**UNIVERSIDADE DE SÃO PAULO**

**Programa Interunidades de Pós-Graduação em Bioinformática - USP**

HELENA BEATRIZ DA CONCEIÇÃO

# **Origens e Potencial Funcional de Retrocópias no Genoma de Humanos e de Outros Animais:** uma abordagem em larga escala de identificação de retrocópias em diversos genomas animais e o estudo do seu padrão de expressão em tecidos normais humanos

Versão Original da Tese defendida

São Paulo

Data do Depósito na SPG:

**03/10/2023**

HELENA BEATRIZ DA CONCEIÇÃO

# **Origens e Potencial Funcional de Retrocópias no Genoma de Humanos e Outros Animais:** uma abordagem em larga escala de identificação de retrocópias em diversos genomas animais e o estudo do seu padrão de expressão em tecidos normais humanos

*Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para obtenção do Título de Doutor em Ciências.*

*Área de Concentração: Bioinformática*

*Orientador: Dr. Pedro Alexandre Favoretto Galante*
Hospital Sírio-Libanês

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPESP (Processo 2018/13613-4)

São Paulo

2023

Ficha catalográfica elaborada com dados inseridos pelo(a) autor(a)
Biblioteca Carlos Benjamin de Lyra
Instituto de Matemática e Estatística
Universidade de São Paulo

---

Conceição, Helena Beatriz da
Origens e Potencial Funcional de Retrocópias no Genoma de Humanos e Outros Animais: uma abordagem em larga escala de identificação de retrocópias em diversos genomas animais e o estudo do seu padrão de expressão em tecidos normais humanos / Helena Beatriz da Conceição; orientador, Pedro A. F. Galante. - São Paulo, 2023.
93 p.: il.

Tese (Doutorado) - Programa Interunidades de Pós-Graduação em Bioinformática / Instituto de Matemática e Estatística / Universidade de São Paulo.
Bibliografia
Versão corrigida

1. Bioinformática. 2. Retrocópias. 3. Identificação. 4. Evolução. 5. Funcionalidade. I. Galante, Pedro A. F.. II. Título.

---

Bibliotecárias do Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2: Maria Lúcia Ribeiro CRB-8/2766; Stela do Nascimento Madruga CRB 8/7534.

**FOLHA DE AVALIAÇÃO**

(Esta página foi intencionalmente deixada em branco)

Dedico esse trabalho aos meus pais,

Rita de Cássia Godo e Antonio Paulo da Conceição,

à minha irmã Heloísa Cristina da Conceição e

ao meu companheiro de vida, Gustavo Martins d’Oliveira.

*“I thought I was going to get a simple experiment,*

*a simple event with a simple answer...*

*I got a most complicated answer.”*

— Barbara McClintock, in her Nobel Lecture on 8 December 1983

# AGRADECIMENTOS

Agradeço primeiramente ao meu orientador, Pedro Alexandre Favoretto Galante, pela confiança e por enxergar em mim o potencial de fazer ciência de qualidade. Agradeço principalmente por me guiar nessa jornada que parece estar apenas começando, e por ser uma inspiração e exemplo como pessoa e cientista.

Agradeço às agências de fomento à ciência, CNPq, CAPES e FAPESP (número do processo 2018/13613-4), cujo apoio é absolutamente vital para o desenvolvimento do país e que tornaram possível esse trabalho.

Agradeço à USP pela minha formação acadêmica desde o bacharelado e, em especial, ao Programa Interunidades em Bioinformática pela oportunidade de realizar o doutorado. Agradeço ao Instituto de Ensino e Pesquisa do Hospital Sírio-Libanês pela infraestrutura.

Agradeço à Eunjung Alice Lee e ao seu laboratório por me receberem em seu espaço de trabalho de forma tão genuína e pela colaboração nos projetos.

Agradeço ao técnico de informática Daniel Ohara por todo o suporte prestado e aos membros do Centro de Oncologia Molecular do Hospital Sírio-Libanês.

Agradeço aos amigos do laboratório Gabriela, Thiago, Fernanda, Felipe, Rafael, Rodrigo, Leonel, Vanessa, Filipe, Nathália, e a todos os outros pela companhia, inúmeros cafezinhos e experiências que vivemos juntos.

Agradeço às minhas queridas gatinhas Vivi, Gunther, Peridote e Zendaya e aos queridos gatinhos Cosmo, Maxwell e Nankin.

Agradeço a Gustavo pela companhia desde o primeiro dia da faculdade e pelo comprometimento que temos um ao outro. Agradeço também à sua família, Eliana, Ricardo e Lígia por me acolherem e se transformarem em minha família também.

Agradeço à Maria Auxiliadora (Dorinha), por cuidar de mim e de minha irmã durante tantos anos e pelo amor que sentimos uma pela outra.

Agradeço à minha família, às minhas avós Rosa, *in memoriam*, e Josefa, meus muitos tios e tias, primos e primas, por uma infância e uma vida muito amorosa e querida. Em especial quero honrar a memória da minha Tia Celina, minha madrinha, pelo amor incondicional.

Agradeço à minha irmã por me compreender como nenhuma outra pessoa no mundo. Agradeço aos meus pais, Rita e Antonio pelo apoio, compreensão, carinho e paciência, por toda a dedicação em construir a vida que sempre quiseram dar às suas filhas. Aos meus pais, à quem devo absolutamente tudo.

# RESUMO

Conceição, H. B. **Origens e Potencial Funcional de Retrocópias no Genoma de Humanos e Outros Animais: uma abordagem em larga escala de identificação de retrocópias em diversos genomas animais e o estudo do seu padrão de expressão em tecidos normais humanos.** 2023. 93p. Tese – Programa Interunidade em Bioinformática. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Retrocópias (também nomeadas de pseudogenes processados) são cópias de genes codificadores originadas por meio do mecanismo de duplicação mediado por RNA e são caracterizadas pela conservação apenas dos exons de seus genes parentais, a ausência de íntrons, frequente presença de cauda poliA integrada ao genoma e ausência de regiões promotoras parentais. Essas características são utilizadas para identificar retrocópias desde a década de 1980, quando muitos genes retroduplicados humanos foram reportados pela primeira vez. No entanto, a busca sistemática por retrocópias tornou-se possível apenas após o sequenciamento e montagem completos do Genoma Humano, que possibilitou o desenvolvimento de tecnologias de sequenciamento mais avançadas, melhorias na anotação do transcriptoma e o surgimento de novas ferramentas computacionais. No início dos anos 2000, as primeiras análises abrangentes para identificar retrocópias no genoma humano e em outras espécies foram realizadas, seguidas por estudos adicionais que se estendem até hoje. No entanto, a literatura sobre identificação e análise funcional de retrocópias ainda carece de análises aprofundadas e abordagens mais abrangentes. Nesta tese apresentamos uma investigação sistemática e abrangente para a identificação de retrocópias em 44 espécies, de humanos a invertebrados. Primeiro, construímos uma *pipeline* para identificar, caracterizar, organizar e disponibilizar via *web* informações sobre as 219.948 retrocópias desses 44 organismos. Todas as informações sobre posição genômica, genes parentais, tamanho das retrocópias, expressão, conservação entre espécies, entre outras, foram organizadas em um banco de dados público, a RCPedia2.0. Em um estudo complementar, investigamos o impacto e potencial funcional das retrocópias transcritas no genoma humano. Para isso, realizamos análises complexas que combinaram dados de sequenciamento de RNA de múltiplos tecidos, dados epigenéticos, e de *Ribosome Sequencing* para, primeiro, elucidar a expressão de retrocópias e como elas podem ser reguladas e depois avaliar suas

funcionalidades. Descobrimos que aproximadamente 50% (cerca de 4.000) das retrocópias presentes no genoma humano são expressas e apresentam seus níveis de expressão regulados em tecidos saudáveis. Cerca de 25% dessas retrocópias são expressas em apenas um tecido (principalmente nos testículos), enquanto que aproximadamente 15% delas são expressas em todos os tecidos humanos investigados. Nossos dados indicam que a força motriz para a expressão dessas retrocópias é a sua localização genômica próxima a genes codificadores de proteína ou a idade (mais antiga) de origem dessas retrocópias. Confirmamos ainda que um subconjunto de retrocópias é traduzido, enfatizando seu potencial funcional. Portanto, nesta tese, destacamos um segmento frequentemente ignorado no transcriptoma humano e de outras espécies: as retrocópias (ou pseudogenes processados). Não apenas revelamos o considerável potencial funcional delas e sua capacidade de gerar inovações genéticas por meio do mecanismo de retrotransposição de genes codificadores, mas também estabelecemos as bases para uma exploração abrangente e específica de cada uma dessas numerosas retrocópias.

Palavras-chave: Bioinformática, Retrocópias, Identificação, Humanos, Animais, Evolução, Expressão, Funcionalidade

# ABSTRACT

Conceição, H. B. **Origins and Functional Potential of Retrocopies in the Genome of Humans and Other Animals: a large-scale approach to identify retrocopies in diverse animal genomes and study their expression pattern in normal human tissues.** 2023. 93p. Tese – Interunity Graduate Program in Bioinformatics. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Processed pseudogenes, also known as retrocopies, are copies of coding genes originating through the RNA-mediated duplication mechanism. They are characterized by the conservation of only the exons of their parental genes, the absence of introns, frequent presence of integrated poly-A tails into the genome, and the absence of parental promoter regions. These characteristics have been used to identify retrocopies since the 1980s when many human retroduplicated genes were first reported. However, the systematic search for retrocopies became possible only after the complete sequencing and assembly of the Human Genome, enabling the development of more advanced sequencing technologies, improvements in transcriptome annotation, and the emergence of new computational tools. In the early 2000s, the first comprehensive analyses to identify retrocopies in the human genome and other species were conducted, followed by additional studies that continue to this day. However, the literature on the identification and functional analysis of retrocopies still lacks in-depth analyses and more comprehensive approaches. In this thesis, we present a systematic and comprehensive investigation for the identification of retrocopies in 44 species, ranging from humans to invertebrates. First, we constructed a pipeline to identify, characterize, organize, and make information about the 219,948 retrocopies of these 44 organisms available via the web. All information, including genomic position, parental genes, retrocopy size, expression, conservation between species, among others, was organized in a public database, RCPedia 2.0. In a complementary study, we investigated the impact and functional potential of transcribed retrocopies in the human genome. For this, we conducted complex analyses combining RNA sequencing data from multiple tissues, epigenetic data, and Ribosome Sequencing to first elucidate retrocopy expression and how they may be regulated, and then evaluate their functionalities. We found that approximately 50% (around 4,000) of retrocopies in the human genome are expressed and have their expression levels regulated in healthy tissues. About 25%

of these retrocopies are expressed in only one tissue (mainly in the testes), while approximately 15% of them are expressed in all investigated human tissues. Our data indicate that the driving force for the expression of these retrocopies is their genomic proximity to protein-coding genes or the (older) age of origin of these retrocopies. We further confirmed that a subset of retrocopies is translated, emphasizing their functional potential. Therefore, in this thesis, we highlight a frequently overlooked segment in the human and other species' transcriptome: retrocopies (or processed pseudogenes). We not only reveal their considerable functional potential and their ability to generate genetic innovations through the mechanism of retrotransposition of coding genes but also lay the groundwork for a comprehensive and specific exploration of each of these numerous retrocopies.

Keywords: Bioinformatics, Retrocopies, Identification, Humans, Animals, Evolution, Expression, Functionality

LISTA DE FIGURAS

LISTA DE ABREVIAÇÕES

| | |
|---|---|
| **5SrRNA** | 5S ribosomal RNA |
| **Ac** | Lócus Ativador |
| **APE** | Apurinic–apyrimidinic endonuclease |
| **BLAST** | Basic Local Alignment Search Tool |
| **CDS** | Coding DNA Sequence |
| **DDBJ** | DNA Data Bank of Japan |
| **DHFR** | Dihidrofolato redutase |
| **DNA** | Ácido Desoxirribonucleico |
| **Ds** | Lócus de Dissociação |
| **Ka** | Taxa de substituições não-sinônimas por sítio |
| **Ks** | Taxa de substituições sinônimas por sítio |
| **LINE-1** | Long interspersed nuclear elements - 1 |
| **LTR** | Long Terminal Repeat |
| **mRNA** | Molécula de Ácido Ribonucleico Mensageira |
| **MT** | Metalotioneína |
| **NCBI** | National Center for Biotechnology Information |
| **ncRNAs** | RNAs não codificantes |
| **NRDB** | Nonredundant Protein Database |
| **ORFs** | Quadros de leitura abertos traduzidos |
| **psi-alpha1** | Pseudogene de alfa globina humana |
| **RCPedia** | Retrocopy Encyclopedia |
| **RefSeq** | NCBI Reference Sequences |
| **RiboSeq** | Sequenciamento de fragmentos processados por ribossomos |
| **RNA** | Ácido Ribonucleico |
| **RPFs** | Fragmentos processados por ribossomos |
| **RPKM** | Reads per kilobase per million |
| **RTC** | Retrocópia |
| **SQL** | Structured Query Language |
| **TrEMBL** | Translated EMBL Nucleotide Sequence Data Library |
| **TRTP** | Target-primed reverse transcription |
| **UCSC** | Universidade da Califórnia, Santa Cruz |

# SUMÁRIO

# INTRODUÇÃO

## Elementos Genéticos Móveis

O estabelecimento dos fundamentos da genética evolutiva na primeira metade do século XX só foi possível graças à redescoberta dos experimentos de Mendel na década de 1900 à luz da teoria da evolução, de Charles Darwin e Alfred Russel Wallace, amplamente aceita pela comunidade científica. Ao longo das décadas de 1930 a 1950, o avanço técnico-experimental consolidou as bases do campo da biologia evolutiva, quando um amplo consenso a respeito dos mecanismos que produzem mudanças evolutivas foi estabelecido. Essa síntese moderna da teoria da evolução, a qual incorpora os princípios da genética, é conhecida como a teoria sintética da evolução, síntese moderna da evolução ou neodarwinismo[1].

Dentre os muitos cientistas que contribuíram para o avanço da genética nesse período, a citogeneticista Barbara McClintock se destaca como uma figura importante para a elucidação dos mecanismos de controle e regulação genética. A Dra. McClintock foi pioneira nos estudos de citogenética na década de 1930, mudando radicalmente a forma como a comunidade científica entendia os padrões genéticos de herança devido à sua descoberta de elementos transponíveis. Ela também foi a primeira cientista a especular - corretamente - o conceito de epigenética – 40 anos antes de o conceito ser formalmente estudado. Ao longo de sua carreira, McClintock utilizou a espécie modelo *Zea mays* (milho) que produz centenas de descendentes em uma única espiga, onde cada grão de milho é um embrião produzido a partir de uma fertilização individual e, portanto, com fenótipos únicos, por exemplo, de cores variadas.

Em 1931, McClintock e Harriet Creighton forneceram a primeira prova experimental de que os genes estavam fisicamente posicionados nos cromossomos, descrevendo o fenômeno do cruzamento e da recombinação genética[2]. As técnicas citogenéticas inovadoras de McClintock permitiram confirmar as ideias propostas por Thomas Hunt Morgan de que traços genéticos passados aos descendentes dependem da troca de material genético entre os cromossomos[3].

Seminalmente, na década de 1940, Barbara McClintock descobriu que a informação genética poderia "se mover" para diferentes locais nos cromossomos,

afetando o comportamento dos genes vizinhos[4]. Essa descoberta foi fruto de experimentos sistemáticos de reprodução envolvendo um fenótipo incomum do milho: a cor arroxeada e muitas vezes fragmentada de alguns grãos de milho, versos a cor “padrão” amarelo claro.

Ao rastrear as alterações de pigmentação no milho e usar microscopia óptica para examinar os cromossomos da planta, McClintock descobriu o que ficou conhecido como sistema Ac/Ds no milho, dois novos *loci* genéticos dominantes e interativos que chamou de Lócus de Dissociação (Ds) e Lócus Ativador (Ac). McClintock não só descobriu que o *locus* do Ds causava a quebra do cromossomo, mas também tinha uma variedade de efeitos nos genes vizinhos quando o Ac também estava presente, controlando os genes que eram realmente responsáveis pela pigmentação do grão de milho. No início de 1948, ela fez a surpreendente descoberta de que tanto o Ds como o Ac poderiam mudar de posição nos cromossomos de cada espécime. McClintock publicou seus achados em um artigo de 1950 intitulado “*The origin and behavior of mutable loci in maize*” e só mais tarde esses *loci* foram classificados como elementos genéticos móveis ou transponíveis. As suas descobertas a respeito dos elementos transponíveis foram aceitas pelos geneticistas que estudavam milho, mas a natureza generalizada desses elementos genéticos móveis e as suas implicações estavam à frente do seu tempo e, durante muitos anos, foram consideradas demasiado radicais pela comunidade científica.

Foi apenas no final dos anos 1960 e 1970, após um grande desenvolvimento do conhecimento sobre a biologia molecular e de novas metodologias de análises que a descoberta de Barbara McClintock foi verificada pelos seus pares como ocorrendo em outros organismos, como vírus e bactérias. Mais tarde, a confirmação de que os transposons eram difundidos entre os eucariotos concedeu reconhecimento ao seu trabalho, levando-a a receber o Prêmio Nobel de Fisiologia ou Medicina de 1983 em reconhecimento a esta e a muitas outras contribuições que ela dera para o campo da genética. O trabalho de McClintock foi revolucionário na medida em que sugeriu que o genoma de um organismo não é uma entidade estacionária, mas está sujeito a alterações e rearranjos. Essa ideia ainda hoje surpreende uma parcela da comunidade científica, destacando a relevância contínua e impactante de suas descobertas.

**Figura 1: Retrato Fotográfico de Barbara McClintock em 1947.** Ela está segurando uma espiga de milho com grãos multicoloridos. Fonte: Barbara McClintock Papers, American Philosophical Society.

## Pseudogenes e o surgimento de novos genes.

O subsequente salto do conhecimento ocorrido entre 1940 e 1980 foi revolucionário para a genética: por exemplo, estabeleceram-se as etapas da replicação de DNA, da transcrição do DNA para RNA mensageiro (mRNA), do código genético e da tradução do mRNA em aminoácidos. Com isso, os genes passaram de conceitos abstratos a entidades moleculares discretas que podiam ser medidas e manipuladas. Adicionalmente, o reconhecimento de que os transposons estão presentes de forma ubíqua (e em grande frequência) nas espécies preparou o terreno para a descoberta das retrocópias ou, como foram inicialmente denominadas, pseudogenes processados.

Desde a criação do termo pseudogene, na década de 1970, sua definição se ampliou e é amplamente aceita para denominar qualquer sequência genômica semelhante a outro gene e defeituosa[5]. Existem duas principais classes de pseudogenes: processados e não processados. Essas classes são definidas pelos mecanismos de origem do pseudogene em questão. Pseudogenes processados (retrocópias) são derivados da retrotransposição de RNAs mensageiros

processados[6,7] enquanto que os pseudogenes não processados são derivados da duplicação segmentar do DNA genômico[8].

A distinção binária entre genes e pseudogenes constitui um tema central na anotação do genoma e, em última análise, é a base para a construção de anotações de referência de genes de um organismo[9]. No entanto, rotular certas regiões genômicas como pseudogenes muitas vezes resulta na sua exclusão ampla de análises genômicas, transcriptômicas e funcionais. Essa exclusão é motivada pela suposição de que tais regiões carecem da capacidade de gerar impactos biológicos significativos. Entretanto, temos evidência justamente do contrário: a formação de pseudogenes processados está em andamento na evolução humana com pelo menos 48 cópias de genes retrotranspostos (retrocópias) que são polimórficas na população humana[10–13].

Apesar da definição de função biológica ser um conceito complexo e muitas vezes controverso em biologia[14], temos um número crescente de casos de regiões anotadas como pseudogenes que posteriormente se descobriu que exibem função biológica. Em adição, no genoma de metazoários são normalmente anotados entre 10.000 e 20.000 regiões como pseudogênicas[15]. Embora muitos pseudogenes possam não ter uma função detectável que atualmente afete a aptidão celular do organismo, sua existência pode desempenhar um papel evolutivo significativo[9]. Um exemplo evidente de papel evolutivo de longo prazo é a redundância genética, onde uma determinada função bioquímica é codificada de forma redundante por dois ou mais genes[16].

O surgimento de novos genes constitui um processo fundamental para a aquisição de novos traços fenotípicos, muitas vezes específicos de linhagens ou espécies[8]. Pelo menos desde a publicação de *“Evolution by Gene Duplication”*, de Susumu Ohno[17,18], a sabedoria convencional tem sido de que novos genes geralmente surgem pela duplicação de genes existentes. Vale ressaltar que a ideia de que a duplicação desempenha um papel destacado no surgimento de novos genes remonta aos primórdios da síntese evolutiva moderna[19–21]. Décadas de estudos moleculares em grande e pequena escala consolidaram a duplicação gênica como a força motriz da origem de inovações evolutivas[22].

Mecanisticamente, várias vias de surgimento de genes foram documentadas. Entre essas vias, destacam-se: i) a origem de novos genes "*ab initio*", a partir de

regiões nunca antes transcritas; ii) a fusão ou "fissão" de genes existentes; iii) a transferência horizontal de genes entre espécies; e iv) a duplicação de genes já existentes. É inegável que a duplicação gênica prevalece como o mecanismo principal em humanos, mamíferos e vertebrados[23]. A duplicação gênica e a formação de pseudogenes estão intrinsecamente interligadas em diversos contextos. Uma vez duplicado, o gene adicional pode seguir diferentes trajetórias evolutivas. Pode adquirir uma nova função, processo conhecido como neofuncionalização, compartilhar a função original com o gene ancestral, resultando em sub funcionalização, ou perder sua função original e evoluir para um pseudogene.

Esta complexa dinâmica evolutiva sublinha a plasticidade genômica e a variedade de caminhos que os genes podem trilhar. Assim, a compreensão dos pseudogenes e sua potencial funcionalidade é essencial para uma visão abrangente da evolução genética, desafiando as premissas anteriores e destacando a complexidade das interações genéticas ao longo do tempo.

## Descoberta das Retrocópias

Em 1977, Jacq e colaboradores cunharam o termo “pseudogene” pela primeira vez para descrever um fragmento de DNA de *Xenopus laevis* que estava truncado na extremidade 5' e apresentava algumas mutações quando comparado ao gene funcional 5S rRNA[24]. Em 1980, Proudfoot e Maniatis publicaram a sequência completa de nucleotídeos de um pseudogene de alfa globina humana (psi-alpha1)[25], o primeiro relato e descrição registrados de tal elemento no genoma humano.

Posteriormente, em 1982, Ueda e colegas[10] relataram, pela primeira vez em humanos, uma nova categoria de pseudogenes: os pseudogenes processados (retrocópias). Esta designação foi atribuída devido à identificação dos autores de que a sequência do pseudogene de imunoglobulina épsilon C3 carece de três sequências inteiras (íntrons) em comparação com o gene funcional. Além disso, apresenta uma sequência adicional rica em A com 31 bases de comprimento na extremidade 3', juntamente com sequências semelhantes a Long Terminal Repeats (LTR) nas regiões flanqueadoras 5' e 3'. No mesmo ano, Wilde, Crowther e Cowan

também publicaram a sequência de três pseudogenes de beta-tubulina humana[26], sendo um deles com íntrons, mas sem uma sequência de codificação terminal e os outros dois sem íntrons, apresentando uma cadeia terminal poli(A) adicional.

Ambos os trabalhos de 1982 discutem e apresentam hipóteses sobre o mecanismo de origem dessa nova classe de pseudogenes, os pseudogenes processados. Ueda e colaboradores descreveram corretamente o reconhecimento da cauda poli(A) de um RNA mensageiro maduro para sua transcrição reversa e integração. Wilde, Crowther e Cowan[26] propuseram que o mecanismo de transcrição reversa de um RNA mensageiro processado deve ser responsável pela origem de sequências poliadeniladas sem íntrons. Na época, ainda não se tinha conhecimento da extensão desses eventos em vertebrados superiores, e não havia evidências claras da existência de uma transcriptase reversa funcional no genoma humano.

Em estudos sobre a família de genes da metalotioneína em humanos, Karin e Richards[27] encontraram mais um exemplo de pseudogene processado: eles publicaram a sequência do gene II A da metalotioneína (MT-IIA) e seu pseudogene, MT-IIB. Eles também observaram variações na população examinada, com alguns indivíduos apresentando uma cópia completa do pseudogene, enquanto outros tinham um fragmento com cerca de metade do tamanho da cópia inteira. Os autores concluíram corretamente que deveria existir um fluxo reverso de informação do RNA para o DNA ocorrendo em células germinativas de mamíferos, representando o primeiro registro de polimorfismo de pseudogenes processados na população humana[27].

No final de 1982, Lemischka e Sharp[28] relataram a descoberta de um pseudogene alfa-tubulina em ratos e foram mais assertivos ao declarar que tal sequência era proveniente de uma cópia do RNA mensageiro maduro transcrito do gene funcional da alfa-tubulina. Essa descoberta marca uma mudança no discurso da comunidade científica, tornando claro que não apenas os pseudogenes processados eram comuns, mas que o mecanismo de origem de tais sequências envolvia inequivocamente o fluxo de informação do RNA para o DNA.

Além disso, todos os artigos acima mencionam a semelhança entre os elementos repetitivos flanqueando os pseudogenes processados (retrocópias) e os elementos *Alu*. Lemischka e Sharp argumentaram corretamente que tanto os

elementos *Alu* quanto os pseudogenes processados compartilham um mecanismo de inserção que envolve (retro)copiar o mRNA para o DNA, resultando em repetições diretas na inserção. A preservação da orientação do mRNA e a necessidade de uma sequência genômica rica em adenina na extremidade 3' do RNA foram igualmente destacadas.

Em 1984 e 1985, Vanin[5] publicou as primeiras revisões sobre os pseudogenes processados, apontando cinco características comuns entre eles: (i) várias “lesões” genéticas resultando em códons de parada prematuros ou mudanças de quadros de leitura; (ii) ausência de íntrons; (iii) presença de cauda poli(A) na extremidade 3' integrada na região genômica; (iv) homologia correspondente ao início e ao final do transcrito do gene funcional e (v) presença de repetições diretas de 7-17 pares de bases flanqueando o pseudogene processado. Embora na época (uma era pré-sequenciamento do genoma humano) apenas alguns pseudogenes processados houvessem sido identificados, Vanin ressaltou que a maioria deles não compartilhava o mesmo cromossomo com seu gene funcional (parental) correspondente, o que também fora observado nos elementos *Alu*, sugerindo um mecanismo de origem comum. A cauda poli(A) foi reconhecida como parte desse mecanismo, emparelhando-se com sequências repetitivas (T/A)n no genoma para iniciar o processo de integração. Vanin propôs que todos os pseudogenes processados tinham origem na linha germinativa e que a transcrição pela RNA polimerase II ou III podia gerar pseudogenes processados, explicando as inserções com características diferentes. Embora inicialmente os pseudogenes processados fossem identificados principalmente em mamíferos, alguns eram tão recentes que eram polimórficos, como o pseudogene 1 de DHFR e o pseudogene II B processado pela metalotioneína.

Nos anos seguintes, mais pseudogenes processados foram identificados em humanos, e hipóteses adicionais foram propostas. A partir dessas e outras publicações, fica claro para a comunidade científica que a existência de tais duplicatas gênicas não era uma ocorrência rara no genoma humano.

## Mecanismo de Retroduplicação

Atualmente sabemos que a maquinaria enzimática necessária para a transcrição reversa do RNA mensageiro e a sua subsequente integração no genoma é fornecida majoritariamente por elementos transponíveis pertencentes às superfamílias de retrotransposons não-LTRs (do Inglês, *Long Terminal Repeats*). Esses elementos codificam uma transcriptase reversa[29] e uma endonuclease/integrase[30], responsáveis por mediar a inserção de retrocópias no genoma. Em diversos mamíferos, a maquinaria de retrotransposição é codificada a partir de elementos LINE-1 (do Inglês, *Long interspersed nuclear elements*), que representam um conjunto de retrotransposons bem variado entre as espécies e o mais abundante em humanos.

Os LINEs, sendo elementos retrotransposons não-LTR, se replicam através do mecanismo conhecido como transcrição reversa anelada por *primer* (do Inglês, *target-primed reverse transcription* - TPRT). Esse processo se inicia com a clivagem do DNA alvo pela atividade de endonuclease apurínica/apirimidínica (apurinic–apyrimidinic endonuclease - APE) codificada pelo próprio LINE-1 e que reconhece diretamente a sequência do sítio alvo[30]. A clivagem expõe um grupo 3’ hydroxyl que serve como âncora para a transcrição reversa local do LINE-1. Estudos cristalográficos detalhados e de mutagênese pontual revelaram ainda que uma alça β-hairpin variável que se projeta do sítio de ligação ao DNA dos APEs entra em contato com o sulco menor do DNA e participa no reconhecimento do ponto de clivagem[31–34]. Durante a TPRT, o anelamento da transcriptase reversa é facilitada pelo pareamento de bases entre o sítio alvo do DNA e a cauda poli(A) do RNA mensageiro e, consequentemente, apenas regiões ricas em T são usadas de forma eficiente durante a TPRT do LINE1[34]. Outra via de inserção independente da atividade de endonuclease (ENi) também já foi reportada, onde a transcrição reversa é iniciada em uma lesão de DNA pré-existente, sem a necessidade de clivagem[35,36]. Entretanto, independente da via de iniciação da retrotransposição, os passos subsequentes de resolução da reação, tais como a síntese da segunda fita de DNA ou a ligação da extremidade 3’ do DNA recém sintetizado ao DNA alvo, carecem de um modelo com evidências robustas na literatura.

Em 1988, Begg[37] e colaboradores mostraram pela primeira vez que existia uma conexão entre pseudogenes processados e a maquinaria LINE-1 em camundongos, observando, adicionalmente, que genomas de eucariotos superiores estão em constante evolução devido a mutações, duplicações e também integração de sequências de DNA a partir de RNA. Posteriormente, Esnault *et al.[38]* demonstraram que a retrotransposição de RNAs mensageiros de genes codificadores em humanos ocorre também por meio da maquinaria dos LINE1. E que, portanto, a inserção das retrocópias (pseudogenes processados) é consequência do redirecionamento *in trans* da maquinaria enzimática de LINE-1 para o processo de retrotransposição de RNAs mensageiros.

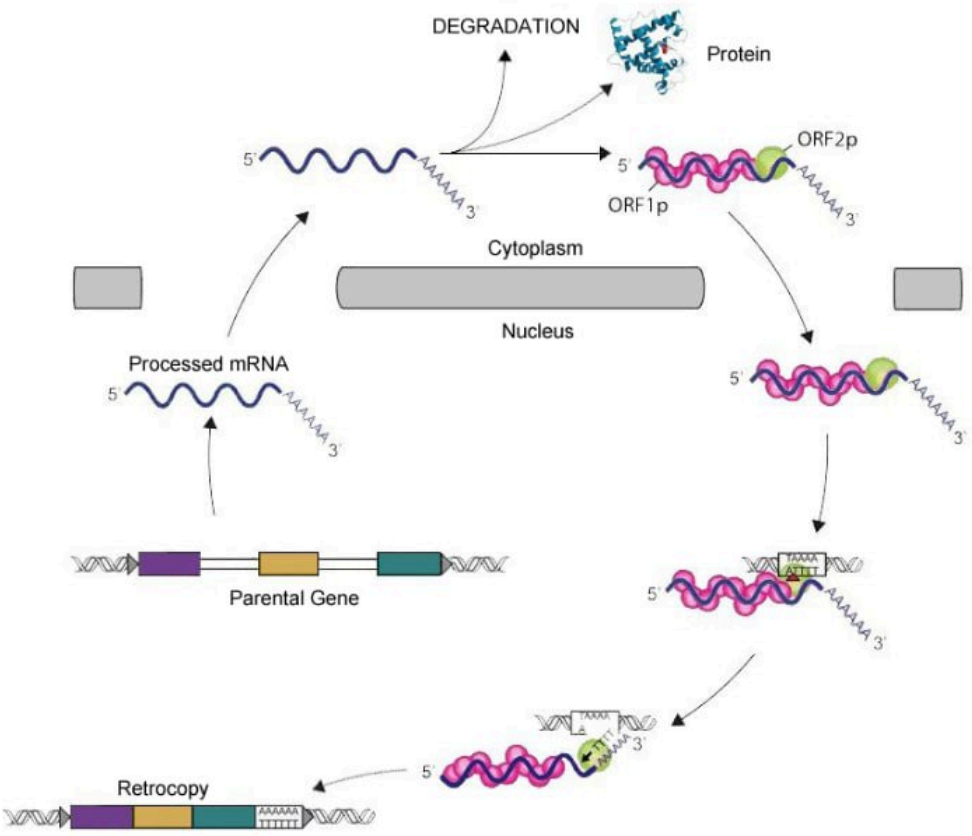



**Figura 2: Mecanismo de Retrotransposição de Transcritos de Genes Codificadores.** O processo de retrotransposição é mediado pela maquinaria enzimática de LINE-1, resultando na inserção de retrocópias em um novo *locus* genômico. Figura adaptada de Viollet *et al.* 2014.

# Capítulo 01 - Retrocópias no Genoma de Animais: uma abordagem em larga escala para a identificação de retrocópias

*Retrocopies in the Animal Genome: a large-scale approach to retrocopy identification*

# Resumo

Neste capítulo, mostramos o desenvolvimento e o uso de uma nova *pipeline* para a identificação sistemática de retrocópias, a qual pode ser aplicada a diferentes espécies. De forma resumida, essa *pipeline* alinha o conjunto completo de sequências codificadoras de mRNA ao genoma de referência da espécie em questão, seleciona alinhamentos com base no tamanho e distância do gene que originou o mRNA (gene parental), busca por junções éxon-éxon oriundas da extremidade mais 3' do mRNA e elimina candidatos compostos majoritariamente por elementos repetitivos. A *pipeline* foi aplicada na busca de retrocópias em 44 espécies, desde primatas a invertebrados. Isso resultou em um catálogo completo de 219.948 retrocópias e 73.569 genes parentais, cujo resultado foi organizado em um banco de dados para uma atualização da RCPedia, que passamos a chamar de RCPedia 2.0. Para explorar a história evolutiva das retrocópias, também desenvolvemos um algoritmo para determinar a ortologia dessas retrocópias entre si. Neste algoritmo, em resumo, recuperamos a sequência genômica em torno de cada retrocópia (3.000 pb a montante e a jusante) e realizamos o alinhamento par a par entre as retrocópias das diferentes espécies usando o alinhador Lastz. Posteriormente, utilizamos filtros de cobertura e identidade de cada alinhamento para determinar a melhor correspondência para cada retrocópia. Descobrimos que, em média, 30% das retrocópias identificadas são conservadas na maioria dos vertebrados (mamíferos e aves). Especificamente, os humanos têm 96,5% de suas retrocópias conservadas com outras espécies, sobretudo primatas (90,11%). Em resumo, desenvolvemos uma abordagem nova e abrangente para identificar retrocópias e, além disso, realizamos uma contribuição importante para o estudo da história evolutiva destas espécies. Esta é uma melhoria substancial, para humanos e outros mamíferos, no conhecimento destas cópias genéticas pouco estudadas.

## 1.1 INTRODUÇÃO

Após a conclusão do sequenciamento e montagem do Genoma Humano [39], bem como o avanço das tecnologias de nova geração[40] e o desenvolvimento de alinhadores de sequências mais eficazes[41], tornou-se possível investigar retrocópias (pseudogenes processados) de forma mais precisa, sistemática e em larga escala. Desde a década de 1980, as características típicas desses elementos têm sido instrumentais na identificação de retrocópias. Essas características incluem a ausência de íntrons, uma cauda de poliadenilação na extremidade 3' da sequência e uma localização distinta no genoma em relação ao gene de origem (gene parental).

A primeira busca sistemática por pseudogenes no genoma humano foi realizada por Harrison et al.[42] em artigo publicado em 2002 após o sequenciamento completo dos cromossomos 21 e 22. O estudo apresenta um método inicial para identificar e caracterizar pseudogenes no genoma humano, partindo do alinhamento de sequências de aminoácidos derivadas do banco de dados SWISSPROT contra a sequência genômica dos cromossomos 21 e 22 humanos usando o alinhador BLAST (Basic Local Alignment Search Tool)[43] para encontrar regiões genômicas com semelhanças às sequências de proteínas conhecidas e com evidências claras de interrupção funcional (como códons de parada ou mudanças de quadro de leitura), minimizando a sobreposição com anotações de genes conhecidos. Os pseudogenes foram classificados em "processados" (derivados de mRNA, sem estrutura de íntrons; isto é, retrocópias) e "não processados" (originados de duplicações de DNA genômico). Essa busca sistemática resultou na descoberta de 454 anotações de pseudogenes, sendo 189 pseudogenes processados (retrocópias), 195 pseudogenes não processados e 70 pseudogenes de imunoglobulinas. A partir dessas anotações, os autores extrapolaram que poderia haver até 20.000 pseudogenes em todo o genoma humano, com pouco mais da metade sendo processados. A principal motivação deste estudo foi enfatizar a importância de caracterizar as populações de pseudogenes processados e não processados para a identificação de novos genes e estudos evolutivos. Não apenas essa é a primeira busca sistemática por pseudogenes, como os autores descrevem uma metodologia de busca baseada em

alinhamentos de sequências codificadoras que também foi utilizada por outros estudos subsequentes.

No mesmo ano, o mesmo grupo de pesquisa (Dr. Mark Gerstein de Yale) [44] realizou a busca específica por pseudogenes processados (retrocópias) de genes codificadores de proteínas ribossomais com base na identidade de sequências, mas agora em todos os cromossomos completamente sequenciados pelo projeto Genoma Humano. As sequências de aminoácidos de 79 proteínas ribossomais foram extraídas do banco de dados SWISSPROT e alinhadas contra a versão de 06 de Agosto de 2001 do genoma humano com o alinhador BLAST. Foram encontrados um total de 2.090 pseudogenes processados e 16 duplicações de genes de proteínas ribossomais. Foram observados que 258 deles não apresentavam desativações óbvias (códons de parada ou mudanças de quadro de leitura), enquanto que 178 eram interrompidos por elementos repetitivos. Os autores observaram que, em média, os pseudogenes processados apresentaram uma truncagem maior na extremidade 5' do que na 3', consistente com o mecanismo de transcrição reversa iniciada por *primers* direcionados ao alvo (TPRT), e que a distribuição dos eventos pelo genoma parece resultar principalmente de inserções aleatórias, proporcionalmente ao tamanho de cada cromossomo humano.

Logo em seguida, em 2003, dois trabalhos de Zhang *et al.[45]* (também do grupo de Gerstein) e Torrents *et al.[46]* realizaram a busca por pseudogenes em todos os cromossomos humanos. No artigo de Zhang e colaboradores, a busca concentrou-se na identificação apenas de pseudogenes processados. O estudo desenvolveu uma *pipeline* baseada na busca por sequências com alta similaridade com proteínas de humanos conhecidas pelo SWISS-PROT ou TrEMBL (Translated EMBL Nucleotide Sequence Data Library) usando a ferramenta BLAST, com identidade superior a 40%, lacunas (*gaps*) inferiores a 60 pares de base, cobertura de região codificadora (*coding DNA sequence* - CDS) superior a 70% e presença de *frameshifts* ou *in-frame stop codons*. Usando essa abordagem, foram identificados 7.819 pseudogenes processados (no genoma humano) e o resultado foi disponibilizado publicamente no repositório http://pseudogene.org, o primeiro banco de dados do tipo. Foi constatado que os pseudogenes processados se assemelham aos seus genes parentais correspondentes, sendo 94% completos em regiões codificadoras, com 75% de similaridade na sequência de aminoácidos e 86% de

similaridade na sequência de nucleotídeos. A distribuição deles pelos cromossomos, assim como no artigo de pseudogenes de proteínas ribossomais, segue um padrão aleatório e disperso, com números aproximadamente proporcionais ao comprimento cromossomal, sugerindo um "bombardeio" de inserções de retrocópias no nosso genoma. Também observou-se que os pseudogenes processados seguem uma relação de lei de potência em relação à sua associação com genes parentais, com alguns poucos genes sendo altamente retroduplicados e a maioria dos genes gerando poucos ou nenhuma retrocópia.

No artigo de Torrents e colaboradores[46] foi desenvolvida uma *pipeline* de identificação de pseudogenes que combina buscas de homologia e um teste de funcionalidade baseado na razão de substituições Ka/Ks (ou dN/dS). A razão Ka (taxa de substituições não-sinônimas por sítio) / Ks (taxa de substituições sinônimas por sítio) é uma métrica amplamente utilizada em genética evolutiva para inferir a natureza da seleção natural atuando sobre um gene ou pseudogene. Os autores utilizaram a sequência de DNA do genoma humano (build30) do NCBI (ftp://ftp.ncbi.nih.gov/genomes/) já mascarada para repetições humanas comuns e excluindo todas as regiões cobertas pelos genes conhecidos e preditos no banco de dados ENSEMBL. Todos os fragmentos não "mascarados" (marcados para não alinharem) e maiores que 100nt foram alinhados contra um banco de dados de proteínas não redundantes (*nonredundant protein database - NRDB*) usando o alinhador BLASTX. Todas as regiões semelhantes a transposons ou proteínas virais foram descartadas. Foram identificadas 19.724 regiões pseudogênicas, das quais 95% foram estimadas como evoluindo de maneira neutra. Uma análise comparativa com o genoma de camundongos mostrou que 70% desses pseudogenes têm uma origem retrotransposicional (processada), e o restante surgiu por duplicação segmentar (não processada).

Em 2005, Khelifi *et al.[47]* desenvolveram um método para anotar pseudogenes processados em genomas completamente sequenciados, armazenando o resultado no primeiro banco de dados dedicado aos pseudogenes processados (retrocópias), o HOPPSIGEN. Neste trabalho, os autores examinaram os genomas completos de camundongos e humanos para encontrar pseudogenes processados gerados a partir de genes funcionais com íntrons. Para cada espécie, alinhou-se ao genoma mascarado (com relação aos elementos repetitivos) as

regiões de CDS de genes com íntrons do ENSEMBL 18.3 (27.156 CDS humanos e 28.696 CDS de camundongo) usando o alinhador TBLASTX para encontrar sequências de DNA semelhantes à CDS funcional. Correspondências positivas (>80 pares de base) foram mantidas, sem limite de similaridade. Os autores designaram os pseudogenes processados como sendo todas as sequências transcritas reversamente e, mais estritamente, que não apresentam um quadro de leitura aberto conservado ou que se sobreponham a genes anotados. Os autores reportaram 5.206 pseudogenes processados no genoma humano e 3.428 pseudogenes processados no genoma de camundongo. Em comparação com estimativas anteriores, o número total de pseudogenes processados foi inferior (subestimado), visto os filtros mais estritos aplicados por estes autores.

Até 2007, todos os artigos publicados de buscas por pseudogenes processados se baseavam em alinhamentos de sequências de aminoácidos contra os genomas de referência, mas na publicação de Sakai et al.[48] pela primeira vez foram alinhadas as sequências nucleotídicas de RNAs mensageiros contra o genoma de referência. A coleção de sequências de mRNA usada neste estudo consistiu em 113.196 sequências de mRNA humano e 101.096 sequências de mRNA de camundongo depositadas na versão 54 do DDBJ (*DNA Data Bank of Japan[49]*). Sequências repetitivas e de baixa complexidade foram mascaradas pelo RepeatMasker com Repbase 7.5. O mapeamento foi realizado usando o BLASTN 2.2.6 para todas as sequências de mRNA contra as sequências do genoma (NCBI build 34 para humanos e build 30 para camundongos). Como resultado, os autores identificaram 7.348 pseudogenes processados em humanos e 6.188 no caso de camundongos.

Nos anos seguintes, a busca por retrocópias em várias espécies foi relatada pelos trabalhos de Yu *et al.* 2007[50], Liu *et al.* 2009[51] e Balasudramanian *et al.* 2009[52], com um leque mais amplo de espécies, entretanto com o número de retrocópias reportadas muito discrepantes com relação à literatura. Não apenas isso, o número de espécies analisadas ainda estava limitado e, em alguns casos, a variabilidade das espécies também implica em uma grande distância evolutiva entre si, tornando a determinação da ortologia e história evolutiva das retrocópias um grande desafio para os artigos publicados até então.

Em 2013, Navarro e Galante[53] publicam a RCPedia (*Retrocopy Encyclopedia*), o primeiro banco de dados dedicado às retrocópias de primatas com o intuito de ressaltar a relevância de retrocópias na evolução de primatas. Para identificar as retrocópias da RCPedia, os autores aplicaram um método de busca baseado no alinhamento de mRNAs anotados pelo RefSeq (*NCBI Reference Sequences*)[54] contra o genoma de referência[55] utilizando o alinhador BLAT[56]. O resultado foi filtrado pela qualidade (*score* superior a 100 e identidade superior a 75%), sendo que apenas os alinhamentos com lacunas de alinhamento inferiores a 15.000 bases foram mantidos. Por fim, os autores verificaram a existência de bordas exon-exon adjacentes na região da retrocópia candidata. Caso o alinhamento passasse por esses filtros, a região genômica correspondente foi considerada uma retrocópia. Como o algoritmo de identificação foi aplicado à seis espécies de primatas com genomas e transcriptomas de tamanhos muito similares, além de composição similar em relação aos elementos repetitivos identificados pelo *Repeatmasker*, os resultados mostraram uma composição relativamente homogênea de retrocópias por espécie. Esse resultado possibilitou aos autores implementarem um método para determinar a ortologia entre as retrocópias baseado em sintenia de seus *loci* genômicos. Assim, foi possível estimar a taxa de origem e fixação de retrocópias durante a evolução de primatas[53].

Por último, no âmbito dos bancos de dados dedicados às retrocópias, vale mencionar o RetrogeneDB, cuja última publicação data de 2014[57], com uma atualização em 2017[58]. O estudo de 2014 abrangeu uma análise de genomas de 62 espécies de animais, com o propósito de identificar retrocópias. A busca foi baseada na comparação entre a sequência genômica de referência e as proteínas codificadas por genes multiexons em uma espécie específica. Diversos critérios foram aplicados para identificar uma região genômica como retrocópia, incluindo um alinhamento com pelo menos 150 pares de bases, uma cobertura mínima de 50% do gene parental, uma identidade mínima de 50% e a perda de pelo menos dois íntrons, entre outros critérios. No total, os autores identificaram 84.808 retrocópias, incluindo 6.277 genes codificadores de proteínas que não haviam sido previamente reconhecidos como retrogenes. Entretanto, os critérios aplicados resultaram em um número de retrocópias consideravelmente menor em cada espécie do que em comparação aos outros artigos publicados.

É importante evidenciar que no começo dos anos 2000 havia uma grande preocupação com a anotação correta de genes codificadores no genoma de humanos, e que, neste contexto, a importância de encontrar pseudogenes servia ao propósito de remover "falsos genes" das anotações genômicas. Nota-se, por exemplo, que na busca por pseudogenes era de grande importância a identificação de *stop-codons*, eventos de *frameshifts* e taxas neutras de substituições (Ka/Ks). Considerando-se o sentido da palavra "pseudogene" como uma cópia não funcional de um gene codificador, exigir a perda de funcionalidade da sequência seria uma imposição justa. Entretanto, se estamos interessados em descobrir sequências oriundas de duplicações ou retrotransposição e entender o impacto dessas sequências na história evolutiva das espécies e como elas podem vir a ganhar funcionalidade ou mesmo serem funcionais sem uma região codificadora, aplicar filtros desse tipo traz graves limitações ou erros. Outro fator a se notar é que a maioria das metodologias se baseiam na busca a partir das sequências de aminoácidos, inclusive o banco de retrocópias RetrogeneDB. Entretanto, o alinhamento a partir do RNA mensageiro maduro de um gene codificador reflete melhor o mecanismo de origem das retrocópias.

Por fim, nota-se que vários avanços foram realizados desde o primeiro sequenciamento completo do Genoma Humano: como a criação de alinhadores mais sensíveis e precisos, melhores anotações dos genes (por exemplo, o projeto GENCODE) e de elementos repetitivos (*RepeatMasker* e outro bancos), melhores montagens dos genomas de referência e maior número de espécies sequenciadas. Além disso, vale ressaltar que até hoje só foram publicados quatro grandes bancos de dados referentes à retrocópias (ou pseudogenes processados): Pseudogene.org (este, não dedicado a pseudogenes processados), HOPPSIGEN (atualmente desativado), RCPedia (focada somente em primatas) e RetrogenesDB (sub representado em relação a retrocópias). Portanto, fica claro que aprimoramentos na metodologia de busca de retrocópias precisam ser realizados para acompanhar tais melhorias e atualizar os bancos de dados existentes.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

Desenvolver uma *pipeline* eficiente e precisa para a identificação e caracterização de retrocópias (fixadas) em diversas espécies.

### 1.2.2 Objetivos Específicos

- Produzir uma *pipeline* para a identificação de retrocópias fixadas no genoma de diversas espécies animais;
- Aplicar a *pipeline* produzida para identificar retrocópias fixadas no genoma de 44 espécies de animais;
- Caracterizar os genes parentais, a extensão da região retrocopiada, a identidade entre retrocópias e genes parentais.
- Determinar a ortologia das retrocópias entre espécies e reportar a conservação entre as espécies;
- Gerar uma anotação genômica confiável para a quantificação de expressão das retrocópias para as espécies aqui analisadas;
- Organizar todas as informações obtidas em um banco de dados completo de retrocópias e seus genes parentais para uso público via interface *web*.

## 1.3 MATERIAIS & MÉTODOS

### 1.3.1. Identificação de Retrocópias Fixadas

Com o intuito de melhorar a anotação das retrocópias em humanos e outras espécies, desenvolvemos uma nova *pipeline* de identificação de retrocópias. Apesar do grupo já possuir uma *pipeline* para a busca de retrocópias fixadas[53], a mesma foi aplicada em alguns primatas e não se mostrou satisfatória quando aplicada a outras espécies, como roedores, devido a uma maior variabilidade entre genes parentais e suas retrocópias, algo encontrado com maior frequência fora da linhagem de primatas.

Tendo em vista essas principais limitações, desenvolvemos uma nova *pipeline* de busca com metodologia de alinhamento diferente e aplicamos novos filtros que garantem maior eficiência computacional e precisão no processo de identificação de retrocópias. Assim como na versão anterior da RCPedia, a busca por retrocópias se inicia com o alinhamento de RNAs mensageiros contra o genoma de referência, conforme ilustrado na Figura 1. Entretanto duas alterações principais foram realizadas: primeiramente, a sequência dos RNAs mensageiros de genes codificadores (XM_ ou NM_) é extraída da posição genômica anotada pelo RefSeq (em arquivo gff) pelo algoritmo gffread[59], para evitar discrepâncias entre a sequência do mRNA e a anotação genômica, uma inconsistência ocasional apontada pelo próprio RefSeq, evitando assim erros na determinação da posição da borda éxon-éxon. Segundo, fizemos uma transição do alinhador BLAT para o alinhador LAST[60], executado com o parâmetro: *lastal -D1000*. O LAST é um software de bioinformática de alto desempenho usado para alinhar rapidamente sequências de DNA ou proteínas em genomas de referência de grande tamanho. Ele emprega um processo de geração de sementes adaptativas que diferem das sementes de comprimento fixo, usadas pelo BLAT, ao ajustar dinamicamente seu comprimento com base nas características das sequências de entrada. As sementes adaptativas tornam o LAST um algoritmo de alta sensibilidade, especificidade, velocidade e eficiência, sendo o mais adequado em tarefas como mapeamento de genoma, predição de genes e genômica comparativa. A escolha do LAST também foi respaldada por sua destacada performance em nossos testes com dados simulados, bem como em experimentos de alinhamento de retrocópias em relação ao genoma de referência. Além dessas mudanças cruciais, incorporamos uma série de filtros suplementares que garantiram uma identificação mais precisa e eficiente de retrocópias. Estes filtros foram implementados por meio de uma variedade de *scripts* em linguagens como Python e Perl e shell *scripts*, juntamente com o uso de ferramentas de bioinformática, como o bedtools[61].

Sobre os filtros: a partir do resultado do LAST selecionamos apenas os alinhamentos com correspondência (*match)* superior a 120 pares de base e distância superior a 200.000 pares de base do gene de origem daquele RNA mensageiro, para remover possíveis duplicações cromossômicas. Em seguida, determinamos a posição da borda éxon-éxon dentro da sequência do RNA

mensageiro de genes codificadores. A partir da anotação genômica atualizada, determinamos as bordas éxon-éxon da sequência do RNA mensageiro. Verificamos, então, se houve o alinhamento da última, penúltima ou antepenúltima borda éxon-éxon, uma vez que a transcriptase reversa processa o RNA mensageiro a partir da cauda poli-A. Também verificamos se esses alinhamentos estavam ocorrendo em regiões altamente repetitivas, baseado na anotação de elementos repetitivos do *RepeatMasker[62]*, com adição do *simpleRepeats* e do *windowMasker* para não mamíferos, excluindo aqueles alinhamentos compostos majoritariamente por elementos repetitivos (superior ou igual a 40%). Outra questão resolvida pela *pipeline* se refere à sobreposição de vários alinhamentos de diferentes RNA mensageiros do mesmo gene em uma mesma região. Nesse caso, é selecionado o alinhamento com a maior correspondência e, em caso de empate, com a maior pontuação *(match/(match+mismatch))*.

Adicionalmente, o algoritmo lida com o fato de que o LAST cria diversos alinhamentos curtos e parciais do RNA mensageiro. Essa característica é o que permite que este algoritmo nos dê alinhamentos de alta qualidade, sem a necessidade de "mascarar" o genoma de referência, mas queremos ter a informação completa das regiões retrocópiadas. A etapa seguinte da *pipeline* reúne alinhamentos contínuos e oriundos de um mesmo RNA mensageiro que estejam até 6.000 pares de base de distância no genoma, permitindo a presença de elementos repetitivos entre os alinhamentos mais curtos.

Uma dificuldade que encontramos se refere às espécies de não mamíferos onde a constituição genômica é consideravelmente diferente, principalmente nos tipos de elementos repetitivos, uma qualidade mais baixa na anotação de genes codificadores e vastas regiões altamente repetitivas, como as regiões pericentroméricas e subteloméricas de *Xenopus laevis[63]* e a maior parte da sequência do cromossomo 4 de *Danio rerio[64]*. Todos esses fatores, em conjunção, resultaram em diversos falsos positivos identificados pela *pipeline*. Para contornar essa dificuldade, optamos então por implementar mais um filtro que em muito pouco afetou os resultados para os mamíferos, mas que aumentou substancialmente a qualidade dos nossos resultados em não mamíferos. Nominalmente: permitimos apenas uma inserção de retrocópia de um mesmo gene parental a cada 500.000 pb. O filtro de distância consegue eliminar muitos dos casos de falso positivo nessas

espécies sem criar *blacklists* que precisam ser conhecidas de antemão, agnóstica de espécie e dispensa a necessidade de eliminarmos esses casos de forma manual.

Outra dificuldade encontrada foi a identificação de retrocópias de genes de grandes famílias gênicas como *ZincFingers*, Tubulinas e Actinas que são, no geral, genes codificadores mais curtos que a média e com muitas cópias próximas a nível genômico (muitas geradas por duplicações *in tandem*). Portanto, retiramos candidatos com sobreposição de 3 ou mais éxons de genes codificadores anotados, ou que tivessem mais de 5 retrocópias com sobreposição a genes anotados da mesma família.

Por fim, verificamos se havia sobreposição de candidatos de genes parentais distintos, selecionando aqueles com maior pontuação, e que em caso de empate, um dos candidatos é escolhido aleatoriamente. Também realizando uma análise manual desses poucos casos, mantendo ocorrências de inserção de retrocópias dentro de retrocópias mais antigas.

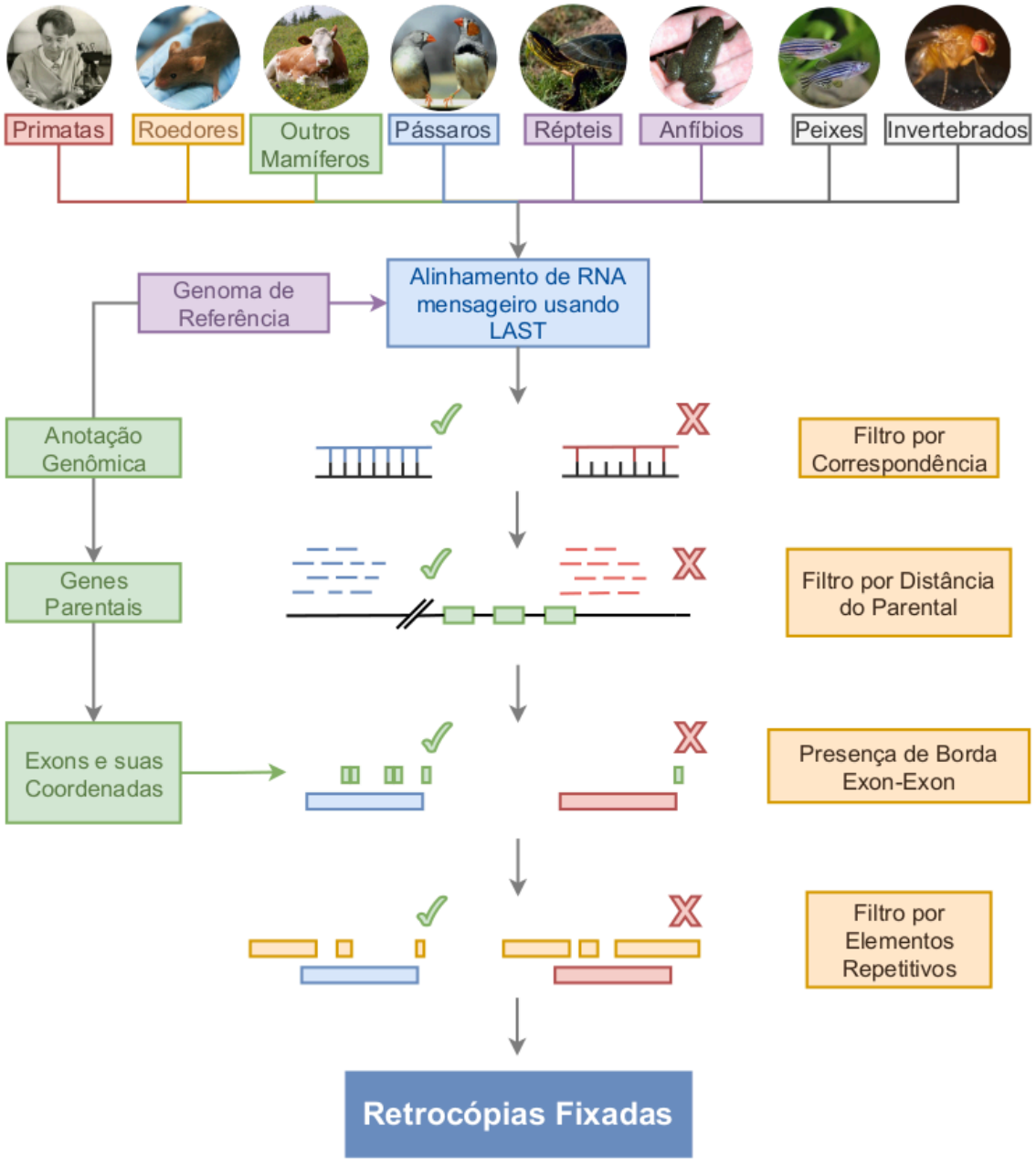



**Figura 1: Fluxograma da Metodologia para Identificar Retrocópias Fixadas**. Em resumo, RNAs mensageiros são alinhados com o alinhador LAST contra o genoma de referência. Os resultados são filtrados pela correspondência e distância do gene de origem do RNA mensageiro (parental). Os candidatos a retrocópias são filtrados com base na presença de bordas exon-exon e na presença de elementos repetitivos.

## 1.3.2. Dados Primários

O FASTA genômico, a anotação de genes codificadores e a anotação das sequências repetitivas foram obtidos do repositório *ftp* do RefSeq (https://ftp.ncbi.nlm.nih.gov/refseq/) ou do Genome Browser (https://hgdownload.soe.ucsc.edu/downloads.html) para as respectivas montagens

genômicas listadas na Tabela 1.

**Tabela 1. Espécies selecionadas para busca de retrocópias**

| Espécie | Montagem |
|---|---|
| Homo sapiens | GCF_000001405.39 |
| Pan troglodytes | GCF_002880755.1 |
| Pan paniscus | GCF_013052645.1 |
| Gorilla gorilla | GCF_008122165.1 |
| Pongo abelii | GCF_002880775.1 |
| Nomascus leucogenys | GCF_006542625.1 |
| Chlorocebus sabaeus | GCF_000409795.2 |
| Macaca fascicularis | GCF_000364345.1 |
| Macaca mulatta | GCF_003339765.1 |
| Papio anubis | GCF_008728515.1 |
| Rhinopithecus roxellana | GCF_007565055.1 |
| Callithrix jacchus | GCF_009663435.1 |
| Microcebus murinus | GCF_000165445.2 |
| Mus musculus | GCF_000001635.27 |
| Rattus norvegicus | GCF_000001895.5 |
| Cricetulus griseus | GCF_003668045.3 |
| Oryctolagus cuniculus | GCF_000003625.3 |
| Sus scrofa | GCF_000003025.6 |
| Bos taurus | GCF_002263795.1 |
| Ovis aries | GCF_002742125.1 |
| Tursiops truncatus | GCF_011762595.1 |
| Equus caballus | GCF_002863925.1 |
| Canis familiaris | GCF_014441545.1 |
| Ailuropoda melanoleuca | GCF_002007445.1 |
| Felis catus | GCF_000181335.3 |
| Choloepus didactylus | GCF_015220235.1 |
| Sarcophilus harrisii | GCF_902635505.1 |
| Monodelphis domestica | GCF_000002295.2 |
| Ornithorhynchus anatinus | GCF_004115215.1 |
| Gallus gallus | GCF_000002315.6 |

| | |
|---|---|
| Meleagris gallopavo | GCF_000146605.3 |
| Taeniopygia guttata | GCF_008822105.2 |
| Melopsittacus undulatus | GCF_012275295.1 |
| Chrysemys picta | GCF_000241765.4 |
| Anolis Carolinensis | GCF_000090745.1 |
| Xenopus tropicalis | GCF_000004195.4 |
| Danio rerio | GCF_000002035.6 |
| Drosophila melanogaster | GCF_000001215.4 |
| Phyllostomus discolor | GCA_014049915.1 |
| Rhinolophus ferrumequinum | GCA_014108255.1 |
| Molossus molossus | GCF_014108415.1 |
| Rousettus aegyptiacus | GCF_014176215.1 |
| Pipistrellus kuhlii | GCF_014108245.1 |
| Myotis myotis | GCF_014108235.1 |

### 1.3.3. Comparação com outros bancos de dados

Para averiguar a qualidade dos dados, realizamos a comparação das retrocópias de humanos e de camundongos identificadas pelo nosso algoritmo com as anotações de retrocópias mais atuais disponíveis. Para humanos, obtivemos as coordenadas dos transcritos anotados como sendo pseudogenes processados no Gencode V40 (https://www.gencodegenes.org/human/release_40.html), das retrocópias da RCPedia 1.0 e do RetrogeneDB 2017, realizando o *liftOver* da versão hg19 do genoma para a versão hg38 e das retrocópias anotadas pelo UCSC Track RetrogenesV9 (https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=1727274138_DqpgJvyo6muAaEh7eNSPs1TGy8NT&db=hg38&c=chr2&g=ucscRetroAli9).

### 1.3.4. Determinação de Ortologia entre as Espécies

Para determinar as retrocópias potencialmente ortólogas, primeiro recuperamos a sequência genômica em torno de cada retrocópia (3.000 pb a montante e a jusante) e realizamos o alinhamento de cada região contra todas as

regiões de retrocópias (3.000 pb a montante e a jusante) das outras espécies de forma par-a-par usando o alinhador Lastz[65]. Os pares foram filtrados com base em uma cobertura de alinhamento acima de 60% e identidade acima de 70% entre as regiões genéticas quando a comparação foi entre primatas e cobertura acima de 50% e identidade acima de 60% quando uma ou as duas espécies não eram primatas, uma vez que esperamos um nível de conservação reduzido. Também verificamos se pelo menos 60% da retrocópia foi alinhada na região “alvo” da outra espécie quando a comparação foi entre primatas ou pelo menos 50% quando uma ou as duas espécies não eram primatas. No caso de mais de uma possibilidade de ortologia, selecionamos o alinhamento de maior cobertura e identidade para cada retrocópia. Para répteis, anfíbios, peixes e invertebrados (cinco espécies) não foi possível encontrar ortólogos por conta da baixa identidade e cobertura dos alinhamentos.

### 1.3.5. Desenvolvimento do Banco de Dados para a RCPedia 2.0

Para organizar e disponibilizar os resultados obtidos de todas as espécies e servir como base para a interface *web* da RCPedia que aqui estamos chamando de RCPedia 2.0 (https://www.rcpediadb.org/) (uma atualização da versão atual: https://www.bioinfo.mochsl.org.br/rcpedia/), desenvolvemos um banco de dados em MySQL, em um sistema de gerenciamento de banco de dados baseado na linguagem SQL. Por ser uma atualização completa, com a atualização de todos os dados anteriores além da inclusão de diversos novos organismos, avaliamos e resolvemos desenvolver *ab initio* uma nova estrutura para o banco de dados.

O desenho do banco de dados foi feito com a ferramenta *MySQL Workbench* (https://www.mysql.com/products/workbench), Figura 2. Em resumo, os dados estão armazenados em 11 tabelas inter-relacionadas, cada uma contendo diversos campos e informações: a tabela principal *rtcs*, contém as informações básicas sobre as retrocópias - nome, espécie, transcrito do gene parental que originou a retrocópia, cromossomo, posição inicial, posição final, fita de inserção, *link* para a visualização no *UCSC Genome Browser*, *ensembl id* e *name* correspondente quando disponível. A tabela *rtc_seq* contém a sequência nucleotídica da retrocópia,

identidade com o parental e, futuramente, o cálculo de Ka/Ks. A tabela *conservation* contém informação sobre a ortologia das retrocópias de forma par-a-par. No restante das tabelas, armazenamos dados relativo às espécies (*species, chromosomes*), aos gene parentais (t*rascripts_parental*, *rnaseq-exp_parental*), aos genes hospedeiros (*rtc_region*), a todos os genes codificadores (*genes*) e aos experimentos de RNA-Seq (*rnaseq-exp_sample*).

Na tabela *rnaseq-exp_rtc* organizamos os dados de expressão. Para humanos, incorporamos os dados processados do GTEx (https://gtexportal.org/home/) e ARCHS4 (https://maayanlab.cloud/archs4/). Para camundongos, também usamos os dados processados do ARCHS4. Para o restante das espécies, selecionamos experimentos disponibilizados pelo SRA para Cérebro, Coração, Rim, Fígado, Ovário e Testículo, escolhendo até 15 amostras por tecido de experimentos de sequenciamento *paired-end*. Para 13 espécies, os experimentos foram selecionados de acordo com o artigo de Fukushima e colaboradores[66]. A quantificação da expressão das retrocópias foi feita através da ferramenta Kallisto[67] com índices construídos a partir das sequências anotadas dos genes codificadores e das sequências das retrocópias anotadas pela RCPedia 2.0.

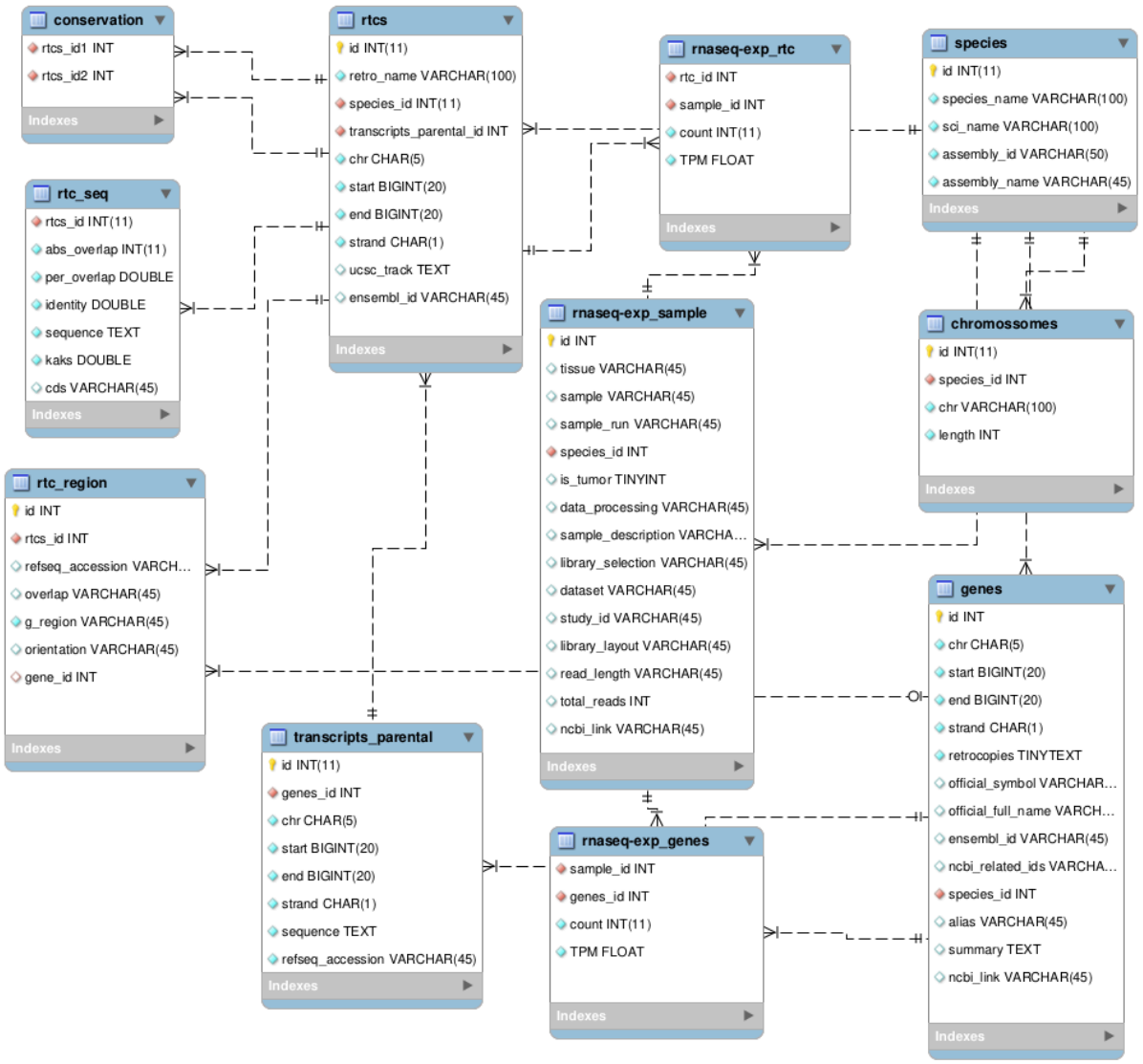



**Figura 2: Desenho do Banco de Dados da RCPedia 2.0.** O banco é composto de 11 tabelas com informações a respeito de todas as retrocópias das 44 espécies: posição, sequência, transcrito de origem, coordenadas de inserção em outros genes, conservação com outras espécies e, caso disponível, expressão em tecidos normais.

## 1.4 RESULTADOS & DISCUSSÃO

### 1.4.1 Panorama geral das retrocópias identificadas

A *pipeline* para identificação de retrocópias foi implementada para 44 espécies, sendo 13 primatas, 4 roedores, 18 espécies de outros mamíferos do porco ao ornitorrinco, 4 aves, 2 répteis, 1 anfíbio, 1 peixe e 1 invertebrado. Essas espécies foram selecionadas uma vez que apresentam uma montagem do genoma à nível de

cromossomo, disponível pelo *UCSC genome browser* ou pelo RefSeq, transcriptoma da mesma montagem anotado pelo RefSeq e anotação dos elementos repetitivos pelo *RepeatMasker*. Um resumo do número de retrocópias identificadas e de genes parentais que deram origem à essas retrocópias por espécie pode ser visto na Figura 3.

O número de retrocópias dos primatas está em linha com o que sabemos a partir dos dados da primeira versão da RCPedia: humanos com 8.080 retrocópias; saguis sendo o primata com mais retrocópias, registrando 11.244 em nossa nova *pipeline*; outros primatas com um número médio em torno de 8.000 retrocópias, conforme ilustrado na Figura 3. No restante dos mamíferos destaca-se o número exacerbado de retrocópias em preguiças (15.858) quando comparado com qualquer outra espécie (média de aproximadamente 6.000) e até mesmo com as espécies mais próximas. Vale destacar que, em comparação com os outros mamíferos, as preguiças não apresentam muito mais genes (23.536), éxons (243.811) ou elementos LINEs (1.938.270). Hipotetizamos que tal diferença se deve ao tipo de subfamília de LINE - possivelmente mais permissiva para retroduplicar mRNAs in trans - que é (ou esteve) ativa nessa espécie e o tempo de atividade dessa subfamília, porém essa hipótese precisa ser testada.

Em outras espécies, há um número menor de retrocópias. Em monotremados, aves, répteis, anfíbios e peixes, o número de LINEs fica entre 100.000 e 200.000 cópias ou abaixo, enquanto que nos mamíferos o número de LINEs é de 1.000.000 a 2.500.000 cópias. Nas moscas *Drosophila*, a espécie com o menor número de retrocópias, existem apenas 8.100 cópias de LINE. Além disso, há uma diferença em termos de composição dos elementos LINE entre as espécies. Por exemplo, enquanto os ornitorrincos (platypus; monotremados) possuem uma predominância de elementos LINE2[68], nos mamíferos a predominância é de LINE1. Aves apresentam poucas cópias de elementos transponíveis no geral, com um número baixo de elementos LINE1 detectáveis[68]. A discrepância no número de retrocópias entre mamíferos, aves, monotremados e anfíbios pode ser uma consequência de uma atividade menos intensa de LINE1 nessas espécies com um menor número de retrocópias. Isso pode também ser atribuído à atividade de LINE2 nessas espécies, uma vez que a maquinaria de retroduplicação desses elementos não é capaz de retroduplicar outros mRNAs *in trans[68]*.

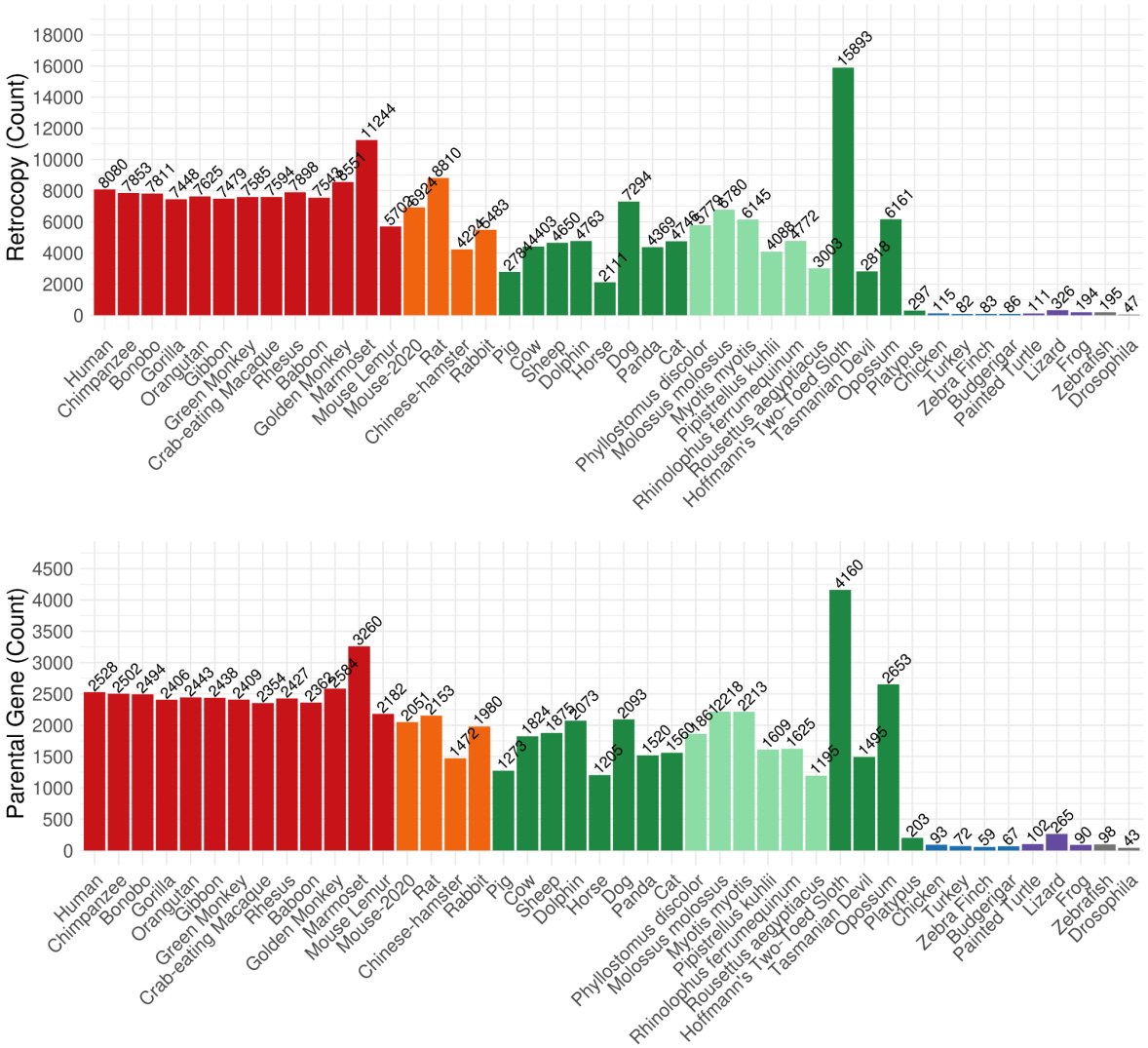



**Figura 3: Número de Retrocópias em 44 Espécies.** A) Número de Retrocópias por Espécies. B) Número de Genes Parentais por Espécies. Em vermelho: grupo dos primatas, em laranja: grupo dos roedores, em verde escuro: outros mamíferos, em verde claro: morcegos, em azul: pássaros, em roxo: répteis e anfíbios; em cinza: peixes e invertebrados.

## 1.4.2 Caracterização das retrocópias identificadas

O tamanho de cada região retrocópiada é calculado com base nas coordenadas das retrocópias, retirando de suas sequências as porções compostas por elementos repetitivos, de forma a representar a sequência do mRNA mensageiro inserido originalmente. Na Figura 4, apresentamos a distribuição do tamanho, em escala logarítmica para melhor visualização, das retrocópias por espécie. Para a maioria das espécies, a mediana gira em torno de 500 a 750 pares de base. Isso é

esperado, pois sabemos que a processividade da transcriptase reversa de L1 é de algumas centenas de nucleotídeos (cerca de 600)[39]. Principalmente para primatas, observamos que o tamanho varia muito pouco entre espécies, o que pode refletir na similaridade do tipo de LINE que esteve ativo na história evolutiva dessas espécies. Curiosamente, preguiça (*sloth*) também apresenta as retrocópias mais longas (mediana > 1000 pb), sugerindo que, de fato, a transcriptase reversa dessa espécie é mais eficiente em termos de capacidade de retroduplicação do a das outras espécies analisadas neste trabalho.

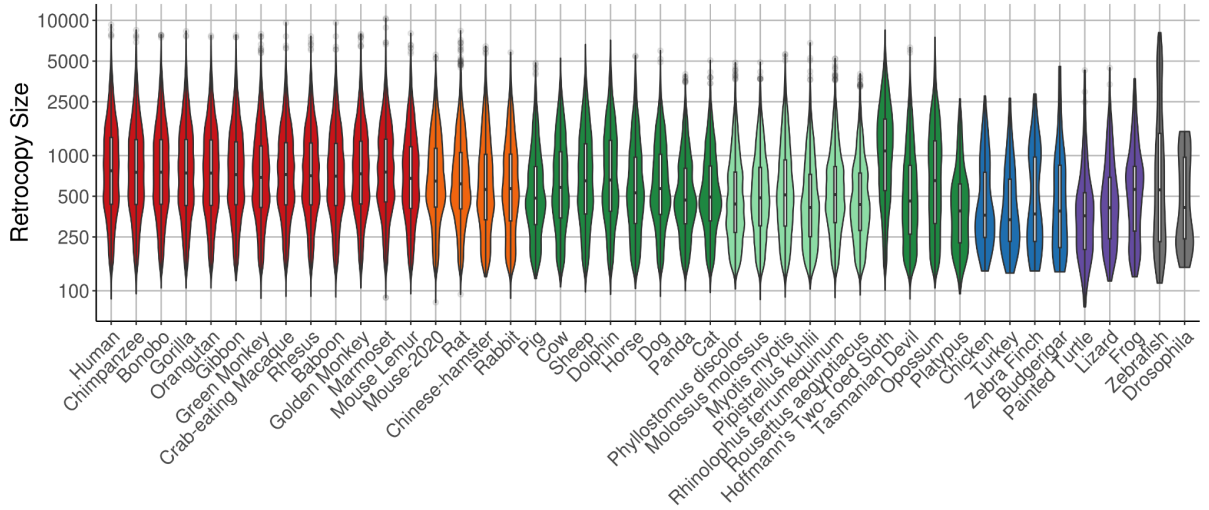



**Figura 4: Distribuição do Tamanho das Retrocópias por Espécie.** Eixo Y em log10. Tamanho representa a sequência de mRNA originalmente retrocopiada. Cores seguem o padrão da Figura 3.

Na Figura 5, podemos observar o panorama do número de retrocópias por gene parental em cada espécie. Optamos por apresentar os dados em escala logarítmica  novamente para melhorar a visualização da distribuição. Fica claro que para a maioria das espécies pelo menos metade dos genes parentais apresentam apenas 1 retrocópia, e que 75% apresenta até no máximo 10 retrocópias, em acordo com as observações dos primeiros artigos de buscas sistemáticas por pseudogenes processados, onde a quantidade de retrocópias por parental segue uma lei de potências.

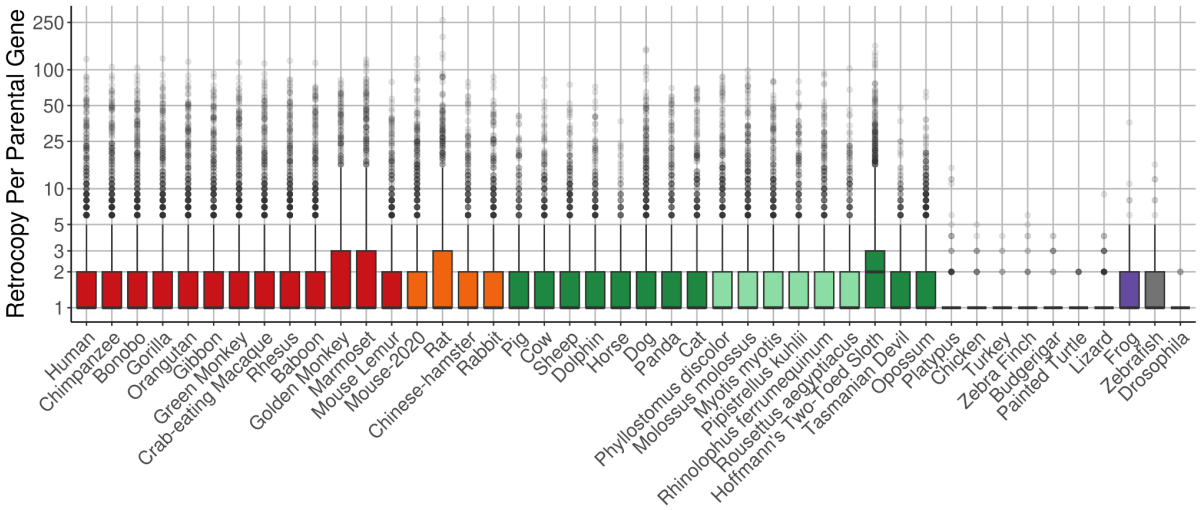



**Figura 5: Número de Retrocópias por Gene Parental por Espécie.** Eixo Y em log10. Cores seguem o padrão da Figura 3.

Com relação à ortologia (Figura 6), descobrimos que, em média, 30% das retrocópias identificadas são conservadas na maioria dos vertebrados (mamíferos e aves). Especificamente, os humanos têm 96,5% de suas retrocópias conservadas com outras espécies, principalmente outros primatas. Em geral, 82% das retrocópias de primatas são conservadas – principalmente com outros primatas. Surpreendentemente, a maioria das retrocópias de roedores são principalmente específicas da espécie, com apenas 10% das retrocópias partilhadas com outras espécies. Para outros mamíferos, descobrimos que em média 30% das retrocópias possuem ortólogos, mas são observados casos proeminentes, como a preguiça que tem mais de 96% de suas retrocópias sendo espécie-específicas.

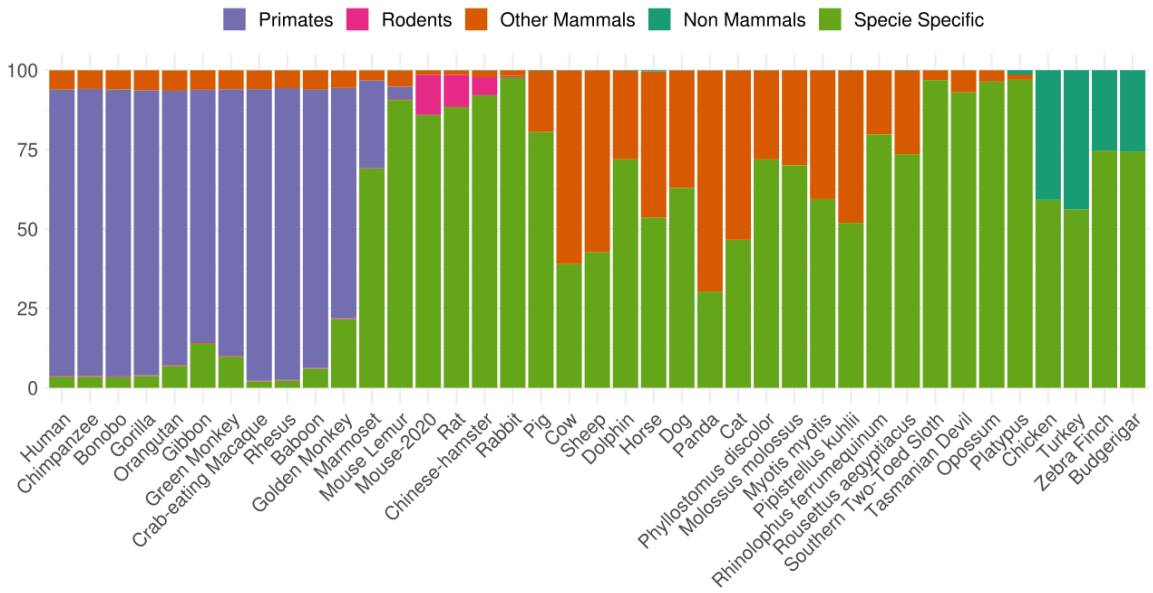



**Figura 6: Ortologia das Retrocópias entre as Diversas Espécies.** Em roxo: porcentagem das retrocópias compartilhadas com primatas. Em rosa: porcentagem das retrocópias compartilhadas com roedores. Em laranja: porcentagem de retrocópias compartilhadas com mamíferos não primatas e não roedores. Em verde escuro: porcentagem das retrocópias compartilhadas com não mamíferos. Em verde claro: porcentagem das retrocópias espécie específica.

## 1.4.3 Comparação com outros bancos de dados

Para avaliar a qualidade dos nossos dados, realizamos uma comparação baseada em coordenadas das retrocópias que identificamos no genoma humano com as retrocópias dos bancos de dados mais atuais: RCPedia 1.0 e RetrogeneDB; e com anotações de pseudogenes processados do Gencode V40 e de retrocópias do UCSC Track RetroGenes V9[69] (Figura 7). Encontramos que 98,6% (7.966/8.080) das retrocópias da RCPedia 2.0 também são identificadas em ao menos uma outra anotação, 94,5% (7.640/8.080) são identificadas em duas ou mais outras anotações, 82,24% (6.645/8.080) são identificadas em três ou mais outras anotações, e 44% (3.568/8.080) retrocópias são encontradas por todas as metodologias.

Em relação à versão anterior da RCPedia, encontramos uma sobreposição de 6901 retrocópias, representando 86% (6.901/8.080) do nosso conjunto. Diferenças entre as metodologias explicam a sobreposição não completa entre as anotações. Dado que na versão anterior da RCPedia foi aplicado o alinhador BLAT na montagem do hg19 e apenas com o filtro de uma única borda éxon-éxon, mudanças no alinhador e a na montagem inerentemente criam diferenças entre as anotações.

Além do mais, o filtro mais permissível da RCPedia 1.0 permite que alguns casos de duplicações cromossomais sejam anotados como retrocópias. Por outro lado, a RCPedia 2.0 é mais estrita, exigindo que uma das bordas mais a 3' do RNA mensageiro esteja presente no evento, o que exclui, por exemplo, retrocópias de transcritos com eventos de *splicing* alternativo (éxon *skipping*). Entretanto, cabe a avaliação de que um filtro de borda mais estrito, ao mesmo tempo que reduz a incidência de duplicações cromossômicas, pode retirar outros casos verdadeiros. Porém, podemos verificar que das 1.179 retrocópias a mais que identificamos, 1.065 (90,33%) foram anotadas por ao menos uma das outras três metodologias e 782 (66,33%) por duas ou mais, certificando uma alta qualidade do nosso dado.

Comparando as retrocópias da RCPedia 2.0 com as retrocópias do RetrogeneDB, apenas 4.254 retrocópias são compartilhadas, porém representando 92,26% (4.254/4.611) das retrocópias do RetrogeneDB. Como adiantamos anteriormente, a *pipeline* de identificação de retrocópias pelo RetrogeneDB é uma metodologia bastante estrita, exigindo que as retrocópias apresentem ao menos duas bordas éxon-éxon do gene parental. Esse filtro implica na seleção apenas dos eventos mais longos e, como sabemos, a processividade da transcriptase reversa do LINE-1 é de algumas centenas de pares de base, espera-se que a maioria dos eventos de retroduplicação sejam truncados e contenham poucos éxons. Portanto, essa condição certamente acaba descartando muitos casos reais de eventos de retroduplicação em que apenas dois éxons foram retroduplicados. É claro que o conjunto de retrocópias do RetrogeneDB é altamente curado, mas quase que na mesma quantidade, a RCPedia 2.0 encontra 3.712 retrocópias não identificadas pelo RetrogeneDB mas identificadas por outras anotações, o que chega a representar um aumento de 87,26% (3.712/4.254) de retrocópias anotadas pela RCPedia 2.0 em comparação ao RetrogeneDB.

Outro grupo que investigamos são os casos que não são identificados pelos bancos de dados dedicados às retrocópias, mas classificados como pseudogenes processados pelo Gencode v40 e/ou Retrogenes V9. A metodologia do Retrogenes V9 se baseia no alinhamento de sequências nucleotídicas do GenBank[70] contra o genoma utilizando o alinhador Lastz. O alinhamento múltiplo das sequências é o suficiente para classificar os eventos como retrocópias, sendo filtrados apenas eventos nos quais os alinhamentos são compostos por mais de 50% de sequências

repetitivas e genes mitocondriais, genes da família de *zinc fingers* e de imunoglobulinas são retirados. Acreditamos que, por esses motivos, o Retrogenes V9 anota 15.491 retrocópias, onde mais de um terço não são identificados por outras metodologias (37,37% - 5.789/15.491), consistindo principalmente de anotações em cromossomos alternativos e retrocópias de genes não codificadores, que são de baixa confiança dada a ausência de bordas éxon-éxon (ambos ausentes em outras anotações ou bancos de dados). Do restante, 1.495 (9,65%) são identificadas em ao menos uma outra anotação, sendo a maioria identificadas apenas pelo Gencode v40 e 8.207 (52,98%) são identificadas por duas ou mais outras anotações, o que representa o conjunto de maior confiança do Retrogenes V9.

Já o Gencode v40 utiliza um conjunto de metodologias computacionais, principalmente o Pseudopipe[71] que se baseia no alinhamento de sequências de aminoácidos contra o genoma de referência utilizando o alinhador BLAST. Alinhamentos que se sobrepõem a genes funcionais são removidos e os alinhamentos restantes são então reunidos em anotações de pseudogenes (processados, duplicados ou ambíguos). Pseudogenes órfãos (sem gene parental evidente) também são identificados através do Pseudopipe, utilizando o conjunto de genes codificadores de proteínas de um organismo diferente como entrada. As anotações do Pseudopipe são conferidas com resultados do Retrogene V9 e da RCPedia 1.0, de acordo com o artigo do Gencode 2019[72]. Essas anotações computacionais são então combinadas com anotações manuais para produzir o conjunto completo de pseudogenes. As anotações de pseudogenes recebem um nível de confiança com base na interseção com anotações manuais. Anotações detectadas tanto pelos *pipelines* computacionais quanto por curadoria manual recebem o nível 1, aquelas detectadas apenas por curadoria manual recebem o nível 2, e as anotações de consenso detectadas pelas *pipelines* computacionais recebem o nível 3 e são disponibilizadas em um arquivo de anotação separado.

No total são anotados 10.154 pseudogenes processados pelo Gencode v40, dos quais 1.277 (12,58%) são exclusivos do Gencode, e 1.061 são compartilhados apenas com o Retrogenes V9 (10,45%). Analisando alguns casos dessas categorias, observamos que cópias de genes mitocondriais (monoexônicos) e duplicações cromossômicas estão sendo erroneamente classificadas pelo Gencode como pseudogenes processados. Em comparação à RCPedia 2.0, 7.022 são

compartilhadas, representando 86,90% das retrocópias identificadas pela nossa *pipeline*.

Podemos concluir que a anotação de retrocópias é uma tarefa que exige muito mais do que o simples alinhamento múltiplo de transcritos de genes. Uma abordagem pouco rígida como a do Retrogenes v9 acaba resultando em uma alta taxa de falsos positivos, enquanto que uma abordagem muito estrita como o RetrogeneDB acaba por não identificar milhares de retrocópias verdadeiras. Diferenças nos alinhadores, genomas e anotações também influenciam no resultado, algo que podemos observar na comparação entre a RCPedia 1.0 e a RCPedia 2.0. Por outro lado, estratégias como as do Gencode de agregar anotações manuais e anotações computacionais dificilmente poderiam ser implementadas em larga escala para outros organismos, além da dificuldade de ponderar a respeito da qualidade das anotações computacionais que exigem um profundo conhecimento do tópico. Nosso principal critério para determinar quais sequências são retrocópias baseia-se justamente na perda de ao menos um dos últimos três íntrons do gene parental, um forte indicativo do evento de retrotransposição iniciado a partir da cauda poli-A. É claro que a adoção do alinhador LAST - que permite o uso do genoma não mascarado -, anotações de transcritos cada vez mais curados e montagens genômicas cada vez mais completas trazem melhorias inerentes à *pipeline* de identificação. A alta taxa de sobreposição da RCPedia 2.0 com os outros bancos de dados (mais de 98%) e o fato de termos identificado cerca de 8000 retrocópias no genoma humano, um número consenso para nossa espécie, do nosso ponto de vista, são indicativos sólidos de que a RCPedia 2.0 fornece uma anotação de alta confiança de retrocópias.

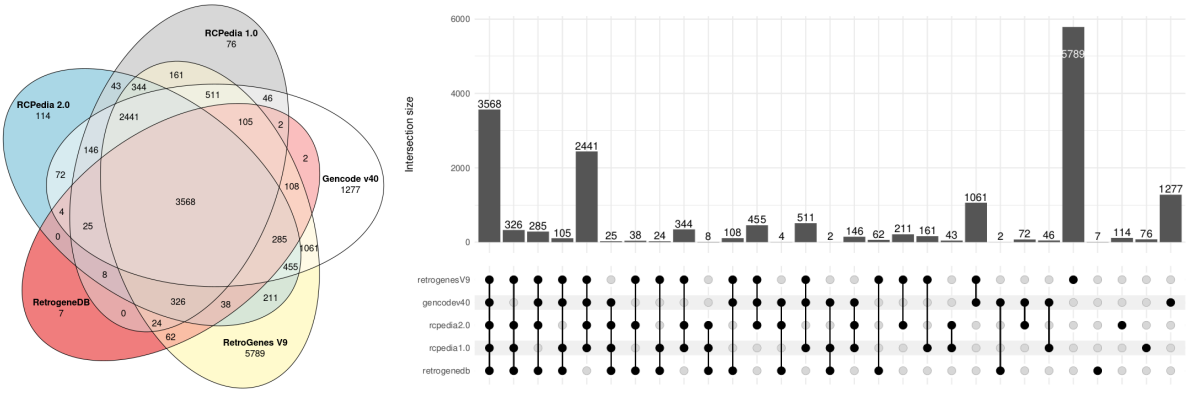



**Figura 7: Análise Comparativa entre a RCPedia 2.0 e 4 Outras Anotações de Retrocópias.** RCPedia 1.0, RetrogeneDB, RetroGenes V9 e Gencode V40. À esquerda, um diagrama de venn em que cada conjunto representa uma anotação. À direita, uma representação em *upsetplot* dos conjuntos do diagrama de Venn.

## 1.5 CONSIDERAÇÕES FINAIS

Em conclusão, este estudo representa um marco importante na pesquisa sobre retrocópias, abordando de maneira abrangente a identificação, conservação e disponibilização de dados sobre esses elementos em uma ampla gama de espécies. As melhorias significativas feitas na reestruturação da *pipeline* de identificação de retrocópias, incluindo a transição para o alinhador LAST e a implementação de filtros suplementares, demonstram um compromisso com a precisão e a qualidade na anotação desses elementos em genomas.

A análise detalhada da conservação das retrocópias entre 44 espécies revela um panorama fascinante de diversidade e semelhanças na atividade retrotransposicional ao longo da evolução. Os primatas, com destaque para humanos e saguis, apresentam números substanciais de retrocópias, indicando um papel significativo desses elementos em suas linhagens evolutivas. No entanto, as preguiças surpreendem com um número exorbitante de retrocópias, sugerindo uma dinâmica evolutiva única associada a uma subfamília específica de elementos LINEs. Essa descoberta ressalta a importância de considerar as subfamílias de retrotransposons ao estudar a evolução das retrocópias.

A análise de ortologia das retrocópias também revela *insights* valiosos sobre a conservação desses elementos. Enquanto os primatas mostram um alto grau de conservação entre suas retrocópias, especialmente entre espécies próximas, os roedores apresentam retrocópias predominantemente específicas de cada espécie.

Vale destacar que o número limitado de espécies de roedores, quanto comparado com primatas, pode estar influenciando este resultado. Porém, é inegável que essa variação no número e na conservação reflete a diversidade de papéis e influências evolutivas das retrocópias em diferentes grupos de mamíferos.

A implementação do banco de dados em MySQL para armazenar e disponibilizar os resultados é um passo importante em direção à acessibilidade e à utilidade da pesquisa sobre retrocópias. Essa ferramenta não apenas facilita o acesso aos dados gerados neste estudo, mas também cria uma plataforma sólida para futuras investigações e análises por outros pesquisadores interessados no tema. A integração da RCPedia 2.0 com esse banco de dados promete ser uma valiosa fonte de informações para a comunidade científica interessada em retrocópias.

Em resumo, este estudo não apenas aprimora nossa compreensão sobre as retrocópias em uma variedade de espécies, mas também estabelece um novo padrão para a pesquisa nessa área. As descobertas aqui apresentadas fornecem uma base sólida para investigações futuras sobre o papel das retrocópias na evolução e na biologia molecular das espécies. À medida que continuamos a explorar os "segredos" das retrocópias, este estudo desempenha um papel fundamental em direcionar nosso conhecimento e expandir nossos horizontes científicos.

# Capítulo 02 - Potencial Funcional de Retrocópias no Genoma de Humanos: estudo dos padrões de expressão em tecidos normais

*Functional Potential of Retrocopies in the Human Genome: study of the expression pattern in normal tissues*

# Resumo

O mecanismo de retrotransposição de RNA mensageiro está ativo durante a evolução dos primatas e pode introduzir uma grande variedade genética entre linhagens e espécies. Por muito tempo, as retrocópias, sendo cópias de RNAs mensageiros e desprovidas de regiões promotoras para transcrição, foram rotuladas como "mortas na chegada". Ou seja, acreditava-se que não possuíam a expectativa de serem transcritas e, consequentemente, funcionais. No entanto, graças à evolução da genômica, transcriptômica, métodos de sequenciamento em larga escala e das análises computacionais, um número crescente de retrocópias funcionais em humanos e outras espécies vem sendo identificado. Este progresso levanta considerações acerca do papel das retrocópias na formação de reservatórios de novos *loci* genômicos potencialmente funcionais e, por conseguinte, no desempenho de atividades cruciais do ponto de vista evolutivo. Neste trabalho, partimos das retrocópias de humanos identificadas na RCPedia 2.0, integrando dados de sequenciamento de RNA em múltiplos tecidos, dados de epigenômica e experimentos de *Ribosomal Sequencing* (RiboSeq) para desvendar a regulação e a expressão de retrocópias, suas características e potencial funcional. Primeiramente, identificamos que aproximadamente 50% (4.038) das retrocópias fixadas no genoma humano são transcritas em tecidos saudáveis. De forma específica, identificamos que aproximadamente 25% das retrocópias são transcritas em um único tecido (principalmente nos testículos) e aproximadamente 15% são "pan-transcritas" em todos os tecidos humanos investigados. Identificamos que a força motriz para essa transcrição é a localização genômica das retrocópias próximas a genes codificadores de proteína. Confirmamos, por dados de RiboSeq que um conjunto de 1.077 retrocópias (26,6%), com potencial codificador, possuem evidência de tradução. Portanto, nossos resultados revelam padrões importantes em termos de expressão de retrocópias, seus possíveis mecanismos de regulação e funcionalidades. De uma forma mais ampla, nossa investigação aponta e destaca um papel importante, mas ainda relativamente negligenciado, das retrocópias em criar novidades genéticas para o transcriptoma humano.

## 2.1 INTRODUÇÃO

Até recentemente, as retrocópias eram classificadas como pseudogenes processados devido à suposição de que a ausência de regiões regulatórias para a transcrição seria um empecilho para a expressão das mesmas[73]. Contudo, análises de transcriptoma já revelaram que centenas (ou milhares) de retrocópias em humanos e camundongos são capazes de gerar transcritos - alguns potencialmente funcionais[74]. A transcrição das retrocópias é possível a partir da obtenção de sequências regulatórias que, em geral, são provenientes de elementos regulatórios já existentes na vizinhança do ponto de inserção da retrocópia no genoma, como mostrado pelos trabalhos de Okamura e Nakai (2008)[75], Fablet et al. (2009)[76] e Navarro e Galante (2015)[77]. Adicionalmente, outras vias de transcrição envolvem sequências ricas em CpG distantes e até mesmo suas próprias regiões em *cis* (um exemplo proeminente é a retrocópia PABP3[78]).

Como discutido no capítulo anterior desta tese, é sabido que todos os vertebrados, especialmente os mamíferos, possuem retrocópias fixadas. Somente o genoma humano contém aproximadamente 8000 retrocópias. Estudos recentes demonstram a importância das retrocópias para a evolução molecular, com papéis importantes em vias celulares complexas[77,78]. Alguns exemplos de retrogenes (retrocópias comprovadamente funcionais) incluem GLUD2, que apresenta função específica em tecidos neurais[79]; e TRIM5-CYPA que é uma quimera com papel na defesa antiviral[80]. Além disso, alterações na expressão de algumas retrocópias (retrogenes) já foram associadas a diferentes tipos tumorais, tais como a sub expressão da retrocópia PTENP1[81] e a super expressão da retrocópia BRAFP1[82]. Um outro exemplo de retrogene relacionado ao câncer é o gene RHOB, um supressor da família Rho GTPase, que surgiu de um evento de retrotransposição nos primeiros estágios da evolução dos vertebrados[83]. Apesar desses estudos pontuais, apenas algumas das 8000 retrocópias fixadas em nosso genoma tiveram sua funcionalidade atribuída até o momento.

Além das retrocópias fixadas, sabemos também que existe um conjunto, ainda não determinado de forma precisa, de retrocópias polimórficas na população humana[13]. As retrocópias polimórficas, também conhecidas como retroCNVs (Copy Number Variants baseados em retroduplicação) ou GRIPs (do Inglês, *Gene*

*Retrocopies In Polymorphism*), são um subconjunto especial de retrocópias que diferem dos pseudogenes fixados pelo fato de não estarem presentes em todos os indivíduos de uma espécie. Nosso grupo foi um dos primeiros a identificar essas retrocópias na população humana[10,13] e o primeiro a publicar uma ferramenta capaz de identificá-las[84]. A presença dessas retrocópias polimórficas em uma população indica que elas podem estar sob diferentes pressões seletivas, contribuir para a variabilidade genética, servir como marcadores em estudos populacionais e estarem associada a traços fenotípicos ou até mesmo associadas à susceptibilidade a doenças [85].

Por fim, também sabemos que as retrocópias podem ser adquiridas somaticamente em contextos patológicos. Alguns estudos demonstram que a reativação de retrotransposons em células somáticas durante a tumorigênese parece induzir a ocorrência de novos eventos de retrotransposição, que podem - assim como a inserção de elementos L1 e *Alu* - perturbar a estrutura e expressão de outros genes. Um exemplo é a inserção da retrocópia do parental TMF1 no gene CYBB, a qual induz um códon de parada prematuro[86]. Em 2014, Cooke e colaboradores[87] usaram 660 amostras de câncer e encontraram 42 eventos *de novo* de retrocópias em 17 amostras tumorais, especialmente em câncer de pulmão de células não pequenas (27/5) e câncer colorretal (11/2). Como discutido pelos autores, esses eventos representam uma nova classe de mutação que pode ocorrer durante o desenvolvimento do câncer, com consequências funcionais potencialmente diversas dependendo do contexto genômico da inserção e do gene parental retrocopiado.

Portanto, é razoável imaginar que muitas retrocópias funcionais ainda não foram bem caracterizadas e que a literatura carece de trabalhos que tenham explorado de forma completa, profunda e em larga escala a expressão e potencial funcionalidade de retrocópias em um número amplo de tecidos humanos. Graças ao desenvolvimento metodológico de tecnologias de sequenciamento de segunda geração (genericamente chamadas de *Next Generation Sequencing*), aliadas a metodologias de biologia molecular e ao desenvolvimento de *pipelines* complexos de análise de bioinformática, atualmente estamos em um momento único para investigar as retrocópias expressas e potencialmente funcionais.

## 2.2 OBJETIVOS

### 2.2.1 Objetivo Geral

Investigar a expressão, o potencial funcional e as contribuições das retrocópias para a origem de novas regiões gênicas no genoma humano.

### 2.2.2 Objetivos Específicos

- Identificar e analisar o perfil de transcrição de retrocópias em 14.910 amostras de 32 tecidos provenientes de indivíduos saudáveis.
- Identificar diversas características intrínsecas das retrocópias transcritas presentes no genoma humano, tais como o tamanho da região transcrita, posição genômica (exônica, intrônica ou intragênica) e a presença de uma fase aberta de leitura;
- Investigar a associação entre a posição genômica das retrocópias e seus níveis de expressão, assim como a abrangência de expressão das retrocópias e suas idades evolutivas.
- Analisar a capacidade e as evidências, por RiboSeq, das retrocópias serem traduzidas.

## 2.3 MATERIAIS & MÉTODOS

### 2.3.1. Dados de Expressão e Quantificação de Expressão de Retrocópias

Para a análise de expressão, utilizamos a lista de retrocópias fixadas em humanos descrita no capítulo anterior desta tese (Capítulo 1 - Origens de Retrocópias no Genoma de Animais: uma abordagem em larga escala de identificação de retrocópias). Avaliamos a expressão das retrocópias nos dados públicos pré-processados do GTEx V8[88]. Em resumo, o consórcio GTEx quantifica a expressão de genes anotados pelo GENCODE 26 (https://www.gencodegenes.org/human/release_26.html) a partir de um trascrito

modelo para cada gene, determinado por um procedimento customizado que exclui exons anotados como "retained_intron" e "read_through" e mescla o intervalo de exons sobrepostos. Exons sobrepostos entre genes diferentes foram excluídos. O código que gera esse modelo colapsado está disponível em: https://github.com/broadinstitute/gtex-pipeline/tree/master/gene_model. A quantificação a nível de gene foi determinada pelo algoritmo RNA-SeQC v1.1.9[89] com os seguintes filtros: *reads* mapeados unicamente, com os dois pares alinhados, permitindo até 6 nucleotídeos de *gap*, alinhados inteiramente dentro de exons ("-strictMode" flag do RNA-SeQC). Um passo a passo detalhado pode ser encontrado na página do github do consórcio GTEx (https://github.com/broadinstitute/gtex-pipeline/tree/master/rnaseq).

Para cada retrocópia da RCPedia 2.0 mapeamos um *gene id* do GENCODE versão 26[72]. Essa informação foi determinada pelo cruzamento das posições das retrocópias com todos os genes anotados utilizando a ferramenta *bedtools intersect[61]* com a condição de que houvesse uma sobreposição de ao menos 63 nucleotídeos (metade da menor retrocópia humana) e 80% de cobertura. Das 8.080 retrocópias fixadas, mantivemos um total de 6.756 retrocópias associadas a uma anotação ("gene") do GENCODE.

### 2.3.2. Seleção de Amostras e Tecidos do GTEx

Também realizamos uma seleção das amostras do GTEx v8. Começando com amostras dos 52 tecidos saudáveis, excluímos as duas linhagens celulares disponíveis e retiramos os tecidos que eram representados por um número muito limitado de amostras (menos de 50 amostras), que é o caso de Bexiga, Cérvix, Tubas Uterinas e Medula Renal. Outros tecidos foram agrupados por serem anatomicamente o mesmo tecido ou por não haver diferença significativa do perfil de expressão geral[90], o caso do Esofago Mucosa e Esofago Muscularis e dos dois tipos de Pele (exposta ao sol e não exposta). Tecidos cerebrais foram agrupados caso fossem parte da mesma estrutura, como os tecidos que fazem parte do Córtex Cerebral (Córtex Cingulado, Córtex e Córtex Frontal) e do Gânglio Basal (Núcleo Caudado, Núcleo Accumbens e Putamen). A amígdala, o hipocampo, o hipotálamo, a espinha dorsal e substância nigra foram reunidos na categoria "Cérebro - Outros".

Adicionalmente, retiramos os tecidos de Sangue Completo, Adiposo Visceral, Glândula Salivar Menor e Glândula Pituitária, uma vez que não haviam experimentos de metilação disponíveis para esses tecidos no ENCODE[91].

No final, analisamos os níveis de expressão em 14.910 amostras de 32 tecidos. A retirada e o agrupamento de certos tecidos nos permitiu ter um melhor panorama dos dados de expressão, além de focar a nossa atenção em tecidos de extrema importância como o Córtex Cerebral.

### 2.3.3. Determinação da Região Gênica de Retrocópias

Sabemos que existem diversos mecanismos de ganho de expressão pelas retrocópias, entre eles o uso de promotores de genes próximos[92]. Para avaliar a influência do local de inserção das retrocópias na sua expressão, separamos as retrocópias de acordo com a região genômica em que estão inseridas. Para determinar tal classificação, utilizamos, quando disponíveis, a anotação do MANE (versão 0.95) correspondentes aos genes codificadores de proteína do GENCODE (versão 26)[72], ou, no caso de genes sem uma versão representativa do MANE, utilizamos a anotação do gene dada pelo GENCODE (v26). Escolhemos fazer essa curadoria para evitar casos em que os genes apresentam um único transcrito mais longo que não é representativo da transcrição usual do gene, evitando falsos positivos de retrocópias intragênicas.

### 2.3.4. Dados de Metilação

Outro fator de influência a se considerar na expressão das retrocópias são as marcas epigenéticas associada à ativação ou repressão da transcrição. O Projeto ENCODE[91] reúne tais dados de metilação para diversos tipos de tecidos. Optamos por examinar os dados de modificação da histona H3 em sua lisina 4 através da trimetilação (H3K4me3), que corresponde ao estado (ativo) de metilação dos promotores. Além disso, incorporamos os dados de trimetilação da histona H3 na lisina 36 (H3K36me3), que marca a metilação do corpo gênico e de uma transcrição ativa. Para os experimentos de H3K36me3, cruzamos as coordenadas das retrocópias com as coordenadas dos picos e selecionamos aqueles com FDR inferior

a 0,05. Para os experimentos de H3K4me3 estimamos os sítios de início de transcrição (TSS, do inglês *Transcription Start Site*) na vizinhança das retrocópias (500 nucleotídeos), e verificamos a presença de picos ao redor desses TSS estimados (1.000 nucleotídeos), selecionando aqueles com FDR inferior a 0,05.

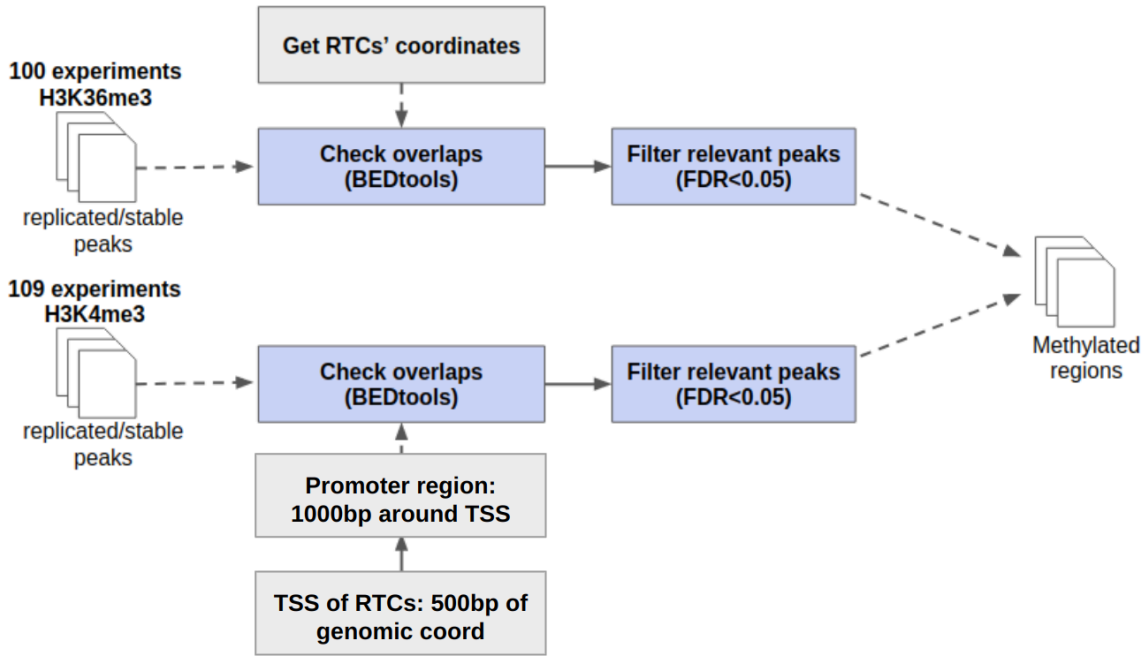



**Figura 1. Processamento dos experimentos de metilação.** Para experimentos de H3K36me3, as coordenadas das retrocópias foram sobrepostas aos picos, selecionando os que têm FDR abaixo de 0,05. Para experimentos de H3K4me3, estimamos os sítios de início de transcrição (TSS) próximos às retrocópias (500 nucleotídeos) e verificamos picos ao redor desses TSS (1.000 nucleotídeos), selecionando os com FDR inferior a 0,05.

## 2.3.5. Cálculo de Abrangência de Expressão

Os padrões de expressão dos genes podem apresentar muitas variações, especialmente quando realizamos uma análise inter-tecidual [93]. Em média, pode-se aparentar que um gene tenha uma expressão intermediária com relação aos outros, mas que na realidade apresenta um alto nível de expressão em um subconjunto dos tecidos e em um nível muito mais baixo ou nem mesmo expressos em outros tecidos. Uma forma de quantificar essa variação inter-tecidual é realizar o cálculo da "abrangência" da expressão gênica, e classificar os genes em: (i) genes de manutenção celular (*housekeeping*), os quais são expressos com níveis de expressão semelhantes em todos ou na maioria dos tecidos analisados; ou (ii) tecido-específicos, os quais são mais expressos em apenas um ou em um pequeno

subconjunto de tecidos. O cálculo de abrangência de expressão é realizado com a seguinte fórmula[94]:

$$\tau = \frac{\sum_{i=1}^{n} (1 - \hat{x_i})}{n - 1};\ \hat{x_i} = \frac{x_i}{\max\limits_{1 \le i \le n} (x_i)}.$$

Onde $x_i$ é a expressão do gene no tecido $i$, $\hat{x_i}$ é a expressão normalizada no tecido $i$, dado o valor máximo de expressão do gene nos tecidos analisados. $\tau$ é portanto a somatória da distância de $\hat{x_i}$ do número 1 (que representa o tecido de maior expressão), dividido pelo número de tecidos ($n$) menos 1. Logo, o valor de $\tau$ próximo de 1 representa um gene mais tecido específico e o valor de $\tau$ mais próximo de 0 representa um gene com expressão mais ubíqua (*housekeeping*).

## 2.3.6. Idade das Retrocópias

A idade das retrocópias foi determinada pela ortologia das retrocópias humanas como descrito na seção 1.3.2. Organizamos a ortologia em cinco grandes grupos: retrocópia específicas de humanos (*Human Specific;* grupo 1), retrocópias compartilhadas com os grandes primatas (*Great Apes;* grupo 2: humanos, chimpanzé, bonobo, gorila, orangotango), retrocópias compartilhadas com os primatas do velho mundo (*Old World Monkeys;* grupo 3: incluindo grandes primatas e gibão, macaco verde, macaco caranguejo, rhesus e babuíno), retrocópias compartilhadas com todos os outros primatas (*All Primates;* grupo 4) e retrocópias compartilhadas com não primatas (*Non Primates;* grupo 5). Em termos de idade, estima-se que as retrocópias do grupo 1 tenham sido originadas há menos de 6 milhões de anos; grupo 2, origem há menos de 16 milhões de anos; grupo 3, origem há menos de 30 milhões de anos; grupo 4, origem há menos de 90 milhões de anos; e grupo 5, origem há mais de 90 milhões de anos.

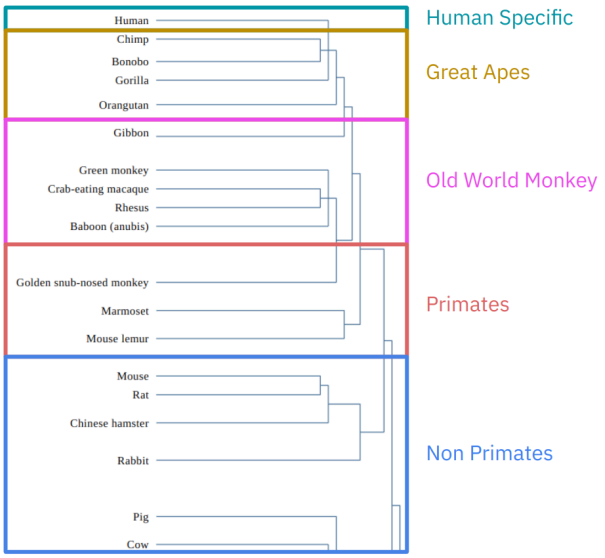



**Figura 2. Idade das Retrocópias de Humanos.** Foram estabelecidos cinco grandes grupos: retrocópia específicas de humanos (*Human Specific*; grupo 1), retrocópias compartilhadas com os grandes primatas (*Great Apes*; grupo 2), retrocópias compartilhadas com os primatas do velho mundo (*Old World Monkeys*; grupo 3), retrocópias compartilhadas com todos os outros primatas (*All Primates*; grupo 4) e retrocópias compartilhadas com não primatas (*Non Primates*; grupo 5).

## 2.3.7. Dados de *Ribosome Sequencing*

As análises de dados de *Ribosome Sequencing* (RiboSeq) foram realizadas com os dados já processados disponíveis no banco de dados RPFdb (versão 2.0)[95]. Para humanos, estão disponíveis três amostras saudáveis de cérebro, oito de rim e uma de músculo esquelético. Devido a baixa quantidade amostral e baixa quantidade de tecidos, essas amostras foram analisadas em conjunto. No banco de dados RPFdb, os dados foram processados da seguinte maneira: após a remoção dos adaptadores de sequência com a ferramenta Cutadapt[96], os dados de Ribo-Seq foram alinhados ao genoma de referência com o alinhador Bowtie2 (v2.3.4.1) [97] e as sequências contaminantes de rRNA e tRNA foram removidas. Em seguida, considerando-se que durante o processo de tradução o ribossomo envolve cerca de 25 a 34 nucleotídeos de um mRNA, apenas leituras com esse tamanho foram mantidas para análise. Finalmente, os dados resultantes foram processados com o algoritmo RibORF [98] para a predição de ORFs traduzidas ativamente.

Experimentos de RiboSeq visam identificar fragmentos de RNA associados a ribossomos. As células são tratadas com cicloheximida para interromper o alongamento dos ribossomos e a RNase I é usada para digerir regiões de RNA não

protegidas por complexos proteicos. Os complexos proteína-RNA são então isolados e os RNAs associados a esses complexos são purificados para sequenciamento. No entanto, como nenhum anticorpo é usado para selecionar especificamente complexos ribossomos-RNA, outros complexos proteína-RNA podem ser detectados. Para distinguir leituras provenientes de complexos ribossomos-RNA, o algoritmo RibORF baseia-se na periodicidade de 3 nucleotídeos (mesmo quadro de leitura) esperada para o movimento do ribossomo durante o processo de tradução, ao passo que leituras provenientes de outros complexos proteína-RNA não-ribossômicos tendem a mostrar uma distribuição mais uniforme ao longo do RNA (característica estacionária desses complexos) e são então eliminadas pelo algoritmo. Após a seleção de leituras potencialmente derivadas de complexos ribossomos-RNA, o algoritmo contabiliza os *counts* por CDS predita. Dentre todas as ORFs candidatas fornecidas pelo banco de dados RPFdb, selecionamos apenas ORFs de RTCs com códons de início ATG para as análises posteriores.

### 2.3.8. Resumo da Metodologia Aplicada

Na Figura 3 resumimos de forma gráfica as principais análises de expressão de retrocópias em dados do GTEx V8. A pesquisa incluiu a identificação de retrocópias associadas a anotações de genes, resultando em 6.756 retrocópias mantidas para análise. Tecidos foram selecionados e agrupados para uma análise mais focalizada, resultando em 14.910 amostras de 32 tecidos. A influência da localização das retrocópias e suas marcas epigenéticas associadas, como H3K4me3 e H3K36me3, na expressão foi avaliada. Além disso, a idade das retrocópias foi categorizada em cinco grupos e analisada. A pesquisa também explorou dados de RiboSeq para identificar ORFs traduzidas ativamente a partir de retrocópias.

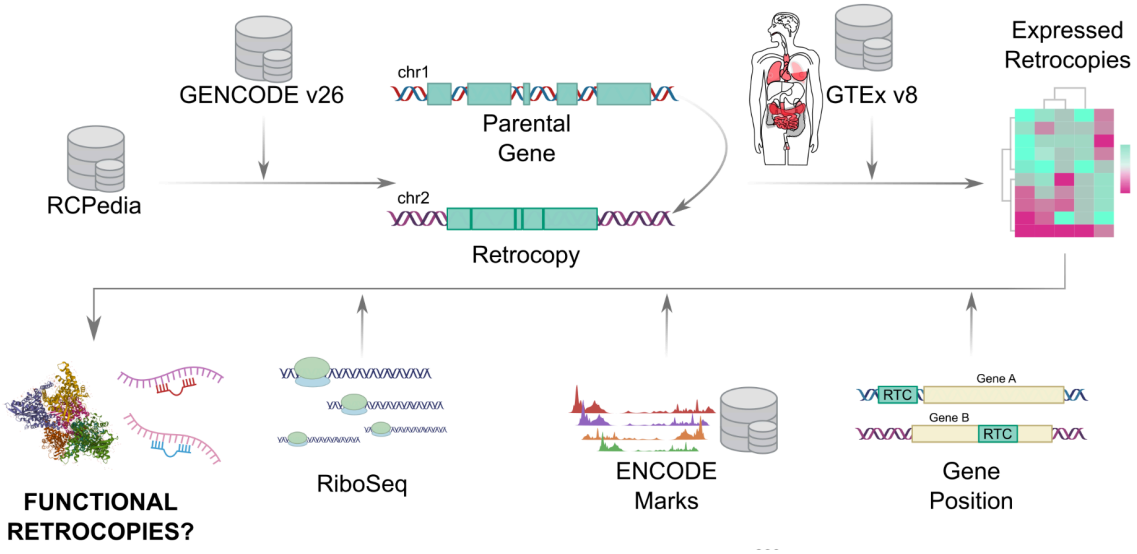



**Figura 3. Visão esquemática do fluxo de trabalho:** Utilizando a RCPedia como nosso banco de dados de retrocópias e o Gencode como nossa anotação de genes de referência, atribuímos a cada retrocópia conhecida um ID único do Gencode sempre que possível. Avaliamos a expressão de cada retrocópia em 32 tecidos humanos do GTEx v8, categorizando RTCs como expressas (expressão em pelo menos um tecido) e não expressas. Além disso, avaliamos a expressão das retrocópias de acordo com a posição em relação aos genes codificadores (intergênicas ou intragênicas), presença de marcas de transcrição (H3K4me e H3K26me) e informações traducionais disponíveis (RiboSeq).

## 2.4. RESULTADOS & DISCUSSÃO

### 2.4.1. Perfil de expressão das retrocópias

Na Figura 4, apresentamos um panorama geral de expressão das retrocópias nos tecidos do GTEx. Escolhemos como corte de expressão o valor de 0,1 TPM, considerando que pseudogenes processados assim como genes não-codificadores são geralmente expressos em uma ordem de grandeza abaixo de genes codificadores e similar aos genes não codificadores. Em uma análise global de tecidos normais, observamos, em média, que o terceiro quartil de expressão se situa em torno desse valor. Obtivemos então 4.038 retrocópias que apresentam expressão (>=0,1 TPM, na mediana) em ao menos um dos 32 tecidos. Por outro lado, 2.718 retrocópias não atingiram esse corte de expressão (>=0,1 TPM, na mediana) e foram consideradas sem evidência de expressão (Figura 4A).

Em seguida, também calculamos a correlação de expressão das retrocópias com seus genes parentais para todos os tecidos (Figura 4B). Essa correlação entre

retrocópias e genes parentais é uma medida importante da precisão de quantificação da expressão de retrocópias, pois espera-se uma ausência de correlação, dado que as retrocópias são controladas por novos promotores. Uma correlação positiva poderia indicar uma contaminação de leituras (*reads*) dos genes parentais "mapeadas", isto é, sendo usadas erroneamente para quantificar as retrocópias. Nosso resultado mostrou um destaque para Testículo, o tecido com menor correlação (*rho* = 0,0194), e Ovário, o tecido com maior correlação (*rho* = 0,1568). Nota-se que a correlação retrocópia-parental por tecido é bem baixa, com mediana de *rho* = 0,1246, Figura Suplementar 1. Também comparamos a distribuição dos valores de *rho* das retrocópias com seus respectivos parentais contra uma distribuição aleatória de valores de *rho*, calculada entre retrocópias com outros genes (que não seus parentais) aleatoriamente selecionados (Figura 4C). Essa comparação resultou em curvas quase que sobrepostas, mostrando que podemos confiar nos dados de quantificação de expressão das retrocópias, uma vez que o nosso conjunto tem um comportamento quase que idêntico ao nosso grupo controle (aleatoriamente escolhido).

Também categorizamos as retrocópias de acordo com o número de tecidos em que elas são expressas (Figura 4D). Observamos que 71,5% (2887/4038) das retrocópias são expressas em dois ou mais tecidos, 45,7% (1845/4038) são expressas em mais de um terço dos tecidos e que 34,82% (1406/4038) são expressas em mais de dois terços dos tecidos. Essa distribuição ressalta que a quantificação de expressão de retrocópias não é apenas um efeito de ruído dos algoritmos de expressão, mas que um número considerável é quantificado em vários tecidos normais humanos com padrões diversos. As retrocópias expressas em apenas 1 tecido chegam a representar 28,5% do total de retrocópias expressas, predominantemente no Testículo (74,5%), mas também no Ovário e Cerebelo. Também observamos que 636 (15,75%) das retrocópias são expressas em todos os tecidos.

Na parte E da Figura 4, apresentamos a distribuição dos níveis de expressão e o número de retrocópias expressas nos diferentes tecidos, ordenadas pela expressão mediana das retrocópias. Embora o comportamento geral das retrocópias entre os tecidos seja pouco heterogêneo, notamos, na parte inferior da Figura 4E, uma considerável variação no número de retrocópias expressas entre os tecidos.

Novamente o testículo se destaca como o tecido com o maior número de retrocópias expressas (pouco mais de 3.000 retrocópias). Acreditamos que esse fenômeno deve-se ao fato de que a cromatina fica relaxada por mais tempo e com mais frequência (sobretudo na espermatogênese) do que em outros tecidos, havendo a transcrição de todo tipo de gene, incluindo retrocópias[99]. Uma outra possível razão é que a expressão gênica no testículo é frequentemente menos restrita em termos de seleção negativa quando comparada a outros tecidos, permitindo que novos genes (no caso, retrocópias) ganhem funções. Isso é importante na seleção sexual e na competição entre espermatozoides, algo que pode impulsionar a rápida evolução de novos genes relacionados à fertilidade e suas funções nos espermatozóides. Por fim, também podemos destacar que a diversidade de papéis que o testículo desempenha, de produção de espermatozoides até a secreção hormonal, permite uma certa plasticidade que pode ser explorada evolutivamente para incorporar novos genes no repertório funcional do órgão. Dessa forma, embora novas retrocópias possam inicialmente ser expressas em diversos tecidos, a especialização funcional pode ocorrer ao longo do tempo, resultando na expressão predominante ou exclusiva dessas retrocópias (retrogenes) nos testículos.

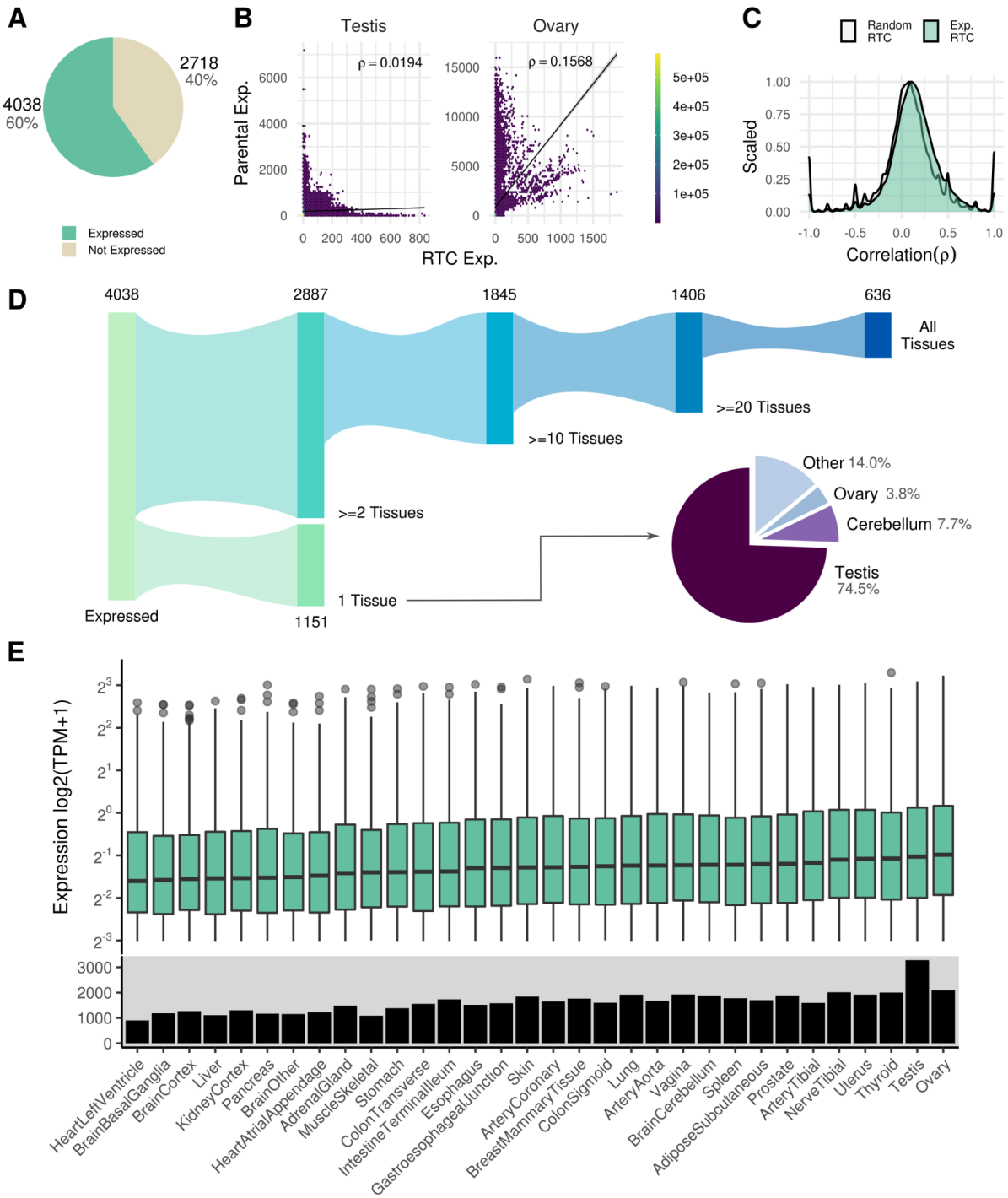



**Figura 4. Padrões de Expressão de Retrocópias em Tecidos (Normais) do GTEx.** A) Total e porcentagem de retrocópias expressas e não expressas (valor de corte de 0,1 TPM na mediana por tecido). B) Correlação de expressão retrocópia vs. parental em Testículo e Ovário. C) Distribuição dos valores de correlação de expressão retrocópias vs parental e do grupo controle parental vs. retrocópia de outro gene. D) Número de tecidos em que as retrocópias são expressas. E) Distribuição da expressão das retrocópias em cada tecido do GTEx (em cima) e número de retrocópias expressas (embaixo).

Também avaliamos  a frequência de expressão das retrocópias nas amostras de cada tecido. Para tal, consideramos uma retrocópia expressa em dada amostra com o mesmo corte utilizado no tecido como um todo (>=0,1 TPM) e calculamos a frequência de expressão como o número de amostras em que a retrocópia é expressa pelo número total de amostras do tecido.

Na Figura 5 apresentamos esse perfil de frequência de expressão das retrocópias por tecido normal. Primeiramente, observamos que as retrocópias se agrupam em três grandes grupos: retrocópias frequentemente expressas (em vermelho), infrequentemente expressas (em azul) e moderadamente expressas (em amarelo). No primeiro grupo observamos que as retrocópias são frequentemente expressas em praticamente todos os tecidos, com exceção dos tecidos cerebrais e tecidos musculares (músculo esquelético e coração). Da mesma forma, retrocópias infrequentemente expressas apresentam um comportamento uniforme em todos os tecidos à exceção de testículo, corroborando as hipóteses de que o ganho de expressão de novos genes ocorre primeiramente neste tecido. O terceiro grupo apresenta uma maior variação de padrões, mas destaca-se o cerebelo e o testículo, onde as retrocópias apresentam um padrão notavelmente distinto, sendo mais frequentemente expressas do que em outros tecidos.

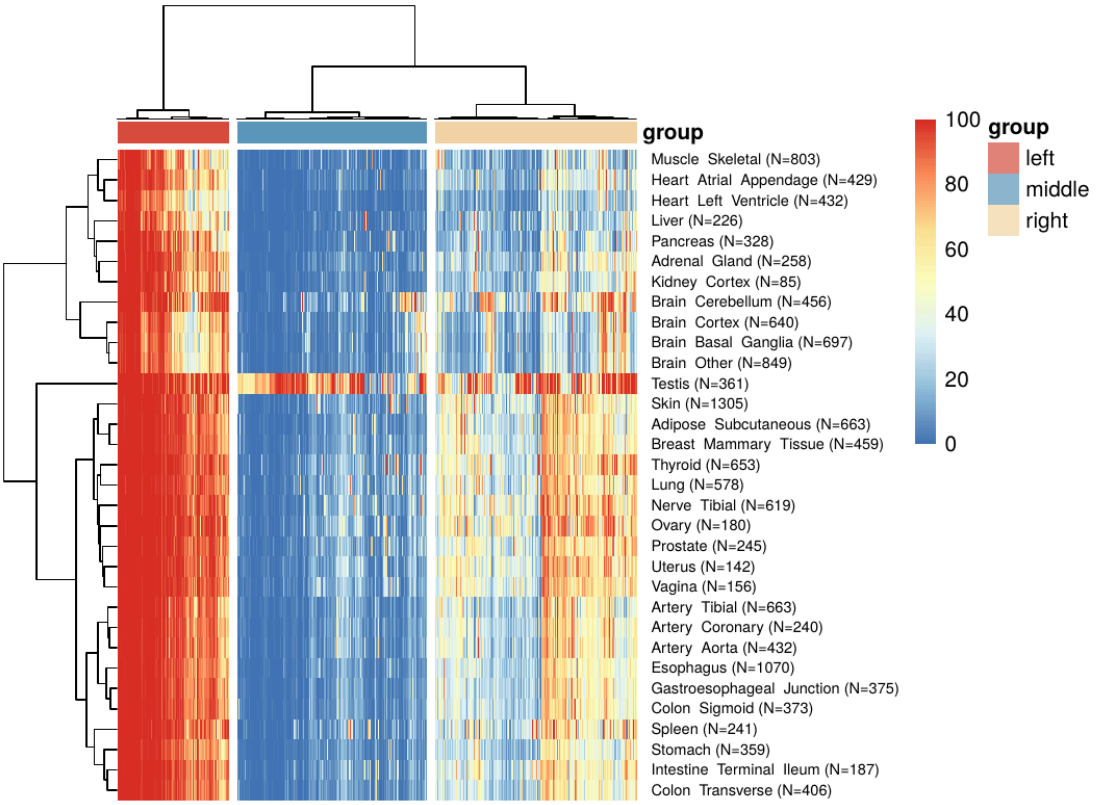



**Figura 5. Porcentagem de Amostras em que Dada Retrocópia é Considerada Expressa por Tecido (Normal) do GTEx.** *Heatmap* onde cada coluna representa uma retrocópia e cada linha representa um tecido normal. A escala (de 0 a 100) representa porcentagem de amostras em que dada retrocópia é expressa em dado tecido. Em vermelho, amarelo e azul, estão representados os três grandes agrupamentos do *heatmap*, respectivamente: expresso em alta frequência, em baixa frequência e em frequência média.

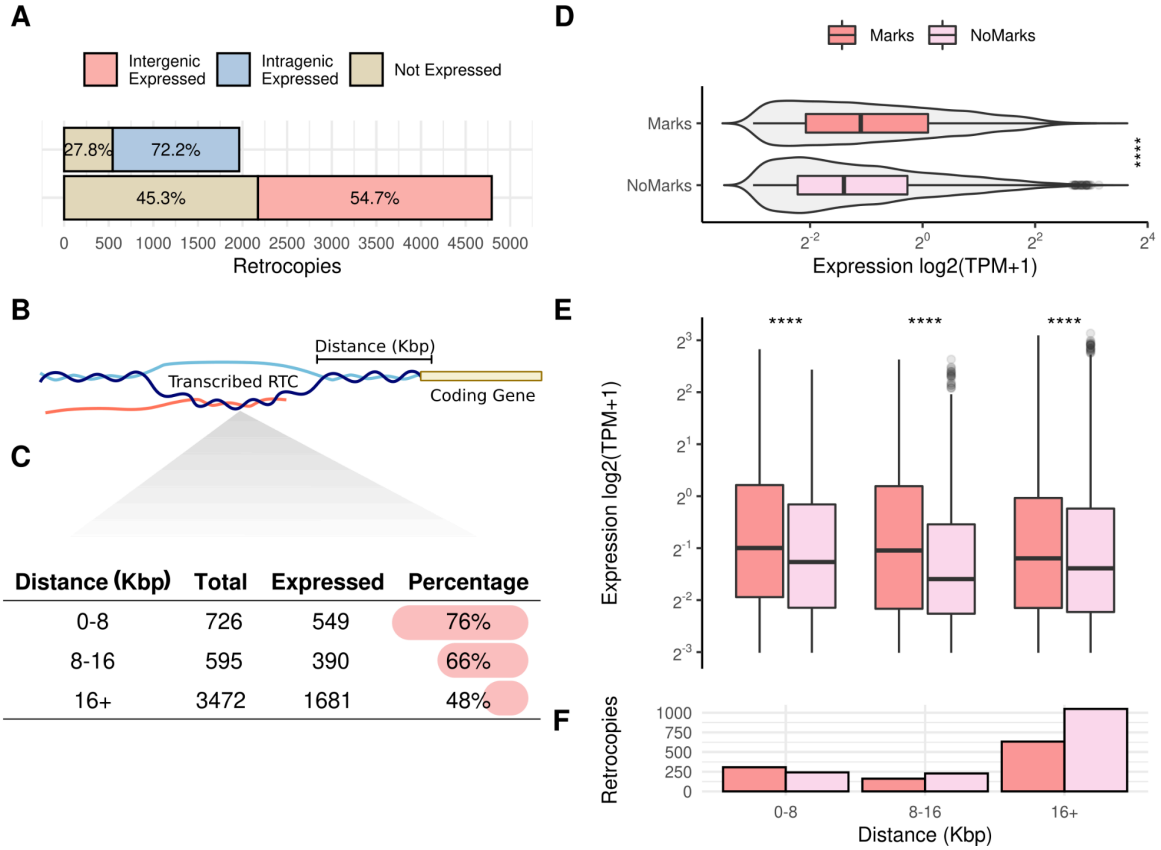
## 2.4.2. Metilação das retrocópias

Sabemos que existem diversos mecanismos de ganho de expressão de retrocópias, entre eles o uso de promotores de genes próximos[92]. Para avaliar a influência do local de inserção das retrocópias na sua expressão, separamos as retrocópias de acordo com a região genômica. Encontramos 4.793 retrocópias intergênicas e 1.963 retrocópias intragênicas, e observamos que 72.2% das retrocópias intragênicas são expressas em comparação a 54,7% das retrocópias intergênicas, uma diferença significativa (p-valor < 0.0001; chi-quadrado = 176, g.l. = 1; Figura 6A). Assumindo que todas as retrocópias intragênicas são transcritas junto com o gene no qual elas estão inseridas e, posteriormente, descartadas (ou não) no processamento pós-transcricional, é lógico esperar que tais retrocópias apareçam mais frequentemente entre aquelas expressas. Além de uma transcrição independente, tais retrocópias intragênicas também podem, eventualmente, formar

transcritos quiméricos com seus genes hospedeiros, algo que não analisamos nesta tese.

Em seguida investigamos as retrocópias intergênicas e expressas, as quais foram estratificadas de acordo com a distância ao início de transcrição (TSS) do gene codificador mais próximo (Figura 6B). Categorizamos-as em 4 classes: RTCs até 8 kb de distância, de 8 a 16 kb e acima de 16 kb de distância (Figura 6C) ao gene codificador mais próximo. Nossos dados mostraram que as retrocópias mais próximas (até 16kb) são significativamente mais transcritas (valor p < 0.0001; chi-quadrado 54.7842, g.l. = 1) (71%) do que aquelas mais distantes (48%), mostrando que além da inserção intragênicas, a proximidade a outros genes codificadores é outro fator importante para o ganho de expressão das retrocópias.

Em seguida, investigamos a relação entre a presença de marcas epigenéticas de ativação da transcrição e a expressão de retrocópias. Na Figura 6D, observamos que existe diferença estatística (W = 1.26e+08, p-value < 2.2e-16; teste de wilcoxon) entre a distribuição da expressão das retrocópias intergênicas que apresentam ao menos um pico de metilação H3K36me3 ou H3K4me3 em ao menos um tecido versus aquelas sem marcações, sendo que o primeiro grupo apresenta a mediana mais elevada. Também verificamos a diferença entre as distribuições levando em conta a posição das retrocópias (Figura 6E-F): para aquelas mais próximas (<8 kb de distância), 307 das 549 expressas apresentam ao menos uma marcação em ao menos um tecido e expressão mais elevada em comparação ao grupo sem marcadores. Portanto, conforme a distância aumenta, apesar do número de retrocópias expressas sem marcadores aumentarem, a mediana do nível de expressão diminui, enquanto que para as retrocópias com marcadores, observamos inclusive um leve aumento da amplitude de expressão com o aumento da distância. Isso mostra que a expressão das retrocópias se comporta como o esperado para genes codificadores, em que sua expressão está sujeita ao estado de metilação de histonas.




| Distance (Kbp) | Total | Expressed | Percentage |
|---|---|---|---|
| 0-8 | 726 | 549 | 76% |
| 8-16 | 595 | 390 | 66% |
| 16+ | 3472 | 1681 | 48% |

**Figura 6. Padrões de Expressão de Retrocópias Intergênicas e Intragênicas em Amostras do GTEx.** A) Total e percentual de retrocópias expressas e não expressas de acordo com a posição genômica. B) Esquema representando uma retrocópia intergênica. C) Tabela com o total e percentual de retrocópias intergênicas expressas de acordo com a distância. D) *Boxplot* da expressão de retrocópias intergênicas de acordo com a presença de ao menos um tipo de marcador (H3K4me3 ou H3K36me3) em ao menos um tecido. E) Distribuição da expressão de retrocópias intergênicas de acordo com a distância de genes codificadores e presença de marcadores. F) Gráfico de barras com o número de retrocópias em cada categoria da parte E.

## 2.4.3. Abrangência de expressão das retrocópias

Em seguida, calculamos a abrangência de expressão (geralmente representada pela letra grega *tau*)[94]. Na Figura 7A, avaliamos a abrangência de expressão das retrocópias de acordo com a posição genômica (intragênica ou intergênica). Observamos que as retrocópias intragênicas apresentam padrões de expressão mais abrangentes (mediana de 0,83) do que as intergênicas (mediana de 0,96). Curiosamente, as retrocópias intergênicas apresentam padrões de expressão mais tecido-específicos quanto mais distantes de genes codificadores - para distâncias maiores do que 16 kb, a mediana é de 0,97.

Em adição à posição genômica, também escolhemos dividir as retrocópias entre aquelas que apresentam baixa expressão (de 0,1 até 0,3 TPM), média expressão (de 0,3 até 0,8 TPM) e alta expressão (acima 0,8 TPM) de acordo com a distribuição dos níveis de expressão das mesmas em todos os tecidos normais (Figura 7B). Os cortes foram escolhidos de acordo com a média dos quantis (de 25 até 50%, de 50 até 75% e acima de 75%).

Na Figura 7C, avaliamos o comportamento da abrangência de expressão das retrocópias de acordo com o seu nível de expressão e sua posição genômica. As retrocópias com baixa expressão são mais tecido específicas (mediana de *tau* variando de 0,93 a 0,98), enquanto que retrocópias com alta expressão são mais abrangentes (mediana de tau variando de 0,73 a 0,90). Em adição, a posição genômica se mantém como forte fator preditivo do padrão da abrangência de expressão, onde retrocópias intragênicas independente do nível de expressão são mais abrangentes do que as retrocópias intergênicas.

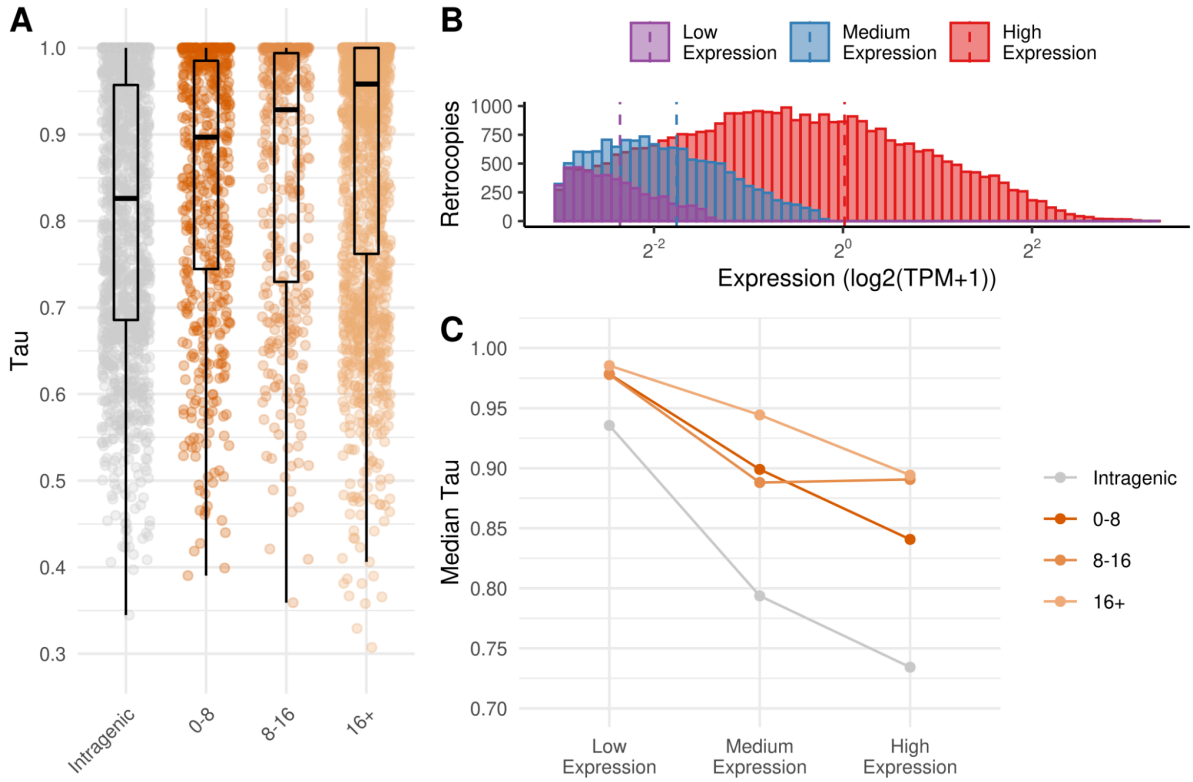



**Figura 7. Abrangência de Expressão de Retrocópias em Amostras de Tecidos Normais do GTEx de Acordo com a Posição Genômica.** A) *Boxplot* da abrangência de expressão (*tau*) das retrocópias de acordo com o contexto genômico. B) Histograma da distribuição de expressão das retrocópias de acordo com a faixa de expressão: baixa, média, alta. C) *Boxplot* da abrangência de expressão (*tau*) das retrocópias de acordo com o contexto genômico e de acordo com a faixa de expressão (baixa, média ou alta).

Para avaliar como a idade das retrocópias afeta seus padrões de abrangência de expressão, utilizamos o resultado de ortologia das retrocópias de acordo com seção 1.4.2. (Figura 8). Observamos que a correlação entre a abrangência de expressão da retrocópia com o seu parental é não significativa para humano específico (rho = 0.1141582, p-value=0.2144, spearman) e para retrocópias compartilhado com primatas do velho mundo (rho = 0.05085292, p-value=0.1968, spearman), mas é correlacionada positivamente de forma significativa para retrocópias compartilhadas com os grandes primatas (*great apes*) (rho = 0.1139344, p-value=0.02183, spearman) e compartilhadas com não primatas (rho = 0.1460327 p-value=0.02549, spearman). Já as retrocópias compartilhadas com todos os primatas (rho = -0.03906066, p-value=0.02709, spearman) apresentam uma correlação negativa significativa. No geral, esses resultados reforçam que os padrões de expressão são distintos entre as retrocópias e os parentais - como deveria ser do ponto de vista biológico - e que alterações na abrangência de expressão podem indicar novas funcionalidades das retrocópias em comparação ao parental. Vale ressaltar que a maioria das retrocópias humanas são compartilhadas entre todos os primatas e que a correlação negativa desse conjunto reflete portanto o comportamento esperado para a maioria das retrocópias. Também para as retrocópias mais jovens (humano específica) a relação não significativa nos garante a qualidade do dado de quantificação, uma vez que a sequência dessas retrocópias ainda são muito similares às sequências dos seus genes parentais. Curiosamente, a melhor correlação entre o *tau* das retrocópias e dos parentais foi observada para as retrocópias mais antigas, possivelmente indicando uma maior funcionalidade, seja com funções mais abrangentes ou especializadas, dessas retrocópias.

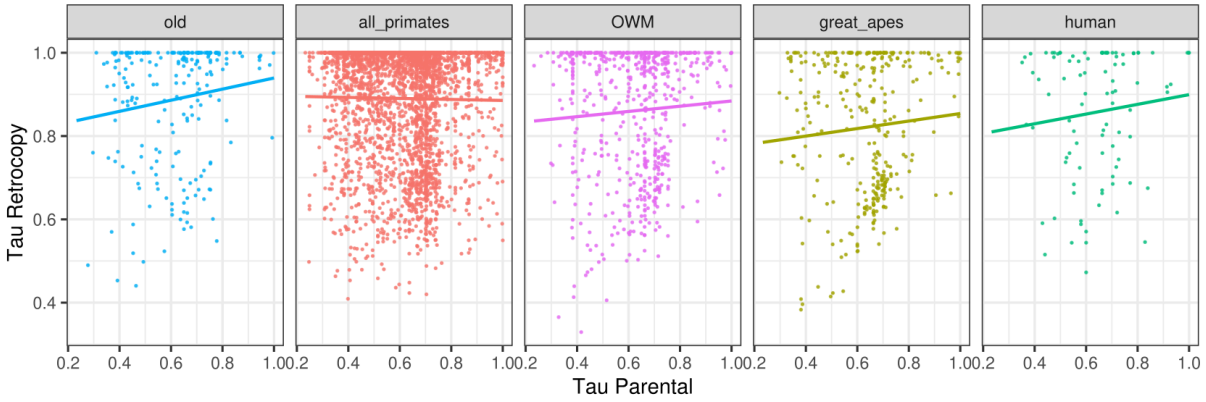



**Figura 8. Scatterplot dos Valores de tau das Retrocópias e de seus Genes Parentais.** Da esquerda para a direita: retrocópias compartilhadas com não primatas, primatas, primatas do velho mundo, grandes primatas e Humano específica.

## 2.4.4. Perfil de RiboSeq das retrocópias

A metodologia de sequenciamento de fragmentos processados por ribossomos (RiboSeq) permite identificar regiões de iniciação do mRNA, regiões de alongamento e áreas de paralisação da tradução. Essa metodologia tem sido utilizada para responder a uma ampla gama de questões, desde a identificação de pequenos quadros de leitura abertos traduzidos (ORFs)[100], quantificação do controle translacional[101] e a elucidação de mecanismos do próprio processo de tradução[102]. Protocolos de Ribo-seq se baseiam em uma ideia simples: fragmentos processados por ribossomos (RPFs) são protegidos da digestão com RNase e podem, portanto, ser isolados e sequenciados. Em resumo, o protocolo Riboseq consiste em (i) tratamento e colheita de células, (ii) *nuclease footprinting* e isolamento de RPF e (iii) preparação e sequenciamento da biblioteca.

O banco de dados RPFdb[95] reúne experimentos disponibilizados publicamente de alta qualidade para muitas espécies e aplica uma *pipeline* de alinhamento, quantificação e predição de ORFs padronizada, considerada como o padrão-ouro na literatura atual[103]. Para humanos, o banco de dados disponibiliza 3 tipos de tecidos saudáveis: 3 amostras de cérebro, 8 de rim e 1 de músculo esquelético.

Dado o nosso conjunto de 4.038 retrocópias expressas em tecidos normais, verificamos que 1.077 (26,9%) também possuem evidência de tradução nas amostras do RPFdb (Figura 9B). Verifica-se que um conjunto distinto de retrocópias é traduzido em cada tecido (Figura 9A): 27 retrocópias em músculo esquelético,

sendo que este é o tecido com menor número de amostras, 709 retrocópias em cérebro e 770 retrocópias em rim. Do total, 407 (37,79%) retrocópias são traduzidas em dois ou mais tecidos, indicando que esse subconjunto pode ser mais abrangente. De fato, a mediana de tau desse conjunto é de 0,83, equivalente às retrocópias intragênicas apesar de 70% (286) dessas retrocópias serem intergênicas. Com relação aos *counts* de RiboSeq, o cérebro é o tecido com a maior evidência de tradução das retrocópias (Figura 9C), com uma distribuição que é significativamente diferente de rim (W = 576565, p-value = 7.432e-12) e de músculo esquelético (W = 12973, p-value = 0.001708).

Além de quantificar as taxas de tradução dos genes, o sequenciamento de Ribo-seq é uma técnica que possibilita a anotação de regiões CDS com base em dados experimentais. Desde os primeiros estudos[104,105], dados de Ribo-seq foram usados para detectar a tradução de quadros de leitura abertos *upstream* (uORFs), o uso de códons de início não canônicos ou a tradução de RNAs presumivelmente não codificantes (ncRNAs). Nesse contexto, é possível realizar a anotação das CDS das retrocópias, traduzidas a partir dos experimentos de Ribo-seq. Essa CDS predita pode então ser comparada com a CDS anotada dos genes parentais. Na figura 9D, observamos a distribuição de cobertura e identidade dessa comparação, onde observamos que as retrocópias são frequentemente truncadas (menos de 50% de cobertura), mas com identidade alta (mais de 75%) em relação a seus genes parentais. Essas mudanças de CDS com relação aos parentais (mais curtas e não idênticas) podem indicar uma alteração da função das retrocópias, incluindo mudanças de domínio proteico. Vale ressaltar que os ribossomos desempenham um papel fundamental na síntese de proteínas, tornando o método Ribo-seq uma ponte poderosa entre técnicas de transcriptômica e proteômica. Esses elos contribuem para aprimorar a anotação de genes e a detecção de proteínas[106].

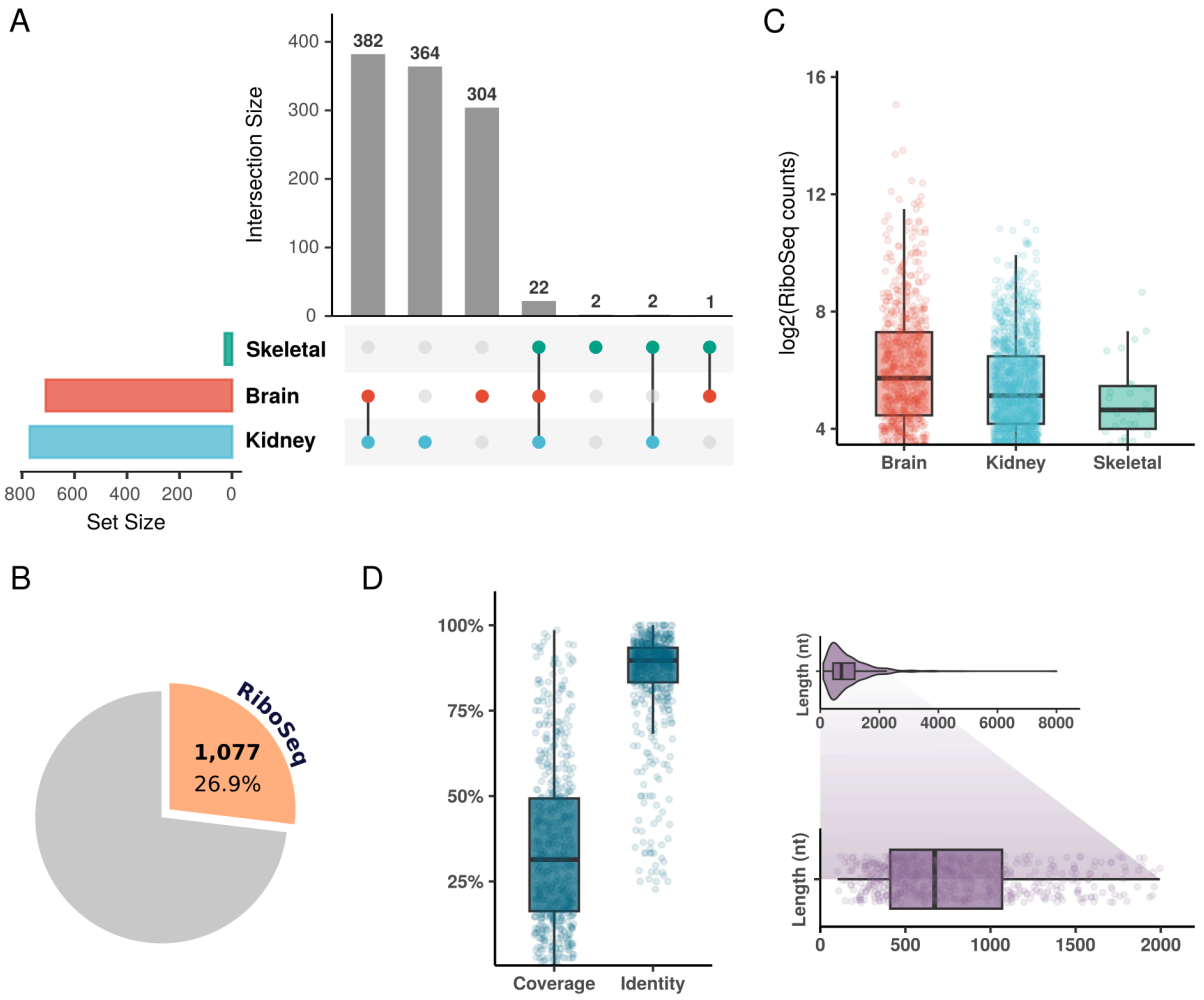



**Figura 9. Retrocópias são Traduzidas em Proteínas.** A) Retrocópias com evidência de tradução ativa por RiboSeq em 3 tecidos humanos: músculo esquelético, cérebro e rim. B) Porcentagem de retrocópias expressas com evidência de tradução ativa por RiboSeq. (C) Distribuição de contagens de RiboSeq que apoiam evidências de tradução para retrocópias em tecidos musculares esqueléticos, cerebrais e renais. (D) Comparação entre ORFs de retrocópia previstas por RiboSeq e ORFs de genes parentais correspondentes: cobertura de alinhamento, identidade e comprimento.

## 2.5 CONSIDERAÇÕES FINAIS

Nossa análise da expressão das retrocópias fixadas em humanos revelou *insights* valiosos sobre esses elementos genéticos e seu papel nos tecidos humanos. Para realizar essa análise, utilizamos uma abordagem criteriosa, começando pela identificação das retrocópias fixadas em humanos, seguida pela atribuição de genes parentais a essas retrocópias com base nos dados do GTEx v8. Isso resultou na seleção de 6.756 retrocópias para a nossa análise.

Uma etapa importante desse processo envolveu a seleção dos tecidos para análise. Optamos por focar em 32 tecidos relevantes e representativos, excluindo tecidos com um número limitado de amostras ou que não apresentaram diferenças significativas nos perfis de expressão. Esse agrupamento nos permitiu obter uma visão abrangente dos padrões de expressão das retrocópias nos tecidos humanos.

É importante reconhecer que a expressão gênica é altamente variável entre os tecidos, e essa variabilidade se estende às retrocópias. Descobrimos que muitas retrocópias exibem padrões de expressão intermediários, sendo expressas em alto nível em alguns tecidos e em níveis muito mais baixos ou ausentes em outros. Esse conceito de "abrangência" de expressão gênica destaca a complexidade dos padrões de expressão e a importância de considerar não apenas os níveis gerais de expressão, mas também a distribuição nos diferentes tecidos.

Além disso, investigamos como a localização genômica das retrocópias e seu estado de metilação podem influenciar sua expressão. Observamos que as retrocópias intragênicas tendem a ser mais expressas do que as retrocópias intergênicas, o que era esperado. Além disso, a proximidade de outros genes também desempenha um papel importante na expressão das retrocópias intergênicas, com retrocópias próximas a genes codificadores exibindo maior expressão.

A análise do estado de metilação das retrocópias revelou que aquelas com marcações de H3K4me3 e H3K36me3 tendem a apresentar níveis mais elevados de expressão. Isso destaca a influência da epigenética na regulação da expressão das retrocópias, semelhante ao que ocorre com os genes codificadores.

Ao explorar a relação entre a idade das retrocópias e seus padrões de expressão, identificamos que a correlação entre a abrangência de expressão das retrocópias e seus genes parentais é geralmente fraca. Isso sugere que as retrocópias podem ter desenvolvido padrões de expressão únicos e independentes de seus genes parentais ao longo da evolução.

Além disso, nossos resultados demonstram que algumas retrocópias são traduzidas em proteínas, como evidenciado pelos dados de RiboSeq em tecidos humanos. Isso desafia a visão convencional de que as retrocópias são principalmente elementos genéticos não funcionais e sugere que algumas delas

podem desempenhar papéis importantes na regulação e diversificação do proteoma humano.

Em resumo, nossa análise aprofundada dos padrões de expressão das retrocópias em uma variedade de tecidos humanos fornece informações valiosas sobre a diversidade e complexidade desses elementos genéticos. Essas descobertas ressaltam a complexidade da regulação genética e destacam o potencial funcional das retrocópias fixadas em humanos, contribuindo para uma compreensão mais profunda da biologia humana.

# REFERÊNCIAS


1. Kutschera U, Niklas KJ. The modern theory of biological evolution: an expanded synthesis. Naturwissenschaften. 2004;91: 255–276.

2. Creighton HB, McClintock B. A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. Proc Natl Acad Sci U S A. 1931;17: 492–497.

3. Coe E, Kass LB. Proof of physical exchange of genes on the chromosomes. Proc Natl Acad Sci U S A. 2005;102: 6641–6646.

4. McCLINTOCK B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36: 344–355.

5. Vanin EF. Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 1985;19: 253–272.

6. Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. FEBS Lett. 2000;468: 109–114.

7. Podlaha O, Zhang J. Processed pseudogenes: the “fossilized footprints” of past gene expression. Trends Genet. 2009;25: 429–434.

8. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 2009;10: 19–31.

9. Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet. 2020;21: 191–201.

10. Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L, et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol. 2013;14: R22.

11. Richardson SR, Salvador-Palomeque C, Faulkner GJ. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. Bioessays. 2014;36: 475–481.

12. Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, et al. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res. 2013;23: 2042–2052.

13. Schrider DR, Navarro FCP, Galante PAF, Parmigiani RB, Camargo AA, Hahn MW, et al. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 2013;9: e1003242.

14. Doolittle WF. We simply cannot go on being so vague about “function.” Genome biology. 2018. p. 223.

15. Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, et al. Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci U S A. 2014;111: 13361–13366.

16. Kafri R, Springer M, Pilpel Y. Genetic redundancy: new tricks for old genes. Cell. 2009;136: 389–392.

17. Ohno S. Evolution by Gene Duplication. Springer Science & Business Media; 2013.

18. Ohno S. So much “junk” DNA in our genome. In “Evolution of Genetic Systems.” Brookhaven Symposium in Biology. 1972;23: 366–370.

19. Bridges CB. SALIVARY CHROMOSOME MAPS: With a Key to the Banding of the Chromosomes of Drosophila Melanogaster. J Hered. 1935;26: 60–64.

20. Muller HJ. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetica. 1935;17: 237–252.

21. Haldane JBS. The Part Played by Recurrent Mutation in Evolution. Am Nat. 1933;67: 5–19.

22. Li W-H, Graur D. Fundamentals of Molecular Evolution. Sinauer Associates; 1991.

23. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003;4: 865–875.

24. Jacq C, Miller JR, Brownlee GG. A pseudogene structure in 5S DNA of Xenopus laevis. Cell. 1977;12: 109–120.

25. Proudfoot NJ, Maniatis T. The structure of a human alpha-globin pseudogene and its relationship to alpha-globin gene duplication. Cell. 1980;21: 537–544.

26. Wilde CD, Crowther CE, Cowan NJ. Diverse mechanisms in the generation of human beta-tubulin pseudogenes. Science. 1982;217: 549.

27. Karin M, Richards RI. Human metallothionein genes--primary structure of the metallothionein-II gene and a related processed gene. Nature. 1982;299: 797–802.

28. Lemischka I, Sharp PA. The sequences of an expressed rat alpha-tubulin gene and a pseudogene with an inserted repetitive element. Nature. 1982;300: 330–335.

29. Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. Science. 1991;254: 1808–1810.

30. Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996;87: 905–916.

31. Weichenrieder O, Repanas K, Perrakis A. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure. 2004;12: 975–986.

32. Maita N, Anzai T, Aoyagi H, Mizuno H, Fujiwara H. Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. J Biol Chem. 2004;279: 41067–41076.

33. Maita N, Aoyagi H, Osanai M, Shirakawa M, Fujiwara H. Characterization of the sequence specificity of the R1Bm endonuclease domain by structural and biochemical studies. Nucleic Acids Res. 2007;35: 3918–3927.

34. Repanas K, Zingler N, Layer LE, Schumann GG, Perrakis A, Weichenrieder O. Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. Nucleic Acids Res. 2007;35: 4914–4926.

35. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2002;31:

159–165.

36. Sen SK, Huang CT, Han K, Batzer MA. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res. 2007;35: 3741–3751.

37. Begg CE, Delius H, Leader DP. Duplicated region of the mouse genome containing a cytoplasmic gamma-actin processed pseudogene associated with long interspersed repetitive elements. J Mol Biol. 1988;203: 677–687.

38. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000;24: 363–367.

39. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.

40. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008;5: 16–18.

41. Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene. 2001;270: 17–30.

42. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, et al. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res. 2002;12: 272–280.

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

44. Zhang Z, Harrison P, Gerstein M. Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome. Genome Res. 2002;12: 1466–1482.

45. Zhang Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 2003;13: 2541–2558.

46. Torrents D, Suyama M, Zdobnov E, Bork P. A genome-wide survey of human pseudogenes. Genome Res. 2003;13: 2559–2567.

47. Khelifi A, Duret L, Mouchiroud D. HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res. 2005;33: D59–66.

48. Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene. 2007;389: 196–203.

49. Fukuda A, Kodama Y, Mashima J, Fujisawa T, Ogasawara O. DDBJ update: streamlining submission and access of human data. Nucleic Acids Res. 2021;49: D71–D75.

50. Yu Z, Morais D, Ivanga M, Harrison PM. Analysis of the role of retrotransposition in gene evolution in vertebrates. BMC Bioinformatics. 2007;8: 308.

51. Liu Y-J, Zheng D, Balasubramanian S, Carriero N, Khurana E, Robilotto R, et al. Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of

retrotrans-positional activity. BMC Genomics. 2009;10: 480.

52. Balasubramanian S, Zheng D, Liu Y-J, Fang G, Frankish A, Carriero N, et al. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biol. 2009;10: R2.

53. Navarro FCP, Galante PAF. RCPedia: a database of retrocopied genes. Bioinformatics. 2013;29: 1235–1237.

54. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–45.

55. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006.

56. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12: 656–664.

57. Kabza M, Ciomborowska J, Makałowska I. RetrogeneDB--a database of animal retrogenes. Mol Biol Evol. 2014;31: 1646–1648.

58. Rosikiewicz W, Kabza M, Kosinski JG, Ciomborowska-Basheer J, Kubiak MR, Makalowska I. RetrogeneDB-a database of plant and animal retrocopies. Database . 2017;2017. doi:10.1093/database/bax038

59. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [version 1; peer review: 2. doi:10.12688/f1000research.23297.1

60. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21: 487–493.

61. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

62. RepeatMasker Home Page. [cited 2 Oct 2023]. Available: http://www.repeatmasker.org

63. Bredeson JV, Mudd AB, Medina-Ruiz S, Mitros T, Smith OK, Miller KE, et al. Conserved chromatin and repetitive patterns reveal slow genome evolution in frogs. bioRxiv. 2021. p. 2021.10.18.464293. doi:10.1101/2021.10.18.464293

64. Chang N-C, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. Genome Res. 2022;32: 1408–1423.

65. Improved pairwise alignment of genomic DNA. 2007. Available: https://search.proquest.com/openview/bc77cca0fb9390b44b9ef572fb574322/1?pq-origsite=gscholar&cbl=18750

66. Fukushima K, Pollock DD. Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. Nat Commun. 2020;11: 4459.

67. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34: 525–527.

68. Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 2015;7: 567–580.

69. Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. Retrocopy contributions to the evolution of the human genome. BMC Genomics. 2008;9: 466.

70. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2020;48: D84–D86.

71. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics. 2006;22: 1437–1439.

72. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47: D766–D773.

73. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20: 1313–1326.

74. Casola C, Betrán E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? Genome Biol Evol. 2017;9: 1351–1373.

75. Okamura K, Nakai K. Retrotransposition as a source of new promoters. Mol Biol Evol. 2008;25: 1231–1238.

76. Fablet M, Bueno M, Potrzebowski L, Kaessmann H. Evolutionary origin and functions of retrogene introns. Mol Biol Evol. 2009;26: 2147–2156.

77. Navarro FCP, Galante PAF. A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biol Evol. 2015;7: 2265–2275.

78. Kubiak MR, Makałowska I. Protein-Coding Genes' Retrocopies and Their Functions. Viruses. 2017;9. doi:10.3390/v9040080

79. Burki F, Kaessmann H. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet. 2004;36: 1061–1063.

80. Sayah DM, Sokolskaja E, Berthoux L, Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature. 2004;430: 569–573.

81. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465: 1033–1038.

82. Karreth FA, Reschke M, Ruocco A, Ng C, Chapuy B, Léopold V, et al. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell. 2015;161: 319–332.

83. Prendergast GC. Actin' up: RhoB in cancer and apoptosis. Nat Rev Cancer. 2001;1: 162–168.

84. Miller TLA, Orpinelli F, Buzzo JLL, Galante PAF. sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa689

85. Wong K, Adams DJ, Keane TM. Go retro and get a GRIP. Genome Biol. 2013;14: 108.

86. de Boer M, van Leeuwen K, Geissler J, Weemaes CM, van den Berg TK, Kuijpers TW, et al. Primary immunodeficiency caused by an exonized retroposed gene copy inserted

in the CYBB gene. Hum Mutat. 2014;35: 486–496.

87. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JMC, et al. Processed pseudogenes acquired somatically during cancer development. Nat Commun. 2014;5: 3644.

88. GTEx Portal. [cited 2 Oct 2023]. Available: https://gtexportal.org/home/

89. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire M-D, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28: 1530–1532.

90. Paulson JN, Chen C-Y, Lopes-Ramos CM, Kuijjer ML, Platig J, Sonawane AR, et al. Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. BMC Bioinformatics. 2017;18: 437.

91. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46: D794–D801.

92. Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 2016;26: 301–314.

93. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348: 648–660.

94. [No title]. [cited 2 Oct 2023]. Available: https://academic.oup.com/bioinformatics/article/21/5/650/220059

95. Xie S-Q, Nie P, Wang Y, Wang H, Li H, Yang Z, et al. RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. 2016;44: D254–8.

96. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

97. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

98. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015;4: e08890.

99. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 2013;3: 2179–2190.

100. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014;33: 981–993.

101. Schafer S, Adami E, Heinig M, Rodrigues KEC, Kreuchwig F, Silhavy J, et al. Translational regulation shapes the molecular landscape of complex disease phenotypes. Nat Commun. 2015;6: 7200.

102. Andreev DE, O'Connor PBF, Loughran G, Dmitriev SE, Baranov PV, Shatsky IN.

Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. Nucleic Acids Res. 2017;45: 513–526.

103. Calviello L, Ohler U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. Trends Genet. 2017;33: 728–744.

104. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324: 218–223.

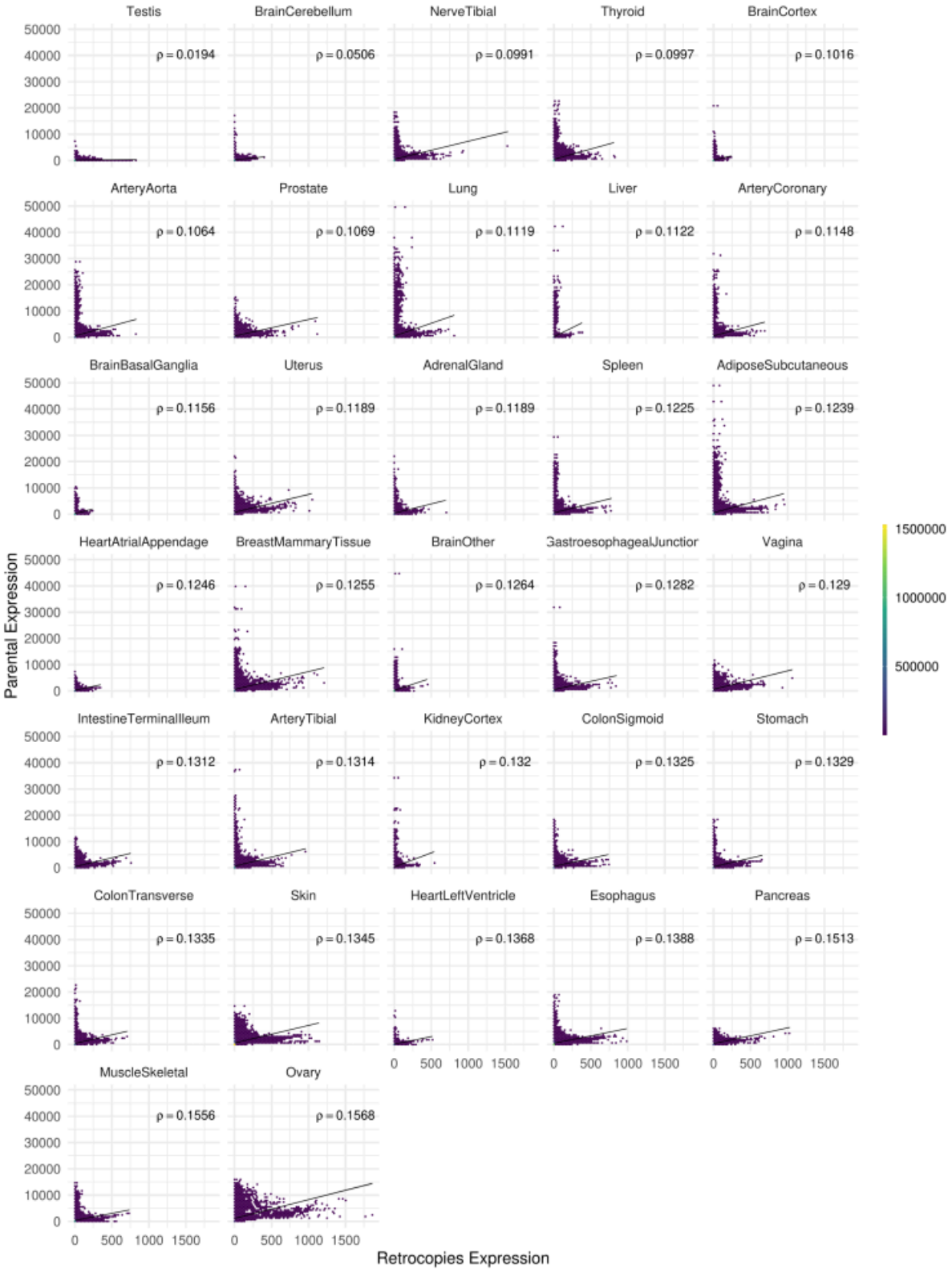
105. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147: 789–802.

106. Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, et al. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res. 2015;43: e29.

# LISTA DE ANEXOS

# ANEXO A

## Figura Suplementar 1



ANEXO B

# SÚMULA CURRICULAR

**DADOS PESSOAIS**
Nome: Helena Beatriz da Conceição
Local e data de nascimento: Jundiaí SP, 21/12/1994

**EDUCAÇÃO**
Universidade de São Paulo, USP, Brasil (2013-2017)
Graduação em Ciências Físicas e Biomoleculares

University of Nottingham, England, United Kingdom (2015-2016)
Graduação Sanduíche em Natural Sciences pelo Ciências sem Fronteiras

Universidade de São Paulo, USP, Brasil (2018-atual)
Doutorado direto em Bioinformática

**OCUPAÇÃO**
Bolsista Institucional ICMC - USP. Monitora da disciplina de Cálculo II, 2014-2015
Bolsista Ciências sem Fronteiras, CNPq, 2015-2016
Bolsista de Doutorado, CAPES, 2018-2019
Bolsista de Doutorado, FAPESP, 2019-2023
Bolsista PAE (Programa de Aperfeiçoamento de Ensino) da USP como monitora da disciplina MAC0113 Introdução à Computação para Ciências Humanas ministrada pelo professor Flavio Soares Correa da Silva, 1º Semestre de 2022

**PUBLICAÇÕES**
Barreiro, R., Guardia, G. D., Meliso, F. M., Lei, X., Li, W.-Q., Savio, A., Fellermeyer, M., **Conceição, H. B.**, Mercuri, R. L., Landry, T., Qiao, M., Blazquez, L., Ule, J., Penalva, L. O. F.; Galante, P. A. (2023). **The paralogues MAGOH and MAGOHB are oncogenic factors in high-grade gliomas and safeguard the splicing of cell division and cell cycle genes**. RNA Biology, 20(1), 311–322. https://doi.org/10.1080/15476286.2023.2221511

Rafael L. Mercuri, **Helena B. Conceição**, Gabriela D. A. Guardia, Gabriel Goldstein, Maria D. Vibranovski, Ludwig C. Hinske, Pedro A F Galante. (2023). **Retro-miRs: Novel and functional miRNAs originated from mRNA retrotransposition.** Mobile DNA 14, 12 (2023). https://doi.org/10.1186/s13100-023-00301-w - **Primeira autoria compartilhada**

Miller, T.L.A., **Conceicao, H.B.**, Mercuri, R. L., Santos, F.R.C., Barreiro, R., Buzzo, L. J., Rego, F.O., Guardia, G.D.A., Galante, P.A.F. (2023). **Sandy: A user-friendly and**

**versatile NGS simulator to facilitate sequencing assay design and optimization.** bioRxiv. https://doi.org/10.1101/2023.08.25.554791 - **Primeira autoria compartilhada**

ANEXO C

# Certificados de Participação em Congressos



X-meeting 2018

14th International Conference of the AB3C São Pedro, October 24th - 26th

**Certificate of Poster presentation**

This certifies that the work entitled **Retrocopies and Genetic Expression**, authored by ***Helena Beatriz da Conceicao, Gabriela Guardia and PEDRO A F GALANTE*** was presented by Helena Beatriz da Conceicao during the Poster session of the X-Meeting 2018 - 14th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in São Pedro - Brazil between 24th and 26th October of 2018.

São Pedro, 26th October 2018.

Alan M. Durham
AB3C President

Raphael Tavares da Silva
X-meeting 2018 Poster chair

Robson Francisco de Souza
X-meeting 2018 Poster chair



X-meeting 2019

**Certificate of Poster presentation**

This certifies that the work entitled **Human Retrocopies and Genetic Expression in Tumor and Normal Tissues**, authored by ***Helena Beatriz da Conceicao, Gabriela Der Agopian Guardia and Pedro A F Galante*** was presented by Helena Beatriz da Conceicao during the Poster session of the X-Meeting 2019 - 15th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (AB3C), held in Campos do Jordão - Brazil between October 30th and November 01st, 2019.

Campos do Jordão, 01st November 2019.

Ney Lemke
AB3C President

Alexandre Paschoal
Poster Chair

CSH

# CERTIFICATE OF ATTENDANCE

## Helena Beatriz da Conceicao

Was a participant in the virtual Cold Spring Harbor Laboratory Conference

***Transposable Elements***

October 6 – October 9, 2020

She gave a virtual poster presentation on entitled:

*"Identification of fixed retrocopies in mammalian genomes"*

Shannon Silva, Senior Conference Coordinator

**To Whom It May Concern:**

Helena Beatriz da Conceicao presented a poster titled *Identifying Retrocopies in Mammalian Genomes* at the Virtual Mobile DNA Conference June 8-9, 2021. For more information re the FASEB events, please visit www.faseb.org

Sincere Regards,
Joanna U. Engstrom, VEMM

Conference Manager
301-634-7009 (O)
Email: jengstrom@faseb.org
Web: www.faseb.org

FASEB

FASEB
Federation of American Societies For Experimental Biology

# CERTIFICATE OF PARTICIPATION

**This certifies that the following individual presented at a FASEB Science Research Conference:**

The Mobile DNA Conference: Evolution, Diversity, and Impact

Helena Beatriz da Conceição
***Name***

IME - University of São Paulo and IEP - Hospital Sírio Libanês
***Affiliation***

RETROCOPIES: ALIVE AND FUNCTIONAL GENE COPIES.
***Abstract Title***

June 5, 2022 - June 9, 2022
***Conference Dates***

**Frank Krause, CAE**
**Executive Director, FASEB**

## CERTIFICATE OF ATTENDANCE

**Helena Beatriz da Conceicao**

Was a participant at the Cold Spring Harbor Laboratory Conference

**The Biology of Genomes**

May 9, 2023 to May 13, 2023

And presented the poster

*Retrocopies—GENE copies identified in vertebrates and invertebrates' genomes and their orthology*

CSH

Samantha Mastronardi, Conference Coordinator