilizamos os parâmetros padrão e -LTRStruct para a classificação de elementos transponíveis LTR. Já o EDTA (Extensive de-novo TE Annotator, disponível em (https://github.com/oushujun/EDTA)), tem como objetivo a construção de bibliotecas de sequências de TEs não redundantes, que podem ser subsequentemente usadas para gerar anotações de TE *de novo*. Composto por ferramentas: LTR_FINDER (XU, Z. et al,2007), LTRharvest e LTR_retriever para a descoberta de retrotransposons LTR; TIR-learner (SU, W. et al,2019) para descoberta de transposons TIRs e HelitronScanner (XIONG, W. et al,2014) para transposons Helitron, além de conter a ferramenta RepeatModeler. A escolha dessas duas ferramentas se deu pelos algoritmos de anotação específicos de algumas superfamílias de TEs, e pelo aumento desses elementos encontrados, auxiliando na escassez de anotação do TERL.

2.2.4 Fase C Criação de biblioteca de dados de TE descritos na literatura

A criação da biblioteca de dados descritos na literatura se deu por meio de duas pesquisas, sendo: FISHTEDB (SHAO, F. et al, 2018); e a pesquisa de HILSDORF, 2021. O FISHTEDB é composto por 63 espécies de peixes, como já mencionado anteriormente. Para a criação deste banco de dados, foi utilizado a ferramenta RepeatModeler e TEClass. Na pesquisa de HILSDORF e colaboradores (2021) foi realizada a montagem e anotação dos genes da espécie *Colossoma macropomum* e a anotação dos elementos transponíveis presentes na espécie, utilizando a ferramenta RepeatModeler e RepeatMasker para a classificação de TEs.

2.2.5 Fase D: Criação de uma biblioteca única, padronizada, e não-redundante de TEs em peixes

Para a retirada da redundância dos bancos de dados gerados (etapas A - B e C, figura 7), foi utilizado a ferramenta cd-hit (WEIZHONG,L. 2006, v4.8.1). Um programa projetado para agrupar sequências genômicas semelhantes em "clusters" para reduzir a redundância nos dados. Utilizando um algoritmo de agrupamento, onde as

sequências que compartilham uma porcentagem específica de identidade são agrupadas juntas.

Aplicado após TEScore, uma métrica criada pelo grupo de pesquisa da UTFPR-CP, para ponderar a confiabilidade do nosso pipeline. Sendo os valores em porcentagem da classificação feita por todas as abordagens utilizadas. Sendo a equação:

$$TEScore = QIP \div QP$$

onde, Quantidade de identificação por programa (QIP), quantidade de TEs de que cada programa encontrou sozinho, ou o que classificou igual entre as ferramentas. Quantidade de programas (QP), sendo quatro ferramentas utilizadas.

Os cálculos foram feitos por meio de um código criado, onde se coloca os valores de classificação das sequências de cada ferramenta, e os valores de intersecção dos diagramas de Venn, para no final se ter o resultado do TEScore.

Assim gerando o banco de dados final visando obter informações sobre esses elementos constando a classe, ordem e a superfamília, criando um conjunto de dados que permitirão auxiliar em futuros estudos.

3. RESULTADOS E DISCUSSÃO

Os resultados a serem apresentados serão os obtidos na Fase A (TERL) e Fase B (EDTA e RM2), e as comparações com os dados da Fase C (Literatura e FISHTEDB). Importante ressaltar que na Fase D foi feito o script para o cálculo do TEScore.

As espécies de peixes, que utilizamos para a ferramenta TERL, estão descritas na tabela S3, presente no material suplementar. Já os resultados das superfamílias presentes nas ordens selecionadas, encontradas pelas ferramentas TERL, EDTA, RepeatModeler2, pelo banco de dados FISHTEDB, e o trabalho de Hildsdorf, estão especificadas na tabela do material suplementar (S4).

3.1 Anotação de TE em peixes

A partir dos dados descritos na subseção 2.2.2, representamos os resultados por meio de um HeatMap (figura 9). Mostrando as espécies analisadas (eixo y), classificadas pelas ordens dos elementos transponíveis (eixo x), sendo LTR (*Long Terminal Direct Repeats*), LINE (*Long Interspersed Nuclear Element*) e TIR (*Terminal Inverted Repeats*). Quanto mais escura a região do mapa, maior é a quantidade de TEs que a espécie possui na ordem representada. Na ordem LTR estão incluídas as superfamílias: Copia, Gypsy, ERV e Bel Pao, enquanto que na ordem LINE está a superfamília L1 e na TIR estão hAT e Mariner.

No material suplementar, é possível acessar a figura (Figura S5) completa do heatmap contendo todas as espécies. Para uma melhor observação, filtramos o heatmap com espécies que mais nos chamaram a atenção, por apresentarem maior quantidade de TEs, e a semelhança dessa quantidade entre espécies do mesmo gênero (Figura 9).

A espécie *Danio rerio* demonstrou maior quantidade de LTR comparado às outras espécies analisadas, e uma quantidade diferente de elementos TIRs. A espécie ancestral celacanto (*Latimeria chalumnae*) possui maior quantidade de LINE comparado aos demais peixes. Observamos também no heatmap, em que espécies do mesmo gênero possuem quantidades de elementos transponíveis parecidas, como é no caso das espécies *Oncorhynchus kisutch, O. mykiss, O. tshawytscha*; *Takifugu flavidus, T. rubripes.* Podemos observar que nestas espécies não se tem muitos elementos da ordem LINE sendo bem distribuída entre as espécies, e possui mais elementos encontrados na ordem LTR, já os TIRs são bem distribuídos a quantidade.

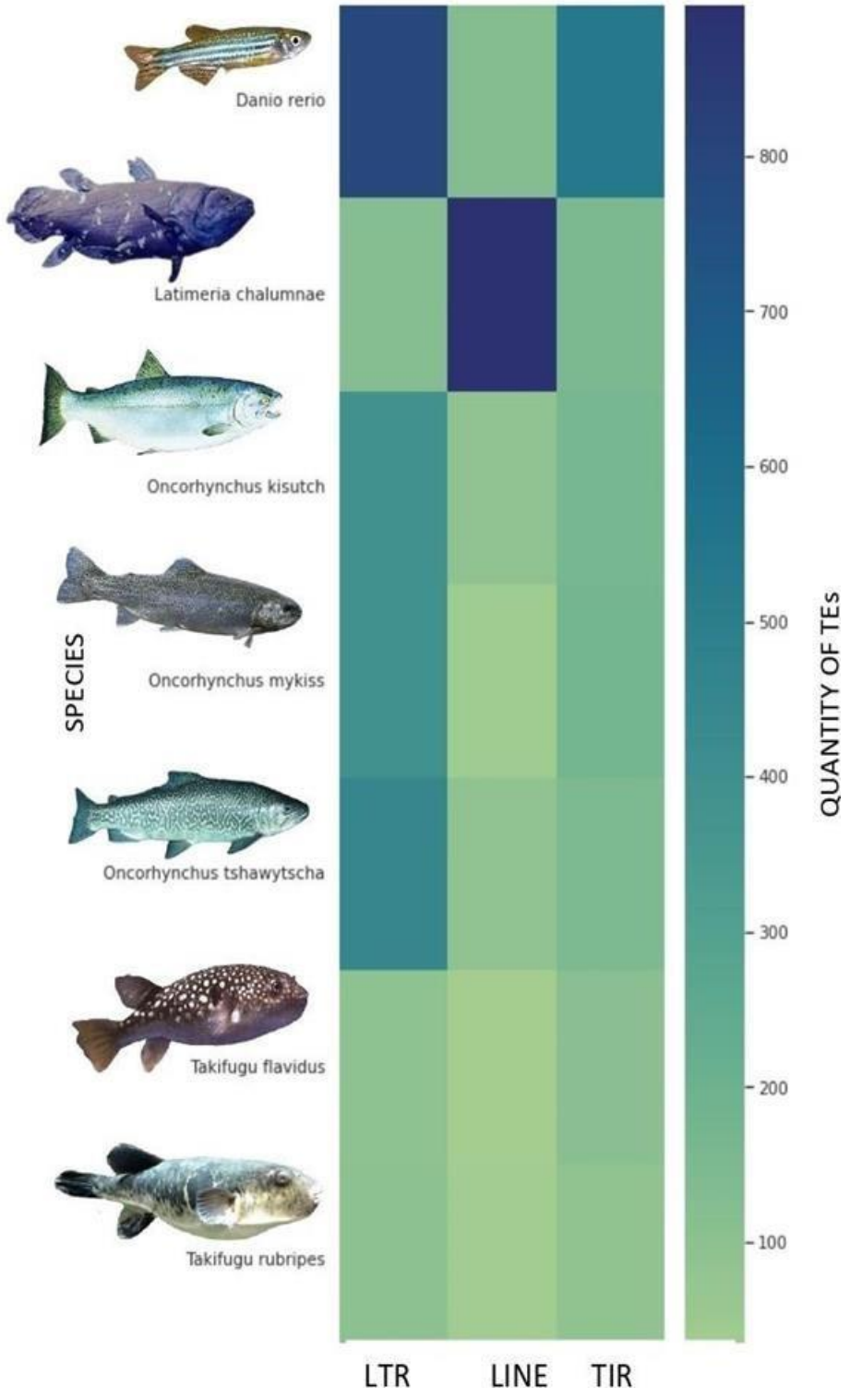
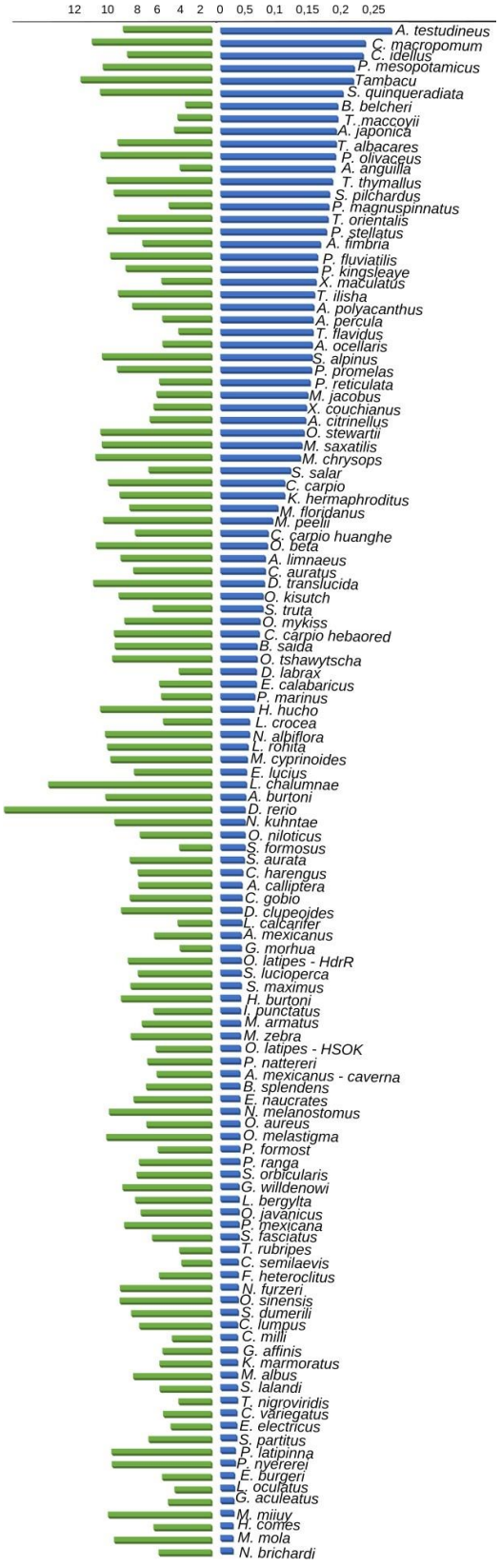


Figura 9: HeatMap - TERL. No eixo Y temos as espécies analisadas, no eixo X a classificação dos elementos transponíveis pelas ordens. Quanto mais escuro a cor, a quantidade de elementos transponíveis encontrada naquela espécie é maior, quanto mais clara, menor é esta quantidade, como demonstra a legenda da cor no lado direito da imagem.

Para verificar as anotações de TE pela ferramenta TERL, montamos o gráfico (figura 10),que apresenta o tamanho dos genomas em GB e a porcentagem da quantidade de TE. Além de realizar o gráfico de regressão linear para obter o R^2 e o p-values, com os mesmos dados. Notamos que a quantidade de TE está relacionada com o tamanho do genoma, já que o gráfico apresentado teve um R^2=0,812 e o p-values=1,537E-55, assim valores dentro do esperado, já que pode-se concluir que há pouca variabilidade e boa precisão nos dados.



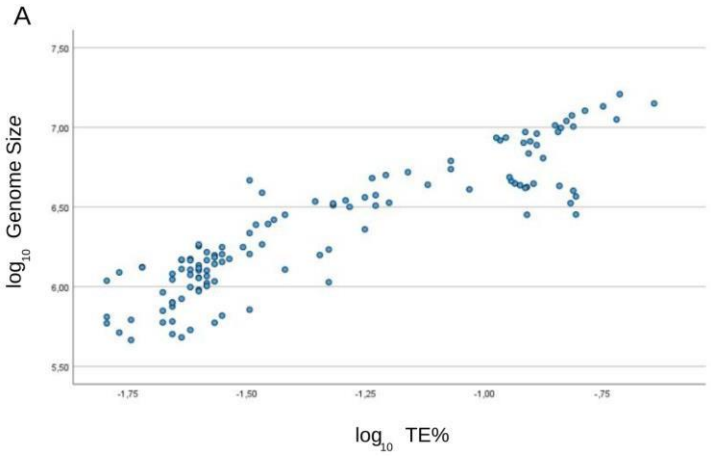


Figura 10: Relação entre o tamanho do genoma e o conteúdo de TEs. O gráfico de barras em azul representa o tamanho do genoma (em GB) e, em verde, a distribuição dos elementos transponíveis nos genomas analisados (em porcentagem). À direita (A), normalizamos o tamanho do genoma e TE usando log(10), e então relacionamos (Pearson) o tamanho do genoma por elementos transponíveis. Usando todos os 142 transcriptomas.

Com os resultados das ferramentas EDTA e RepeatModeler2 (subseção 4.2.3), representamos os dados por meio de heatmaps (figura 11 e 12). No eixo Y apresentamos as espécies de peixes, e o eixo X as ordens de TEs classificadas.

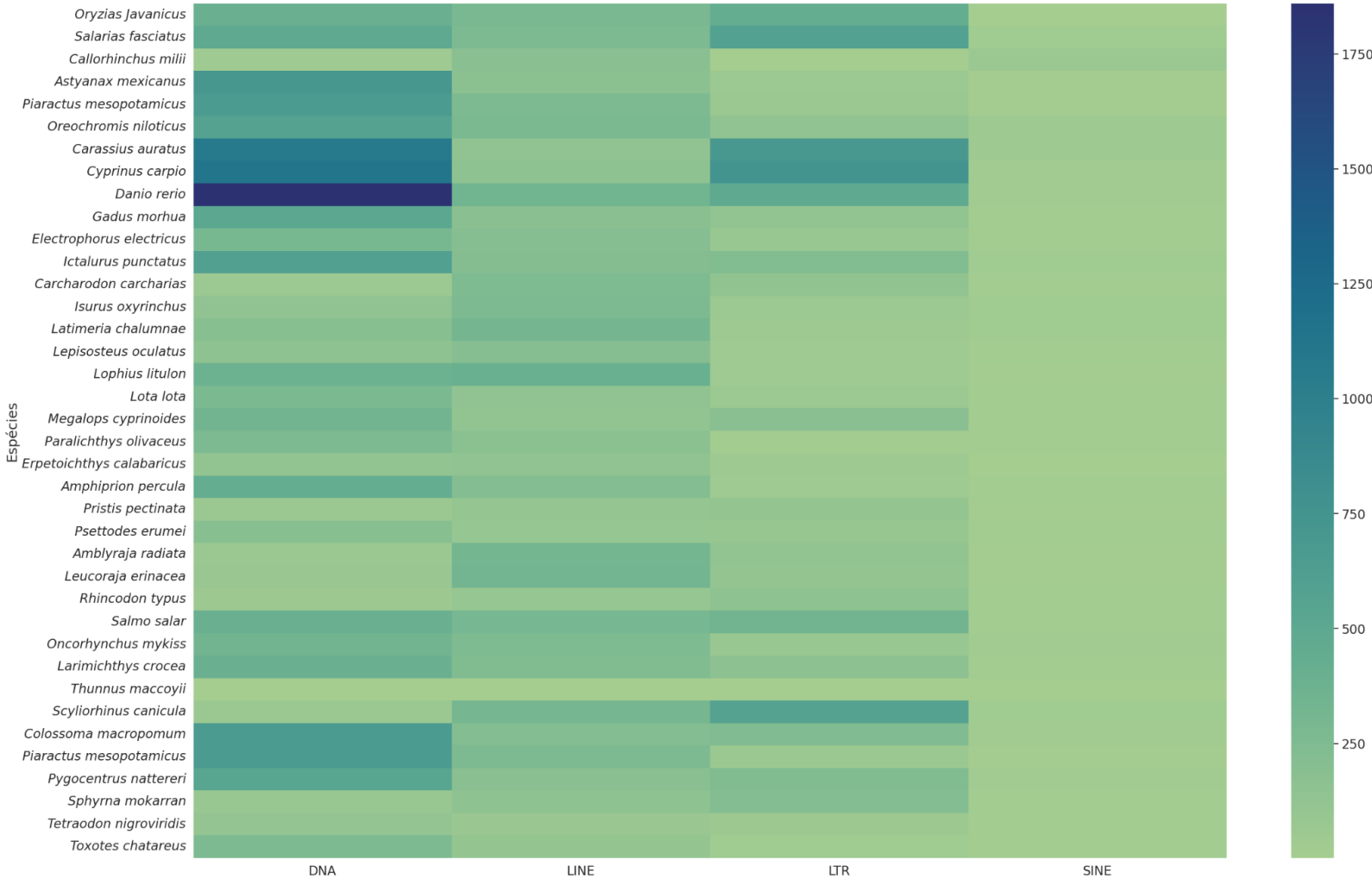


Figura 11: HeatMap - RepeatModeler2. No eixo Y temos as espécies analisadas, no eixo X a classificação dos elementos transponíveis pelas ordens (DNA-LINE-LTR-SINE). Quanto mais escuro a cor, a quantidade de elementos transponíveis encontrada naquela espécie é maior, quanto mais clara, menor é esta quantidade, como demonstra a legenda da cor no lado direito da imagem.

Demonstrado por ordem de família das espécies, observamos que as espécies obtiveram uma maior quantidade de TEs na ordem DNA, com bastante diferenciação nas cores, sendo o valor mais alto da espécie *Danio rerio* = 1860, e a espécie com menor valor *Thunnus maccoyii* = 67. Já as ordens LINE, sendo a média entre as espécies de 203 TEs e SINE com a média de 27, tiveram resultados bem parecidos entre os peixes.

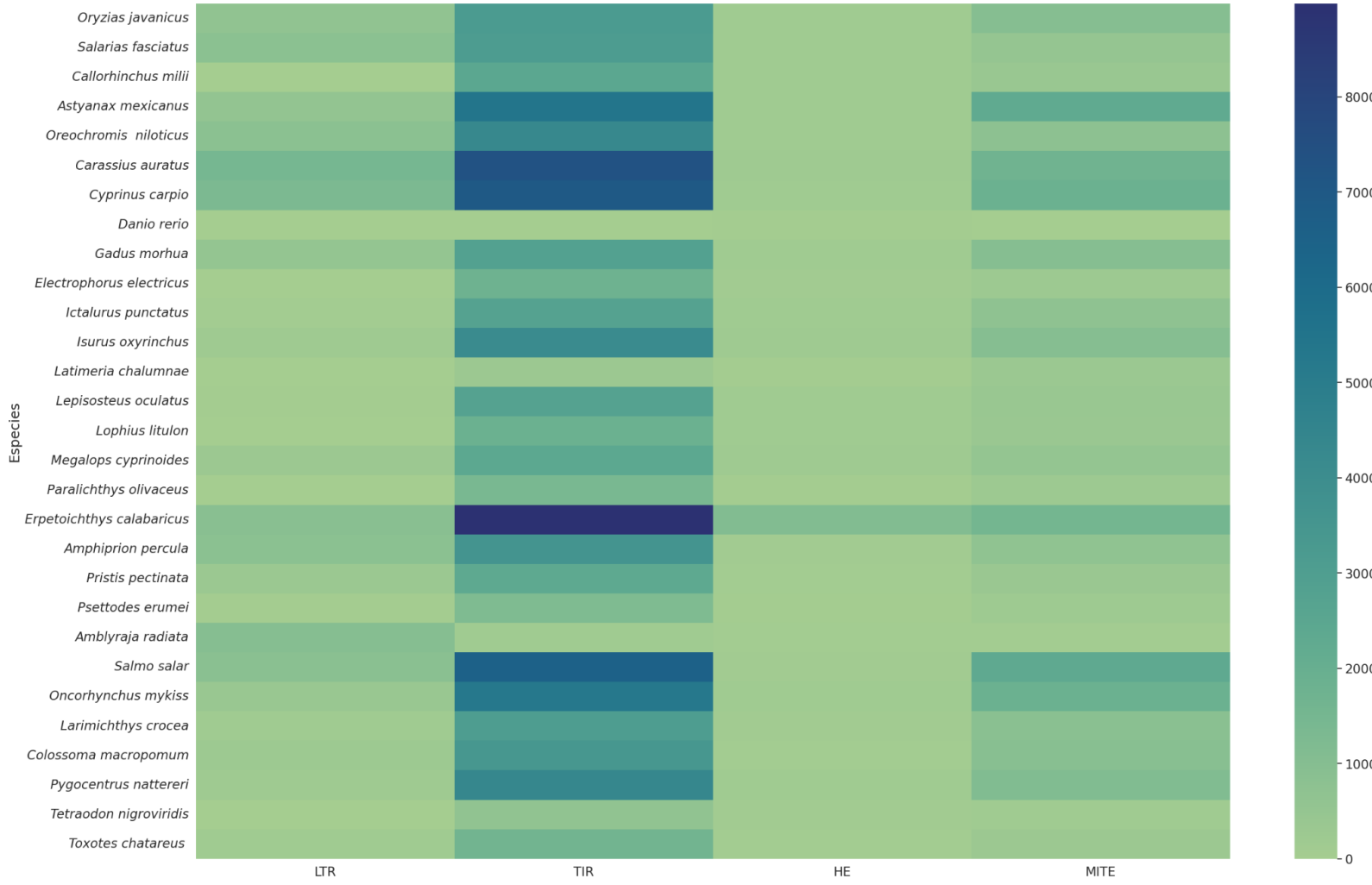


Figura 12: HeatMap - EDTA. No eixo Y temos as espécies analisadas, no eixo X a classificação dos elementos transponíveis pelas ordens (LTR-TIR-Helitron-MITE). Quanto mais escuro a cor, a quantidade de elementos transponíveis encontrada naquela espécie é maior, quanto mais clara, menor é esta quantidade, como demonstra a legenda da cor no lado direito da imagem.

Pela classificação do EDTA, vimos que a ordem TIR é a que possui maiores números, com diferenciação nas cores entre os peixes, sendo o peixe com maior valor deste elemento *Erpetoichthys calabaricus*= 8980, e com o menor valor *Danio rerio* = 13. Dentre as superfamílias presentes na ordem TIR, a CACTA possui os maiores valores. Já na ordem Helitron, observamos que se destaca a espécie *Erpetoichthys calabaricus,* por apresentar a maior quantidade de TEs entre as espécies.

3.2 Comparação entre as ferramentas e dados públicos

Realizamos a comparação dos resultados dos elementos transponíveis entre as ferramentas TERL, EDTA, RepeatModeler2, e o banco de dados FISHTEDB com as espécies de peixes que aparecem em todas as ferramentas e no banco de dados (Figura 13).

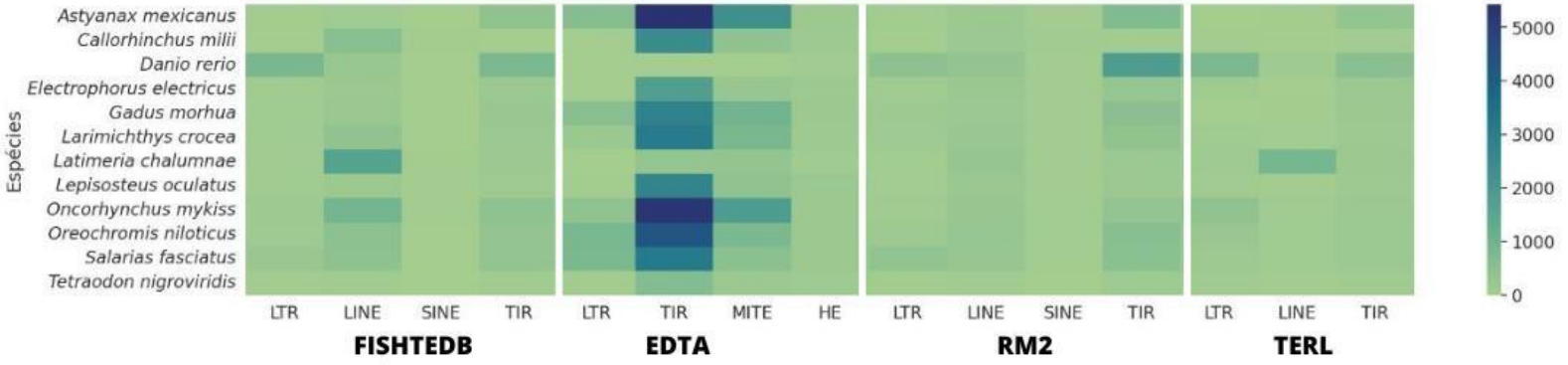


Figura 13: Heatmap - Comparação das ferramentas e o banco de dados FISHTEDB. No eixo Y temos as espécies analisadas presente entre todas as ferramentas e o banco de dados. No eixo X a classificação dos elementos transponíveis pelas ordens (LTR-LINE-SINE-TIR-MITE-HE). Quanto mais escuro a cor, a quantidade de elementos transponíveis encontrada naquela espécie é maior, quanto mais clara, menor é esta quantidade, como demonstra a legenda da cor no lado direito da imagem.

Notamos no geral, pela diferenciação de cores, que há uma diferenciação na quantidade de TEs, presente nas ordens: TIR onde a ferramenta EDTA classificou muito mais; e LINE onde o banco de dados FISHTEDB possui quantidade maior. As demais observamos que possui uma quantidade similar entre elas. Para verificar melhor essa diferenciação entre eles, selecionamos seis espécies de peixes, para realizar uma análise mais profunda, sendo: *Danio rerio*, *Latimeria chalumnae*, *Colossoma macropomum*, *Gadus morhua*, *Astyanax mexicanus* e *Oreochromis niloticus*.

3.2.1 Danio rerio

Para a primeira análise de comparação, fizemos um heatmap com as classificações por ordem das ferramentas e do FISHTEDB (Figura 14).

Notamos que há uma diferenciação perceptível na ordem TIR pela ferramenta RepeatModeler2, classificando mais que as demais. Observamos também que os resultados do TERL com o banco de dados FISHTEDB são parecidos. Porém, a ordem LINE possui uma diferenciação entre eles, o que não ocorre, se compararmos o FISHTEDB com o RepeatModeler2. O EDTA possui resultados distintos, mas classificou mais elementos da ordem Helitron do que os demais.

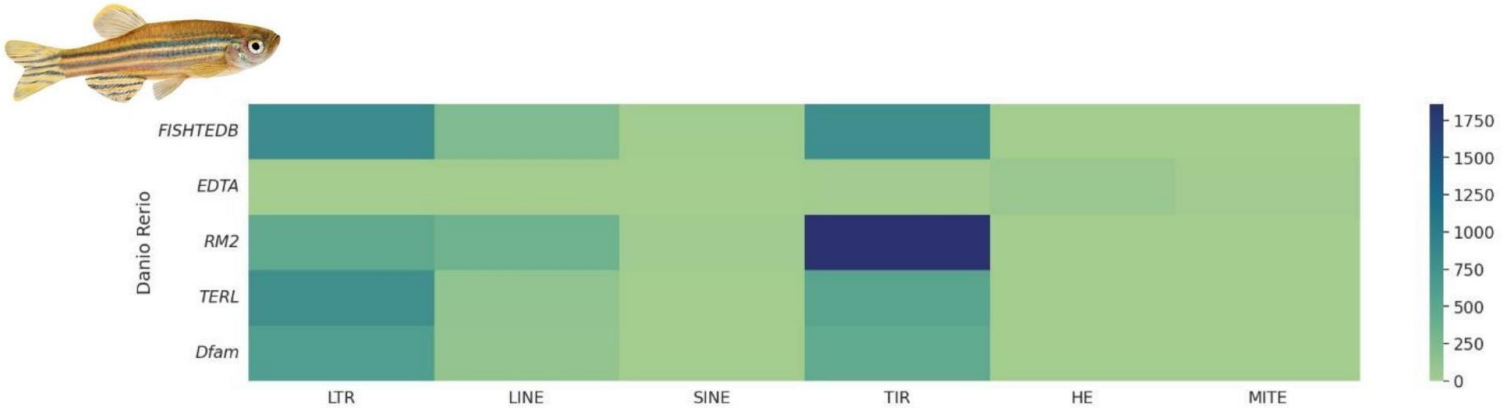
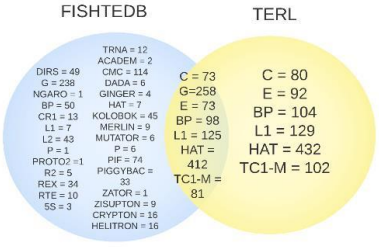


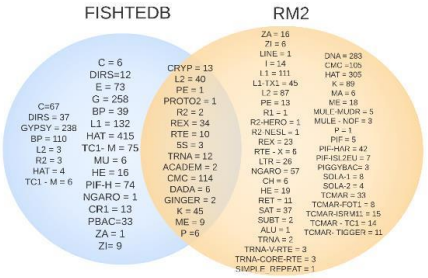
Figura 14: Heatmap - Comparação da espécie *Danio rerio*. No eixo Y temos as ferramentas (TERL, RM2, EDTA) e o banco de dados FISHTEDB. Eixo X a classificação dos elementos transponíveis, pelas ordens: LTR, LINE, SINE, TIR, HE, MITE.

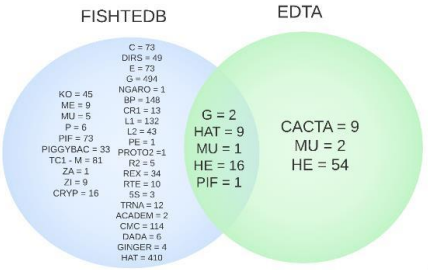
Para verificar quais das classificações foram iguais entre eles, ou seja, presente na mesma coordenada, realizamos essa análise por meio do software GFFCompare. Com os resultados, montamos um diagrama de Venn para melhor observação (Figura 15).

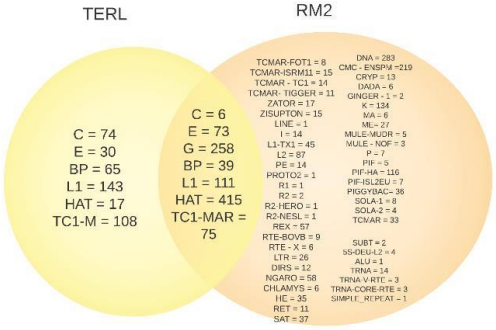
Para essa visualização deixamos diagramas dos resultados entre cada ferramenta, e entre todas elas. Como já analisado pelo heatmap, a ferramenta EDTA foi a que obteve menos resultados na classificação de TEs, comparado às demais. Ao todo as ferramentas encontraram elementos exclusivamente, e em especial a ferramenta RepeatModeler2, classificou mais diversas superfamílias.

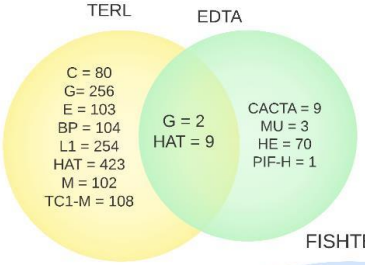
Danio rerio

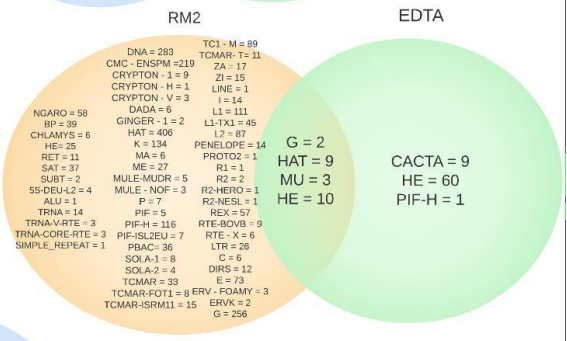


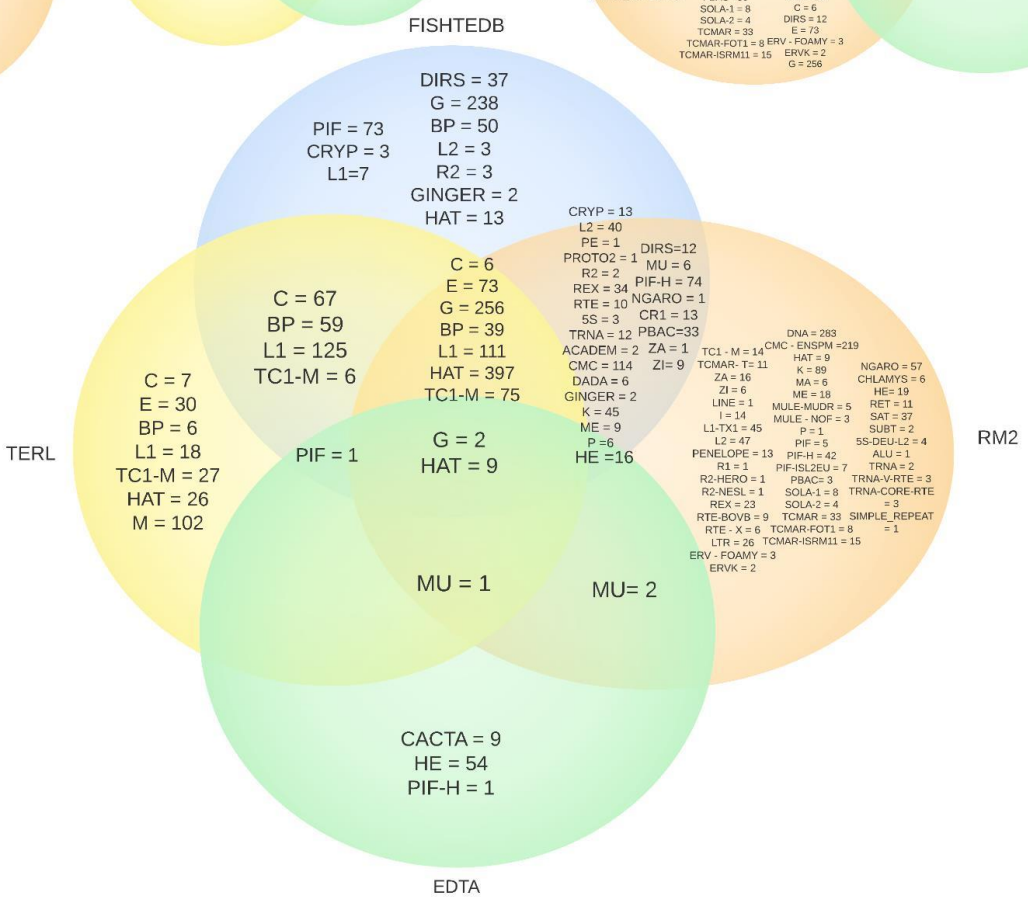












LEGENDA:
C= Copia
E= ERV
G=Gypsy
BP= Bel Pao
M= Mariner
PE= Penelope
HE=Helitron
MU=Mutator
CAC=CACTA
K = KOLOBOK
ME = Merlin
ZA = Zator
ZI=Zisupton
PBAC = PiggyBac
RM2 = RepeatModeler2

Figura 15: Diagrama de Venn - *Danio rerio*. Comparação de quantidade de classificação em superfamília nas abordagens citadas. Nas áreas de intersecção, apresenta os valores encontrados na mesma coordenada entre elas. Fora da intersecção, valores que apenas foram apresentados em uma ferramenta.

Com os valores das análises comparativas na mesma coordenada, pudemos calcular o TE-Score (Figura 16). TE-Score é a porcentagem de que cada ferramenta classifica TEs entre elas, e o que cada uma encontrou sozinha.

Observamos que a maior porcentagem obtida, foi a classificação entre uma ferramenta, com 52.2%. Vemos que todas as ferramentas apenas encontraram 0,3% de toda a classificação de TEs nessa espécie. O que havíamos visto pelo diagrama, com as quantidades diversas de TEs encontrados individualmente em cada abordagem.

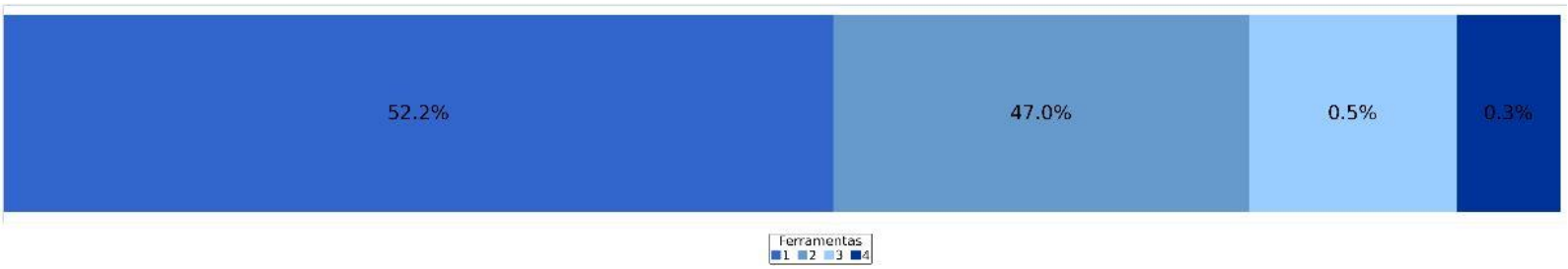


Figura 16: TE-Score *Danio rerio*. Apresentação da porcentagem de cada ferramenta, sendo na ordem uma ferramenta, duas ferramentas, três ferramentas e todas as ferramentas.

3.2.2 Colossoma macropomum

Para a análise da espécie nativa brasileira Colossoma macropomum, comparamos os resultados das ferramentas e o trabalho de Hildsdorf, que realizou a anotação de TEs dessa espécie. Nessa espécie em específico, não possui dados no banco de dados FISHTEBD.

Para isso, fizemos o heatmap de cada uma das abordagens (Figura 17). Observamos que no geral as ferramentas encontraram valores parecidos na ordem LTR, HE, PE e SINE. Já na ordem LINE o trabalho Hildsdorf teve maiores classificações, e na ordem TIR possui valores semelhantes ao do EDTA.

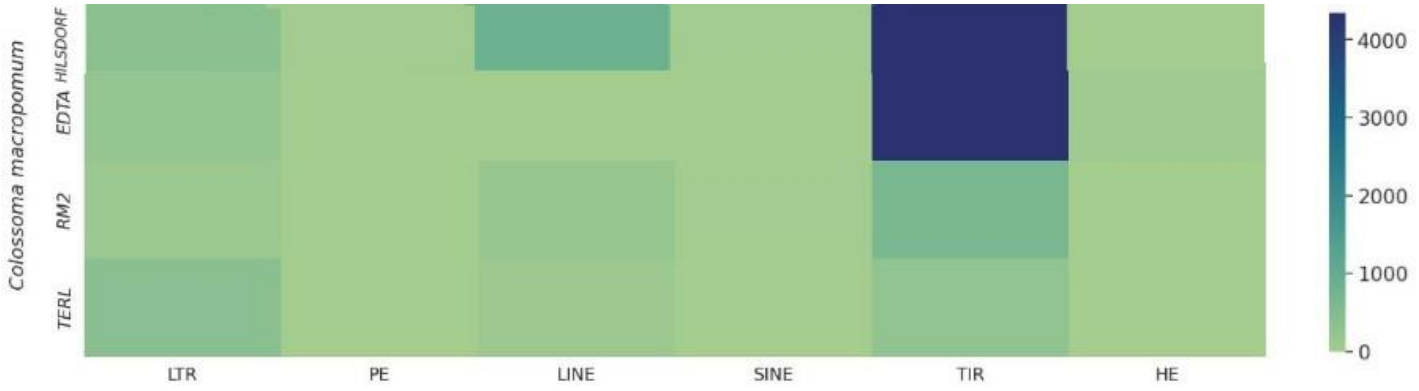


Figura 17: Heatmap - *Colossoma macropomum.* Eixo Y representando as abordagens utilizadas. Eixo X mostrando as ordens de elementos transponíveis classificadas. Quanto mais escura a cor, maior é a quantidade de elementos encontrados. No canto direito, há a legenda com as cores e a quantidade respectivamente.

Para uma melhor visão sobre os dados, fizemos o Diagrama de Venn (Figura 18). Mostrando a quantidade de superfamílias classificadas na mesma coordenada, entre os métodos citados.

No geral, observamos que todas as abordagens identificam superfamílias que não estão na mesma coordenada, sendo o RM2 a ferramenta que possui maiores valores dessa anotação. Já as na mesma coordenada, notamos que em questão de quantidade, as ferramentas TERL e EDTA são as que possuem maiores valores em cada superfamília. Hilsdorf e RM2, são as que encontram superfamílias mais distintas.

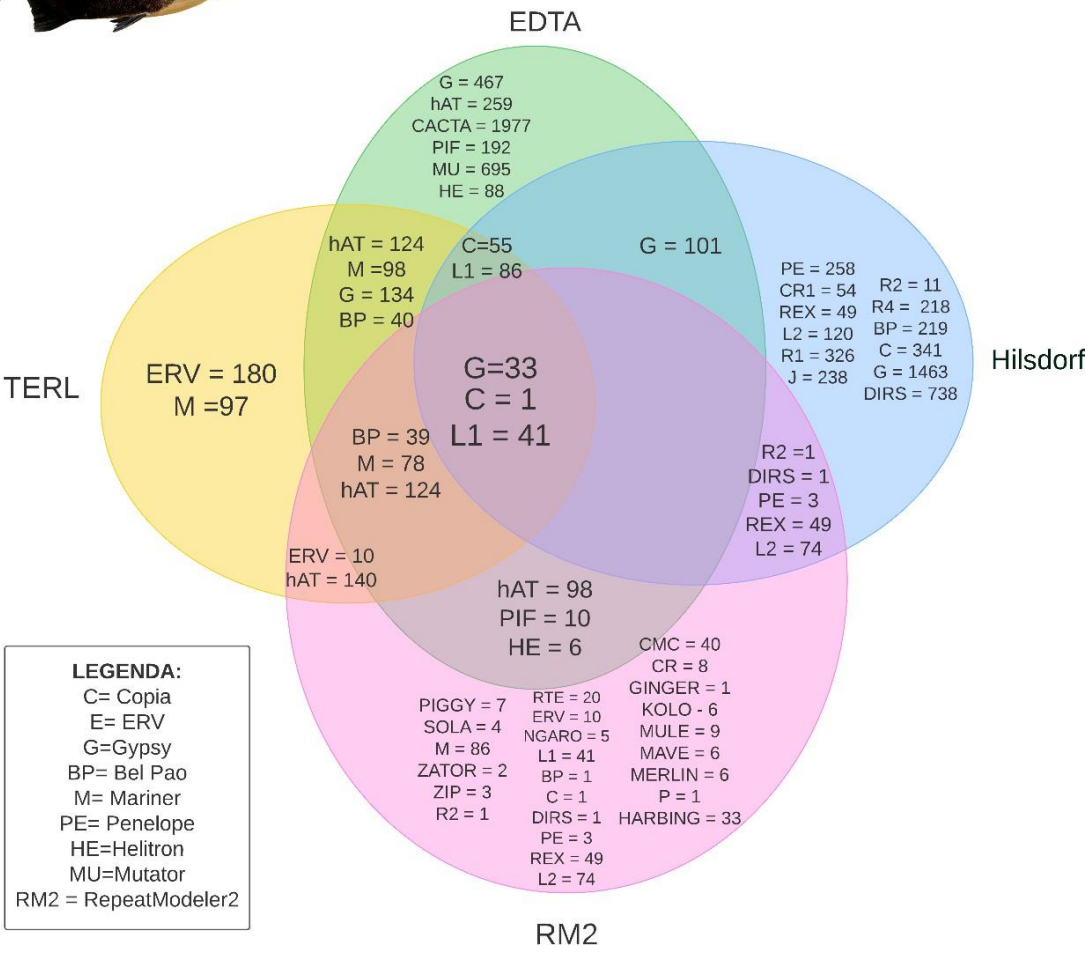


Figura 18: Diagrama de Venn - *Colossoma macropomum*. Comparação de quantidade de classificação em superfamília nas abordagens citadas. Nas áreas de intersecção, apresenta os valores encontrados na mesma coordenada entre elas. Fora da intersecção, valores que apenas foram apresentados em uma ferramenta.

Com os resultados das classificações na mesma coordenada, calculamos o TE-Score, para verificação em porcentagem dessa análise, apresentada na Figura 19. Podemos ver que os resultados entre todas as ferramentas foi de 0,8% ao todo. Já o que cada uma encontrou separadamente, temos 86,2% do valor total.

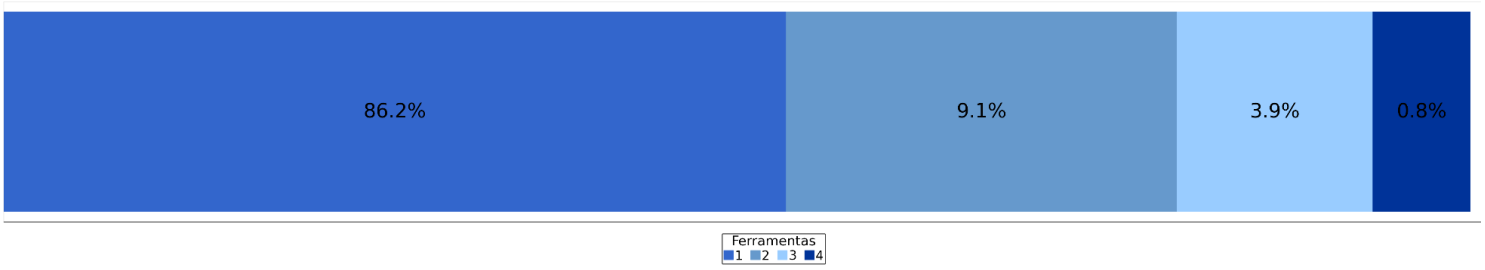


Figura 19: TE-Score *Colossoma macropomum.* Representando a porcentagem dos resultados encontrados entre as 4 abordagens, entre 3, em 2 e o que foi classificado separadamente. A legenda abaixo apresenta as respectivas cores das ferramentas, sendo na ordem crescente.

3.2.3 Gadus morhua

Com os resultados obtidos pelas ferramentas e o banco de dados (Figura 20), observamos que o EDTA foi a ferramenta que mais classificou elementos transponíveis, com destaque na ordem TIR, que obteve a coloração mais escura. Entre as ferramentas FISHTEDB e RM2, a classificação está muito próxima.

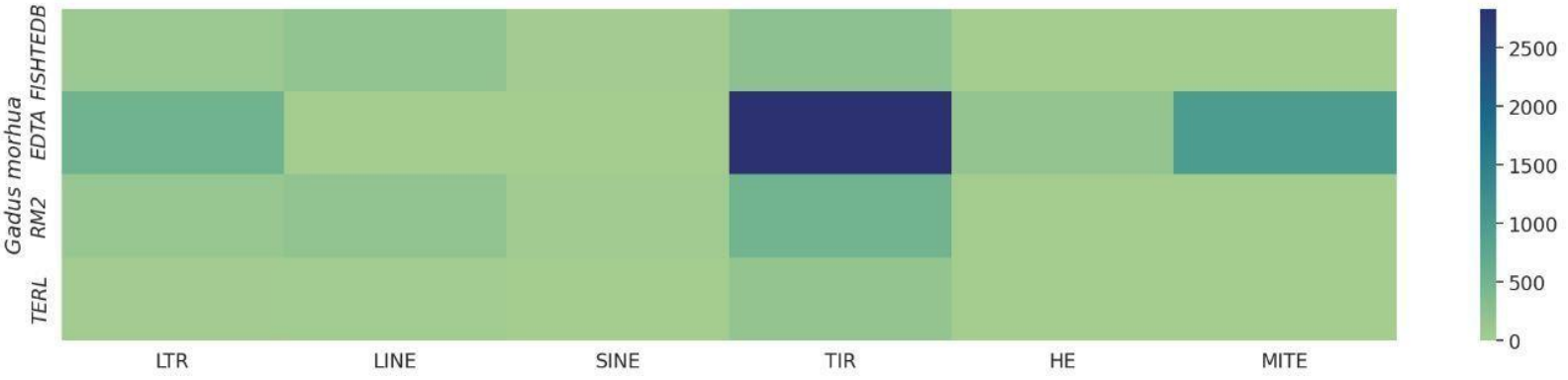


Figura 20: Heatmap - *Gadus morhua*. Eixo Y representando as abordagens utilizadas. Eixo X mostrando as ordens de elementos transponíveis classificadas. Quanto mais escura a cor, maior é a quantidade de elementos encontrados. No canto direito, há a legenda com as cores e a quantidade respectivamente.

Para melhores análises, fizemos a comparação da classificação presente nas mesmas coordenadas. Com o resultado, fizemos um diagrama de Venn (Figura 21). Notamos que as ferramentas utilizadas em nosso trabalho encontram TEs presentes no FISHTEDB. E ainda encontraram mais superfamílias presentes no genoma da espécie.

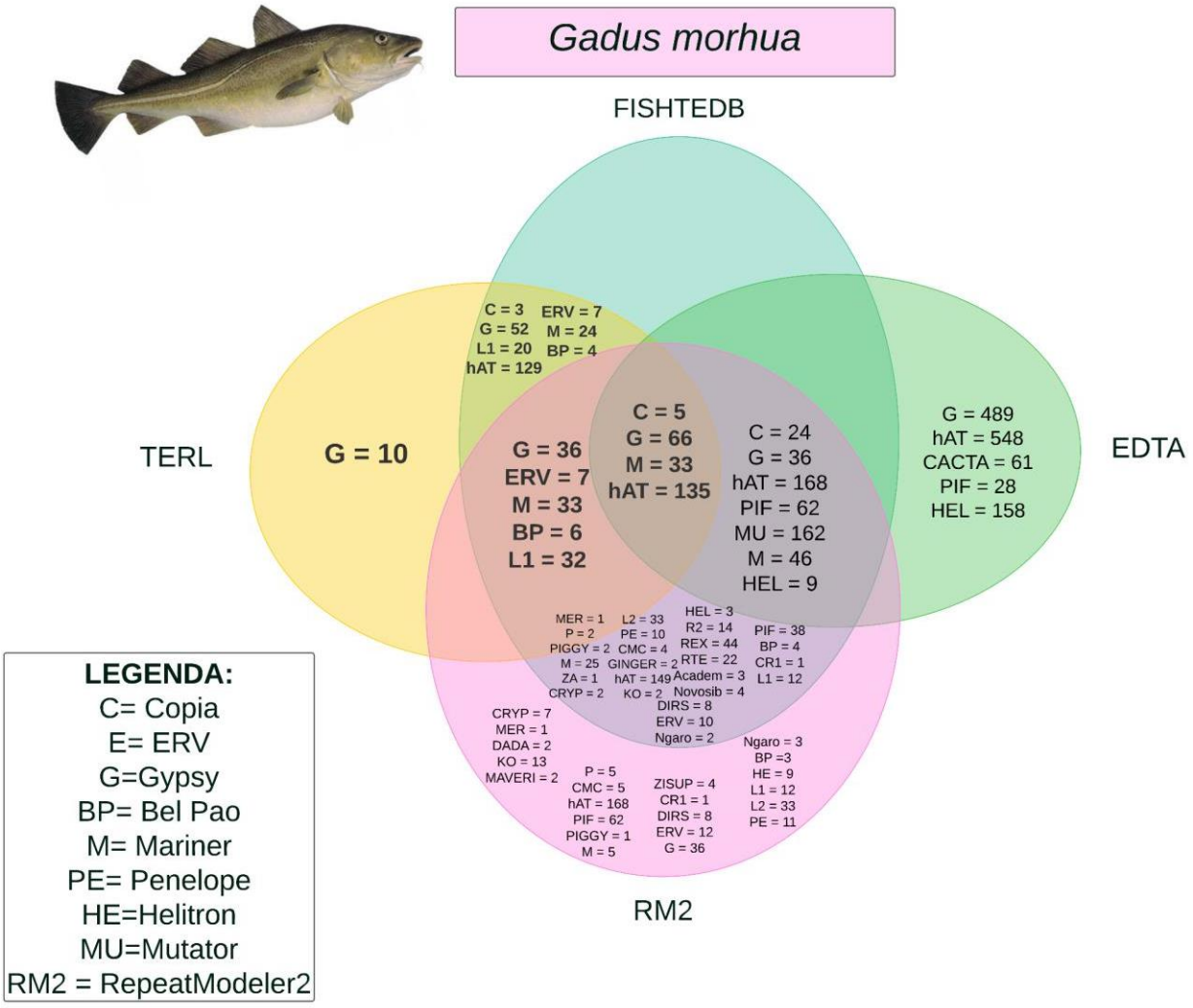


Figura 21: Diagrama de Venn - *Gadus morhua*. Comparação de quantidade de classificação em superfamília nas abordagens citadas. Nas áreas de intersecção, apresenta os valores encontrados na mesma coordenada entre elas. Fora da intersecção, valores que apenas foram apresentados em uma ferramenta.

Com os resultados do diagrama, calculamos o TE-Score dessa espécie (Figura 21). Onde obtivemos as porcentagens de 7,5% de elementos classificados entre todas as abordagens; 19,5% entre três ferramentas; 19,7% entre duas e 53,3% em cada uma delas.

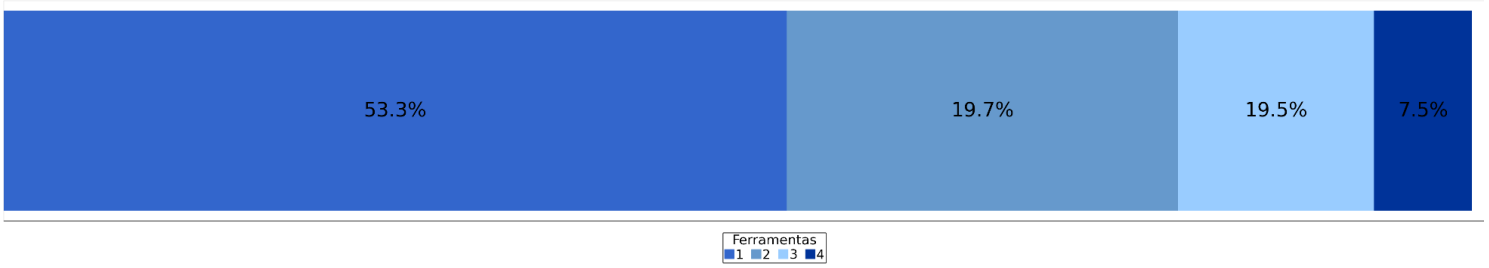


Figura 22: TE-Score *Gadus morhua*.Representando a porcentagem dos resultados encontrados entre as 4 abordagens, entre 3, em 2 e o que foi classificado separadamente. A legenda abaixo apresenta as respectivas cores das ferramentas, sendo na ordem crescente.

3.2.4 Latimeria chalumnae

Para a análise mais detalhada da classificação dos TEs do celacanto, fizemos o heatmap a seguir (Figura 23). Reparamos que há uma grande presença de TEs na ordem LINE, e que o banco de dados FISHTEDB encontrou um pouco a mais que o TERL, como mostrada pela diferenciação de cor, e isso pode ser explicado pois a ferramenta TERL só possui a superfamília L1 presente nesta ordem. Mas as demais ferramentas, no resultado do LINE, possuem um grande diferencial de coloração. As demais ordens vemos uma semelhança.

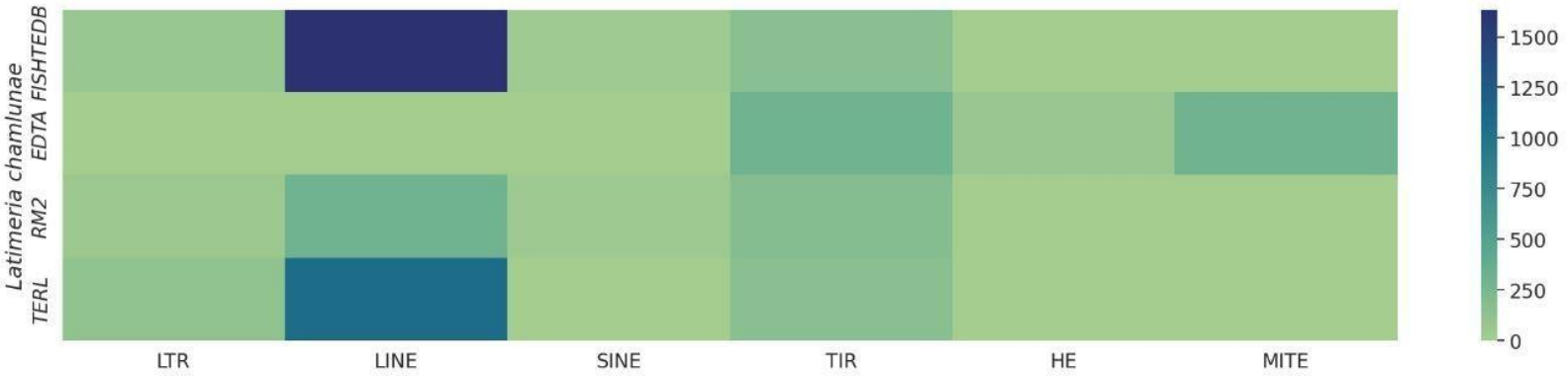


Figura 23: Heatmap - *Latimeria chalumnae*. Eixo Y representando as abordagens utilizadas. Eixo X mostrando as ordens de elementos transponíveis classificadas. Quanto mais escura a cor, maior é a quantidade de elementos encontrados. No canto direito, há a legenda com as cores e a quantidade respectivamente.

Para melhor verificação dessas diferenças obtidas nos resultados, montamos o diagrama Venn (Figura 24), baseado na classificação presente na mesma coordenada. Observamos que em todas as abordagens foi classificado TEs diversos, sendo o RepeatModeler 2 com mais diferenças. Todas encontraram em conjunto TEs da superfamília hAT e Tc1-Mariner, e entre elas quem obteve maiores resultados foi o TERL, FISHTEDB e RM2.

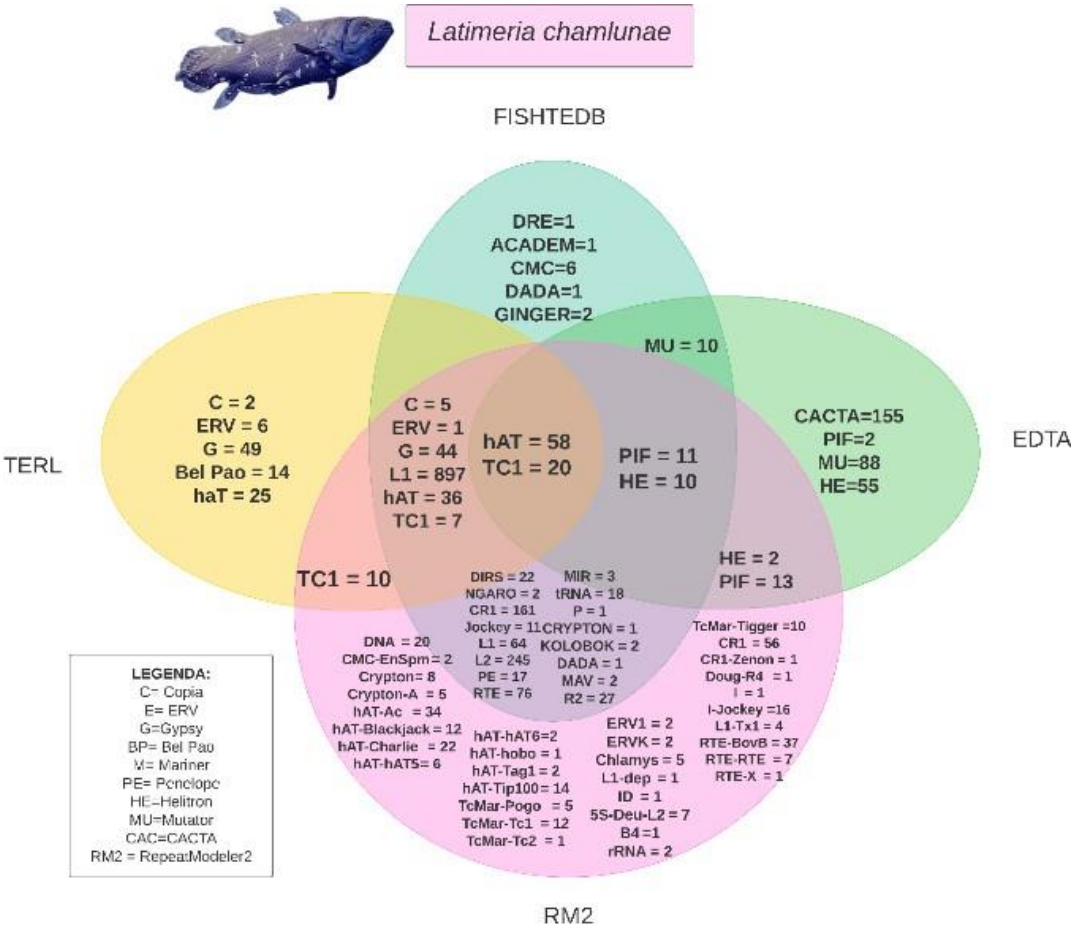


Figura 24: Diagrama de Venn - *Latimeria chalumnae*. Comparação de quantidade de classificação em superfamília nas abordagens citadas. Nas áreas de intersecção, apresenta os valores encontrados na mesma coordenada entre elas. Fora da intersecção, valores que apenas foram apresentados em uma ferramenta.

Com os resultados do diagrama, realizamos o cálculo do TE-Score para esta espécie (Figura 25). Que obteve em apenas uma ferramenta 28,5% da classificação de TEs no total, entre duas ferramentas 27,7%, entre três 40,7% como já previsto pelo diagrama, por ter sido uma alta quantidade de classificação em conjunto, e entre todas 3,1%.

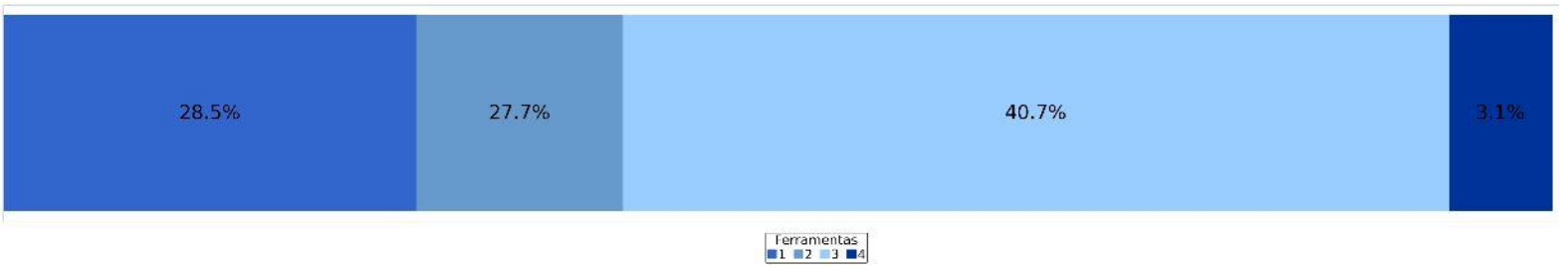


Figura 25: TE-Score *Latimeria chalumnae.* Representando a porcentagem dos resultados encontrados entre as 4 abordagens, entre 3, em 2 e o que foi classificado separadamente. A legenda abaixo apresenta as respectivas cores das ferramentas, sendo na ordem crescente.

3.2.5 Oreochromis niloticus

Com os resultados da tilápia do Nilo (*O. niloticus*), montamos o heatmap (Figura 26). A ferramenta EDTA, foi a que obteve mais classificação entre as ordens encontradas, sendo LTR, TIR, MITE as que mais possuem quantidade de elementos transponíveis presente. As demais seguem um mesmo padrão de cor, sendo parecidas nos resultados.

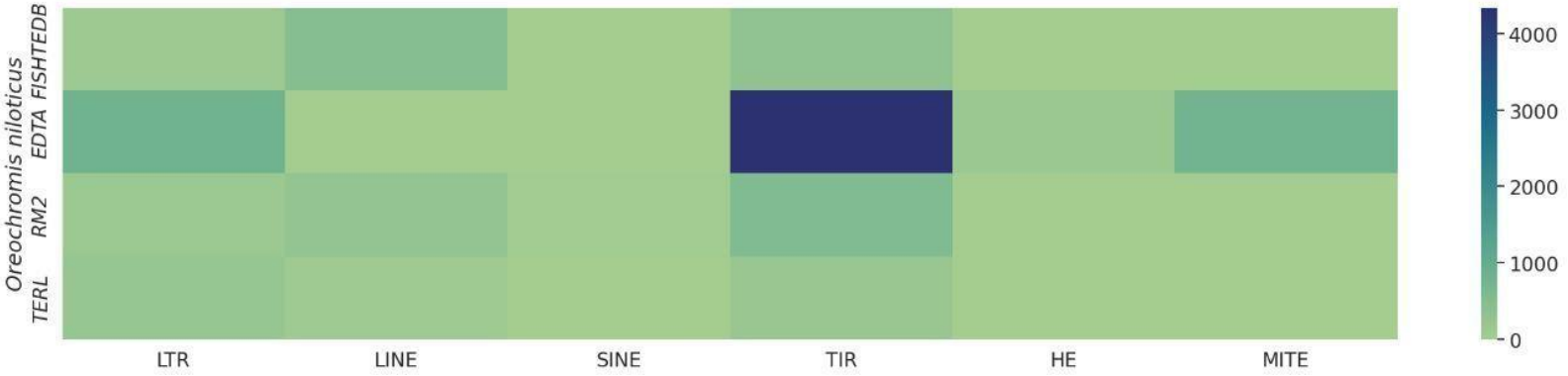


Figura 26: Heatmap - *Oreochromis niloticus.* Eixo Y representando as abordagens utilizadas. Eixo X mostrando as ordens de elementos transponíveis classificadas. Quanto mais escura a cor, maior é a quantidade de elementos encontrados. No canto direito, há a legenda com as cores e a quantidade respectivamente.

Para melhor observação, fizemos o diagrama de Venn (Figura 27), com as classificações entre eles. Há uma grande quantidade de TEs classificadas entre todas as abordagens. RM2 foi a ferramenta que mais classificou TEs na mesma coordenada que as demais, e entre as ferramentas, vimos a intersecção de RM2 com o TERL, RM2 com EDTA e RM2 com FISHTEDB. Sendo a única ferramenta presente em todos os pontos de intersecção de classificações iguais.

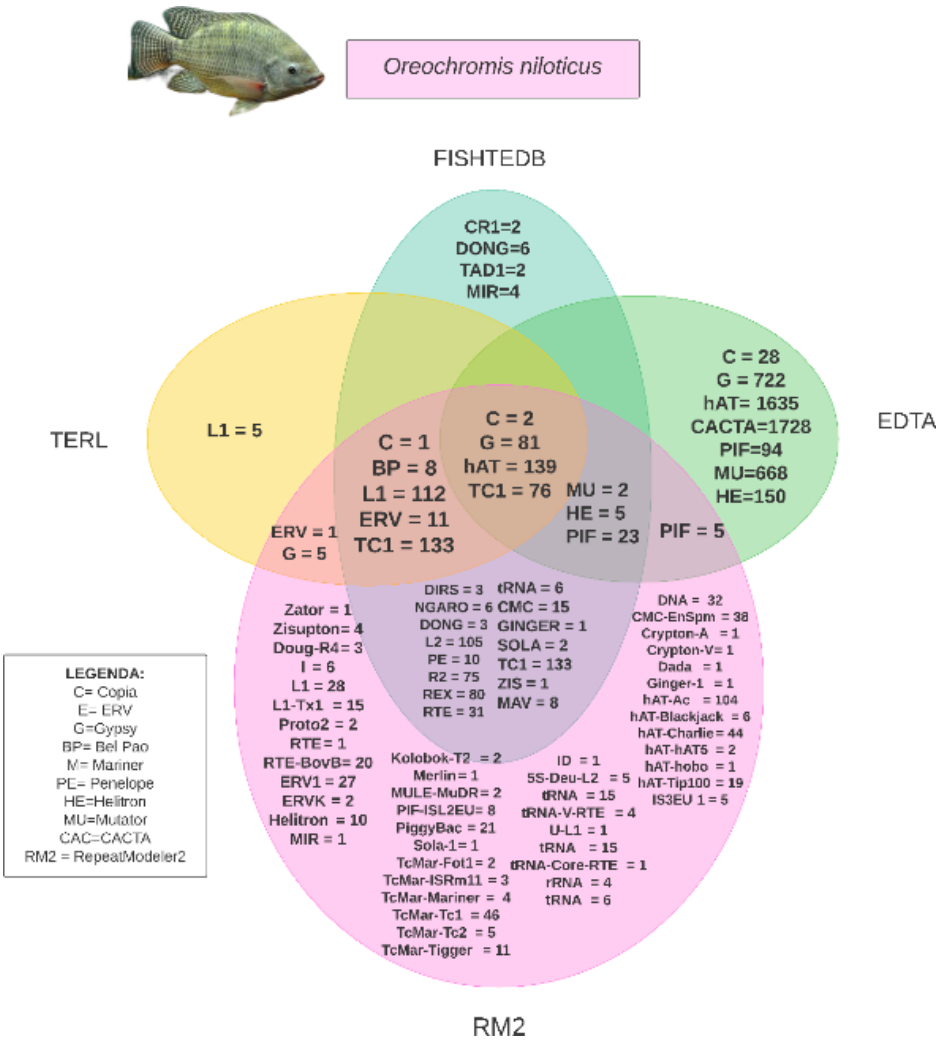


Figura 27: Diagrama de Venn - *Oreochromis niloticus*. Comparação de quantidade de classificação em superfamília nas abordagens citadas. Nas áreas de intersecção, apresenta os valores encontrados na mesma coordenada entre elas. Fora da intersecção, valores que apenas foram apresentados em uma ferramenta.

Calculamos o TE-Score da tilápia com base nos resultados do diagrama de Venn (Figura 28). Obtivemos excelentes valores de classificação de uma ferramenta, sendo 83,7% do total. Entre duas foi de 7,4%, de três 4,4% e entre todas 4,5%.

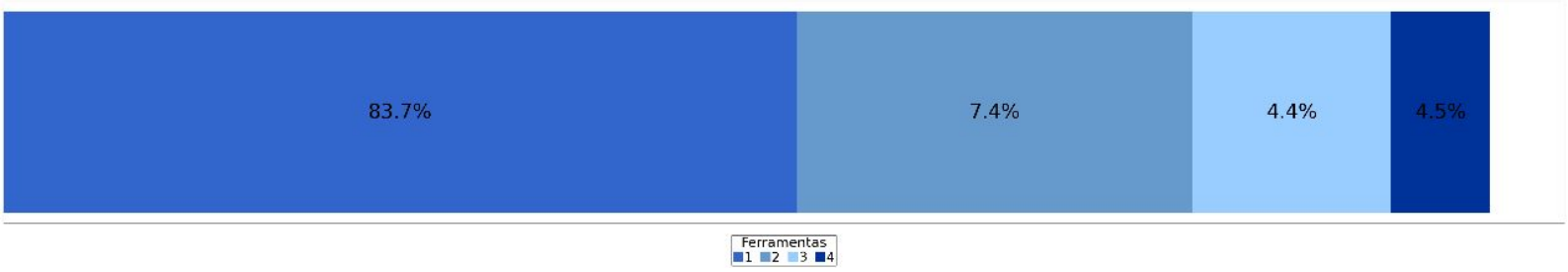


Figura 28: TE-Score *Oreochromis niloticus*. Representando a porcentagem dos resultados encontrados entre as 4 abordagens, entre 3, em 2 e o que foi classificado separadamente. A legenda abaixo apresenta as respectivas cores das ferramentas, sendo na ordem crescente.

3.2.6 Astyanax mexicanus

Com as classificações das ferramentas obtidas na espécie Lambari, produzimos o heatmap (Figura 29) com os resultados. Notamos que a ferramenta EDTA, possui valores mais altos nas quantidades de TEs encontrados. O TERL e FISHTEDB possuem classificação muito parecida entre as ordens. As ferramentas RepeatModeler 2 e TERL, tem a ordem TIR bem próxima. No geral, observamos que a ordem SINE é a menos encontrada nessa espécie.

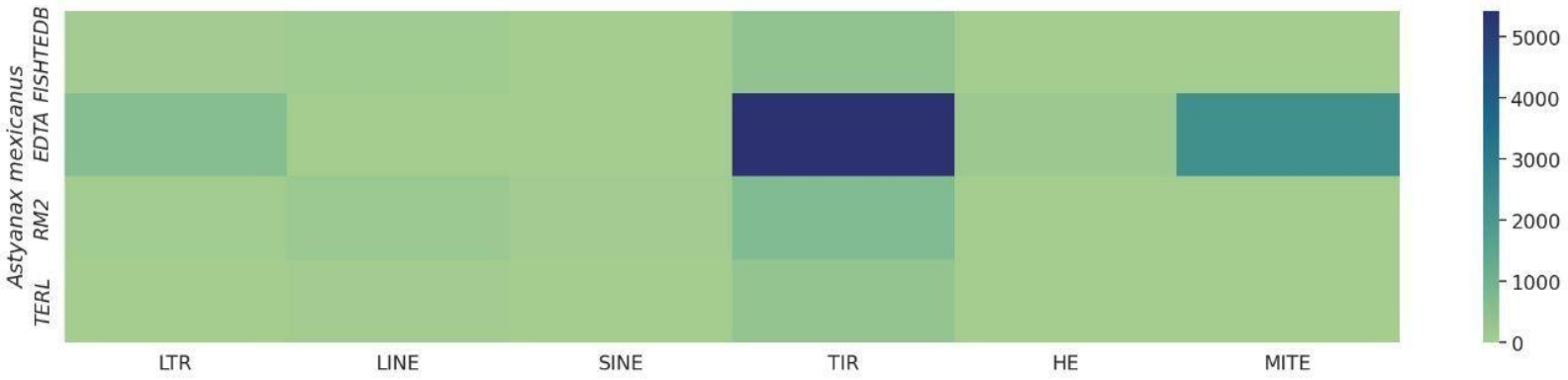


Figura 29: Heatmap - *Astyanax mexicanus*. Eixo Y representando as abordagens utilizadas. Eixo X mostrando as ordens de elementos transponíveis classificadas. Quanto mais escura a cor, maior é a quantidade de elementos encontrados. No canto direito, há a legenda com as cores e a quantidade respectivamente.

Para melhor investigação dos resultados, montamos o diagrama de Venn (Figura 30), com os valores das classificações em superfamílias, presentes nas intersecções, de cada ferramenta utilizada. Onde observamos que todas as abordagens encontram uma grande quantidade de TEs apenas eles, sendo o FISHTEDB o que encontrou menos, apenas uma superfamília MIR. Entre todas as abordagens, foram classificadas superfamílias Tc1-Mariner, hAT e Gypsy. Há ainda, uma grande quantidade de classificação entre RM2 e o FISHTEDB.

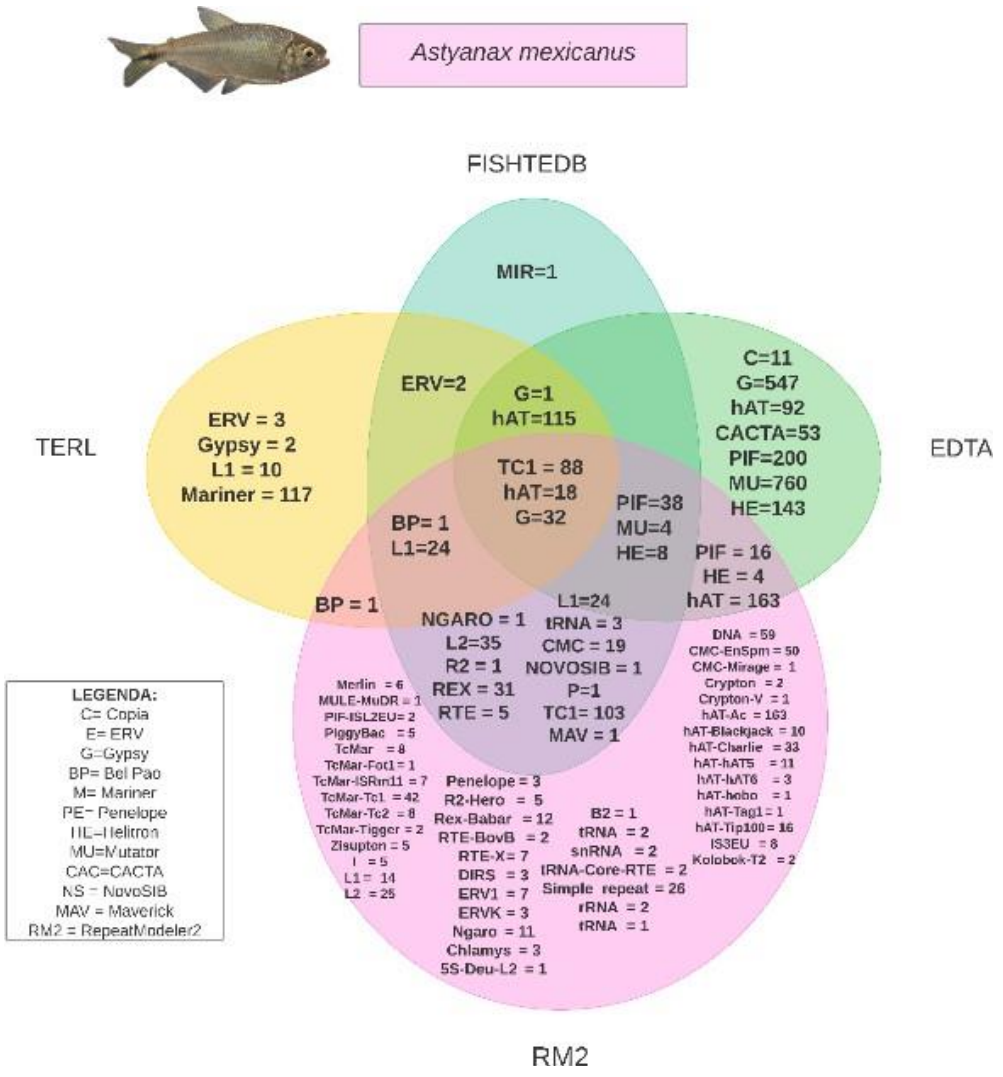


Figura 30: Diagrama de Venn - *Astyanax mexicanus*. Comparação de quantidade de classificação em superfamília nas abordagens citadas. Nas áreas de intersecção, apresenta os valores encontrados na mesma coordenada entre elas. Fora da intersecção, valores que apenas foram apresentados em uma ferramenta.

Com os resultados do diagrama, calculamos as porcentagens do TE-Score (Figura 31). Tendo resultados de classificação das ferramentas sozinhas, sendo 77,4%, entre duas 12,6%, entre três 2,3% e entre todas em conjunto 7,8%.

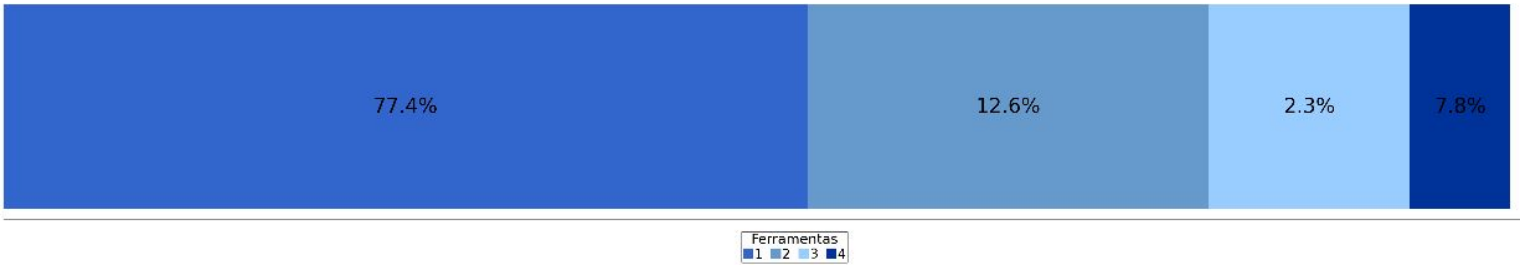


Figura 31: TE-Score *Astyanax mexicanus*. Representando a porcentagem dos resultados encontrados entre as 4 abordagens, entre 3, em 2 e o que foi classificado separadamente. A legenda abaixo apresenta as respectivas cores das ferramentas, sendo na ordem crescente.

CONCLUSÕES

Este artigo apresentou a anotação de TE em genomas de peixes completamente sequenciados. No caso, dados de transcritos de 142 genomas foram analisados pelo classificador TERL. E dados de genomas, de 37 espécies de peixes selecionadas, foram analisadas pelas ferramentas EDTA e RepeatModeler2.

A análise comparativa dos TEs encontrados, entre nossa abordagem e a literatura (FISHTEDB; Hilsdorf em tambaqui), as quantidades de superfamílias e ordem, o quanto cada abordagem e literatura têm em comum ou identificaram, e a relação entre tamanho de genoma versus quantidade de TE são as principais contribuições discutidas.

O resultado completo de todas as ferramentas, dados e a comparação foi feito para as espécies *Gadus morhua*, *Danio rerio*, *Latimeria chalumnae*, *Colossoma macropomum*, *Oreochromis niloticus* e *Astyanax mexicanus*, observando que as ferramentas encontraram igual e a mais do que havia na literatura. Por fim, TERL identificou um total de 64.014 TEs, todos classificados em 7 superfamílias, sendo Gypsy e Copia com maior diversidade nas sequências. EDTA, identificou 15.257 TEs, classificados em 15 superfamílias, e RepeatModeler2 identificou 28.595 TEs em 116 superfamílias. Acredita-se que este trabalho poderá contribuir para uma discussão do papel dos TEs em peixes.

Referências Bibliográficas

ALVES, A. L., VARELA, E. S., MORO, G. V.; KIRSCHNIK, L. N. G. **Riscos genéticos da produção de híbridos de peixes nativos**. Embrapa Pesca e Aquicultura-Documentos (INFOTECA-E), 2014.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; et al. **Basic local alignment search tool**. Journal of Molecular Biology, v. 215, n. 3, p. 403–410, 1990.

ARAUJO - LIMA, C.A.R.M.; GOULDING, M. 1998. **Os frutos do Tambaqui: ecologia, conservação e cultivo na Amazônia**. Sociedade Civil Mamirauá/CNPq/Rainforest Alliance, Brasil. 186p.

ARIA, M., & CUCCURULLO, C. **Bibliometrix: An R-tool for comprehensive science mapping analysis**. Journal of Informetrics , 11(4), 959-975,2017.

BAO Z., EDDY S. R., **Automated de novo identification of repeat sequence families in sequenced genomes**. Genome Res. 12, 1269–1276, 2002.

BARTLEY, D. **Review of the status of aquaculture genetics**. Aquaculture in the Third Millenium, p. 137-166, 2001.

BIÉMONT, C., VIEIRA, C. Junk **DNA as an evolutionary force**. Nature 443: 521-524, 2006.

BIET, E., SUN, J., DUTREIX, M**. Conserved sequence preference in DNA binding among recombinant proteins: abnormal effect of ssDNA secondary structure**. Nucleic Acids Research. 27: 596-600, 1999.

BOHNE, A., BRUNET, F.,GALIANA, D., SCHULTHEIS, C.,VOLFF, J.N. **Transposable elements as drivers of genomic and biological diversity in vertebrates.** Chromosome Research 16: 203-215, 2008.

CASAL, C. M. V. **Global documentation of fish introductions: The growing crisis and recommendations for action.** Biological Invasions, 8: 3-11,2 006.

CHARLESWORTH, B.; LANGLEY, C. H**. The evolution of self-regulated transposition of transposable elements.** Genetics, 112: 359-383,1986.

CNGBdb - China National GeneBank DataBase. Cngb.org. Disponível em: <https://db.cngb.org/>. Acesso em: 17 fev. 2023.

CRESCÊNCIO, R. **Ictiofauna brasileira e seu potencial para criação. In: Espécies nativas para piscicultura no Brasil**. UFSM, ed. Santa Maria, 2005, p.23-26.

CRUZ, M. H. P.; DOMINGUES, D. S.; SAITO, P. T. M.; PASCHOAL, A. R.; BUGATTI, P. H. **TERL: classificação de elementos transponíveis por redes neurais convolucionais.** Briefings in Bioinformatics, Vol. 22, 2020.

DAIRIKI, J.K.; SILVA, T.B.A. Revisão de Literartura: **Exigências nutricionais do tambaqui-compilação de trabalhos, formulação de ração adequada e desafios futuros. Embrapa Amazônia Ocidental.** Manaus, AM, 2011.

DOMINGOS, R. M. da C.; LASP-HIV1-RESTool: **Desenvolvimento de uma ferramenta de bioinformática para análise de resistência do HIV-1 aos antirretrovirais**. 2010.

DUVERNELL, D. D.; PRYOR, S. R. ; ADAMS, S. M. **Teleost Fish Genomes Contain a Diverse Array of L1 RetrotransposonLineages That Exhibit a Low Copy Number and High Rate of Turnover**. Journal of Molecular Evolution, v. 59, n. 3, p. 298–308,

2004.

ELLINGHAUS D., KURTZ S., WILHOEFT U., **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons**. BMC Bioinformatics 9, 18, 2008.

ENGEPESCA. Disponível em< https://www.engepesca.com.br/post/principais-dados-da-pesca-brasileira-em-2021-e-perspectivas-para-2022> Acessado em 26 de Janeiro de 2022.

Ensembl. Species List. Ensembl.org**.** Disponível em: <http://www.ensembl.org/info/about/species.html>. Acesso em: 15 ago. 2022.

FLYNN, J. M.; HUBLEY, R.; GOUBERT, C.; et al. **RepeatModeler2 for automated genomic discovery of transposable element families**. Proceedings of the National Academy of Sciences, v. 117, n. 17, p. 9451–9457, 2020.

FROESE, R., PAULY D. Editors. 2022. **FishBase**. World Wide Web electronic publication. www.fishbase.org, version (08/2022).

FURANO, A. **L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish**. Trends in Genetics, v. 20, n. 1, p. 9–14, 2004.

GASPARINO, E.; OSÓRIO, P. de S.; SOARES, M. A.M.; MARQUES, D. S.; BLANCK, D.V., LUIZETTI, F**. Uso da bioinformática na diferenciação molecular da Entamoeba histolytica e Entamoeba díspar**, 2008.

GenBank TSA. Disponível em < https://www.ncbi.nlm.nih.gov/genbank/tsaupdate/>, acesso em 02 de outubro de 2021.

GÉRY, J. **Characoids of the world**. Publications, Neptune City, New Jersey,1977.

HILSDORF, A. W. S.; ULIANO-SILVA, M.; COUTINHO, L. L.; PINHAL, D. et al. **Genome assembly and annotation of the tambaqui (Colossoma macropomum): an emblematic fish of the Amazon River basin.** 2021.

IBGE – Instituto Brasileiro de Geografia e Estatísitica, 2016. Pesquisa da Pecuária Municipal.

JURKA, J., KAPITONOV, V.V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O., Walichiewicz, J. (2005). **Repbase Update, a database of eukaryotic repetitive elements.** Cytogenetics and Genome Research. 110: 462-467.

KAZAZIAN, H.H. Mobile elements: drivers of genome evolution. Science 303: 1626-1632, 2004.

KIDWELL, M.G. **Transposable elements and the evolution of genome size in eukaryotes.** Genetica 115: 49-63, 2002.

KUBITZA F, CAMPOS JL, ONO EA, ISTICHUK PI. **Panorama da Piscicultura no Brasil: Estatísticas, espécies, polos de produção e fatores limitantes à expansão da atividade**. Panorama da Aquicultura; 22(132):14-23, 2012.

LI, Y.C., KORD, A.B., FAHIMA, T., BERLES A., NERO E. **Microsatellites: genomic distribution, putative functions and mutation mechanisms: a review**. Molecular Ecology 11: 2453-2465, 2002.

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

LIU, Z., LI, P., KOCABAS, A., KARSI, A., JU, Z. **Microsatellite containing genes from**

the channel catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B. Biochemical and Biophysical Research Communication 289: 317-324, 2001.

MARTINS, M.L.; ONAKA, E.M.; MORAES, F.R.; BOZZO, F.R.; PAIVA, A.M.F.C GONÇALVES, A. **Recent studies on parasitic infections of freshwater cultivated fish in the state of São Paulo**, Brazil. Acta Scientiarum, v. 24, n. 4, p. 981-985, 2002.

MOREIRA, A.A.; HILSDORF, A.W.S.; SILVA, J.V.; SOUZA, V.R. **Variabilidade genética de duas variedades de tilápia nilótica por meio de marcadores microssatélites**. Pesquisa Agropecuária Brasileira, v.42, p.521-526, 2007.

National Center for Biotechnology Information. Nih.gov. Disponível em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 17 fev. 2023.

OU S., SU W., LIAO Y., CHOUGULE K., AGDA J. R. A., HELLINGA A. J., LUGO C. S. B., ELLIOT T. A., WARE D., PETERSON T., JIANG N., HIRSCH C. N.; HUFFORD M. B. **Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined**. Comprehensive Pipeline. Genome Biol. 20(1): 275,2018.

PERTEA, G. ; PERTEA, M.. **GFF Utilities: GffRead and GffCompare.** F1000Research, v. 9, p. 304, 2020.

PRICE A. L., JONES N. C., PEVZNER P. A., ***De novo* identification of repeat families in large genomes**. Bioinformatics 21 (suppl. 1), i351–i358, 2005.

SILVA, S.; BONAN, T.C.; SILVA, S.D.M.; et al. **Degradation of different composite resins after acid erosion and brushing. Dental Materials**, v. 28, p. e56, 2012.

SIMONETTI, R. B. **Zebrafish (Danio rerio): futuro do modelo animal em pesquisa biomédica**. Universidade Federal do Rio Grande do Sul Faculdade de veterinária, 2014.

SMIT, **Repeatmasker.org.** Disponível em: <https://www.repeatmasker.org/RepeatMasker/>. Acesso em: 17 fev. 2023.

SHAO, F.; WANG, J.; XU, H.; et al. **FishTEDB: a collective database of transposable elements identified in the complete genomes of fish**. Database, v. 2018, 2018.

SHAO, F.; H., Minjin ; P., Zuogang. **Evolution and diversity of transposable elements in fish genomes**. Scientific Reports, v. 9, n. 1, 2019.

VOLFF, J.N. BOUNEAU, L., OZOUF-COSTAZ, C., FISCHER, C. **Diversity of retrotransposable elements in compact pufferfish genomes**. Trends in Genetics 19: 674-678, 2003.

Capítulo 3

3.1 Conclusão

Em resumo, abordagens computacionais são componentes cruciais para uma identificação *in silico* de TE nos genomas. Embora a identificação automatizada de repetições dispersas seja complexa, existem diversas ferramentas disponíveis na bioinformática que ajudam a caracterizar esses elementos. Com o avanço da tecnologia, é possível esperar que novas ferramentas e abordagens sejam desenvolvidas para aprimorar ainda mais esses estudos. A compreensão mais completa dos elementos transponíveis pode levar a novas descobertas em áreas como a evolução dos genes e a biologia molecular, bem como a aplicações práticas em áreas como a agricultura e a medicina.

A investigação dos papéis dos TEs na evolução do genoma e no impacto nos genes do hospedeiro em peixes pode oferecer insights valiosos para outros vertebrados. Neste estudo, desenvolvemos um pipeline robusto que fornece uma base sólida para estudos funcionais e permitirá a criação de um banco de dados de TEs em peixes. Foram identificados um total de 107.866 TEs em 168 espécies de peixes, classificados em diversas superfamílias, incluindo Cópia, Gypsy, Bel Pao, L1, hAT e Mariner. Nossos resultados destacam a diversidade desses elementos nos genomas de peixes e sugerem que há correlação entre a quantidade de TEs e o tamanho do genoma.

Em particular, com a utilização da ferramenta TERL, podemos notar que as superfamílias Gypsy e Cópia apresentaram maior diversidade do que as outras superfamílias. Na classificação de ordem, notamos que os LTR apresentam uma maior quantidade de elementos transponíveis, e os LINEs apresentam uma quantidade baixa.

Já as ferramentas, RepeatModeler2 e EDTA, encontraram diversas superfamílias de TEs. Ao todo notamos que em classificação de ordem, TIR e LTR apresentaram uma grande quantidade de TEs nas espécies selecionadas.

Diante dos TEs identificados, realizamos a comparação dos resultados com os bancos de dados já existentes, conseguimos observar que a ferramenta TERL, RepeatModeler 2 e EDTA, na maioria das espécies, encontraram maior quantidade de TEs do que já tinha sido encontrado. Para verificar essa sequência de comparação, realizamos o diagrama de venn para verificar quais sequências foram anotadas

igualmente entre os bancos de dados. Sendo feitas para as espécies *Danio rerio*, *Latimeria chalumnae*, *Astyanax mexicanus*, *Oreochromis niloticus*, *Gadus morhua* e *Colossoma macropomum*, verificamos que muitas superfamílias as nossas abordagens encontraram estes elementos que os outros bancos de dados não foram capazes de anotar.

O foco será para o futuro o fechamento desta análise por completo, re-anotando os elementos transponíveis, utilizando a ferramenta RepeatMasker em todos os dados de peixes, atualizando o TEScore para a formação do banco de dados curado final nomeado LorFISHTEDB. Disponibilizando todos os dados via um portal web integrado com o banco de dados LorFISHTEDB.

Do ponto de vista de contribuições, o trabalho produziu: apresentação de trabalho, em pôster, na III Escola Paranaense de Bioinformática 2022; participação nos eventos: II Workshop de Genética e Biologia Molecular- UEL,2022, X-Meeting 2021 e 2023, Simpósio Brasileiro de Bioinformática 2021.

Por fim, é importante ressaltar que o trabalho contou com o apoio da NAPI Bioinformática via Fundação Araucária por meio do convênio 66/2021.

Referências Bibliográficas

ARIA, M. & CUCCURULLO, C. **Bibliometrix: An R-tool for comprehensive science mapping analysis**. Journal of Informetrics, 11(4), pp 959-975, Elsevier,2017.

Associação Brasileira de Piscicultura. **Anuário Peixe BR de 2022 da Piscicultura.** Disponível em <https://www.peixebr.com.br/anuario2022/> acesso em 17 de Janeiro de 2023.

BORGHETTI, N. R. B.; OSTRENSKY, A.; BORGHETTI, J. R. **Aquicultura: uma visão geral sobre a produção de organismos aquáticos no Brasil e no mundo**. Grupo Integrado de Aqüicultura e Estudos Ambientais. Curitiba, 2003.

BIÉMONT C, VIEIRA C. **Genetics: junk DNA as an evolutionary force**. Nature; 443:521–4, 2006.

CANAPA A., BISCOTTI M.A., BARUCCA M., CARDUCCI F., CAROTTI E., OLMO E. **Shedding light upon the complex net of genome size, genome composition and environment in chordate**s. Eur. Zool. J.;87:192–202, 2020.

CARARETO, C. M. A.; VITORELLO, C. B. M.; SLUYS, M. A. V. **Elementos de transposição: diversidade, evolução, aplicação e impactos nos genomas dos seres vivos.** Sociedade Brasileira de Genética, FIOCRUZ. Rio de Janeiro,2015.

CARDUCCI F., BARUCCA M., CANAPA A., CAROTTI E., BISCOTTI M.A. **Mobile elements in ray-finned fish genomes.** Life;10:221, 2020.

CARDUCCI F., BISCOTTI M.A., FORCONI M., BARUCCA M., CANAPA A. **An intriguing relationship between teleost Rex3 retroelement and environmental temperature**. Biol. Lett. 15:20190279, 2019.

CNA, Confederação da Agricultura e Pecuária do Brasil. **Paraná amplia liderança como maior produtor de peixes de cultivo do país**. Disponível em: <https://www.cnabrasil.org.br/noticias/parana-amplia-lideranca-como-maior-produtor-de-peixes-de-cultivo-do-pais>. Acesso em: 9 mar. 2023.

CRAIG, N. L. **Mobile DNA II**. ASM Press. Washington, 2002.

Embrapa Pesca Aquicultura. Disponível em <https://www.embrapa.br/pesca-e-aquicultura> acesso em 17 de Janeiro de 2023.

FLYNN, J. M.; HUBLEY, R.; GOUBERT, C.; et al. **RepeatModeler2 for automated genomic discovery of transposable element families.** Proceedings of the National Academy of Sciences, v. 117, n. 17, p. 9451–9457, 2020.

FREITAS, C. E. C. SOUZA, F. K. S. **O uso de peixes como bioindicador ambiental em áreas de várzea da bacia amazônica**. Revista Agrogeoambiental. Amazonas, 2009.

HILSDORF, A. W. S. et al. **Genome assembly and annotation of the tambaqui (Colossoma macropomum): an emblematic fish of the Amazon River basin.** Cold Spring Harbor Laboratory, bioRxiv, 2021.

KAPITONOV, K. O.; JURKA, M. V. **Repetitive Sequences in Complex Genomes : Structure and Evolution. Annual review of genomics and human genetics**, v. 8, p. 1–21, Maio. 2008.

KOGA A, LIDA A, HORI H, SHIMADA A, SHIMA A. **Vertebrate DNA Transposon as a Natural Mutator: The Medaka Fish Tol2 Element Contributes to Genetic Variation without Recognizable Traces.** Mol Biol Evol, 2016.

LERAT E. et al. **Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs**. Heredity. 104(6):520–533, 2010.

NOVÁK, P.; NEUMANN, P. ;MACAS, J.. **Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data**. BMC Bioinformatics, v. 11, n. 1, 2010.

PAGE, M. J. et al. **The PRISMA 2020 statement: an Updated Guideline for Reporting Systematic Reviews.** British Medical Journal, v. 372, n. 71, p. n71, 29 mar. 2021.

PEREIRA, C. M.; JARDIM, S. S.; SCHUCH, A. P.; LORETO, E. L. S. **Effects of heat and UV radiation on the mobilization of transposon mariner-Mos1**. Cell Stress and Chaperones, v.20, p.841-856, 2015.

SHAO, F. et al, **FishTEDB: a collective database of transposable elements identified in the complete genomes of fish**. Database, v. 2018, 2018.

SHAO, F.; HAN, M. ; PENG, Z. **Evolution and diversity of transposable elements in fish genomes**. Scientific Reports, v. 9, n. 1, 2019.

SMIT, Repeatmasker.org. Disponível em: <https://www.repeatmasker.org/RepeatMasker/>. Acesso em: 17 fev. 2023.

SOUSA, R. G. C.; FREITAS, C. E. C.**The influence of flood pulse on fish communities of floodplain canals in the Middle So-limões River.** Neotropical Ichthyology, v.6, n.2, p.249-255, 2008.

WARREN, I.A.; NAVILLE, M.; CHAPOLIN, D.; Levin, P.; BERGER, C.S.; GALIANA, D.; VOLFF, J.N. **Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates**. Chromosome Research, 23(3), 505-531, 2015.

WICKER, T.; SABOT, F.; HUA-VAN, A.; BENNETZEN, J.;CAPY, P.; CHALHOUB,B.; et al. **A unified classification system for eukaryotic transposable elements.**

Nature Reviews Genetics, v. 8, n. 12, p. 973–982, dez. 2007.