UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA

Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

FÁBIO CASSAROTTI PARRONCHI NAVARRO

A retrotransposição de mRNAs como fator de variabilidade genética no genoma humano e de outros primatas

Tese defendida

São Paulo

Data do Depósito na SPG:

12/08/2014

FÁBIO CASSAROTTI PARRONCHI NAVARRO

A retrotransposição de mRNA como fator de variabilidade genética no genoma humano e de outros primatas

Tese apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioquímica)

Orientadora: Dra. Anamaria Aranha Camargo

Co-orientador: Dr. Pedro A. F. Galante

São Paulo

2014

Ficha Catalográfica
Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP.

Navarro, Fábio Cassarotti Parronchi
N322r A retrotransposição de mRNA como fator de variabilidade genética no genoma humano e de outros primatas / Fábio Cassarotti Parronchi Navarro. -- São Paulo, 2014.
163p.

Tese (doutorado) - Instituto de Química da Universidade de São Paulo. Departamento de Biquímica.
Orientador: Camargo, Anamaria Aranha
Co-orientador : Galante, Pedro Alexandre Favoretto

1. Genoma : Biologia molecular 2. Bioinformatica I. T. II. Camargo, Anamaria Aranha, orientador. III. Galante, Pedro Alexandre Favoretto, co-orientador

574.88 CDD

Folha de aprovação

Dedico esta tese à Camila Olivato Navarro, quem deu força nas horas mais difíceis, aos meus parentes quem sempre apoiaram e nortearam minhas escolhas e aos que acharam que não conseguiria.

AGRADECIMENTOS

Aos que me orientaram nesta jornada, em especial ao Dr. Pedro Alexandre Favoretto Galante quem me mostrou que não há resultado sem trabalho árduo. Quem, me acolheu e deu total liberdade para desenvolver meu trabalho, interferindo somente quando necessário e quem sempre me orientou pelas escolhas da carreira acadêmica. Obrigado!

À Dra. Anamaria Aranha Camargo, pelo apoio, confiança e por ter nos acolhido em um momento difícil. Também agradeço profundamente, por sempre estar disponível quando precisei da ajuda, orientação ou norte.

Ao Prof. Dr. Diogo Meyer, pelas discussões sobre genômica evolutiva.

Ao Gustavo Starvaggi França, pela amizade, discussões, colaborações e quem mostrou o quanto a serenidade pode ser valiosa.

Ao Andrei Rozanski, pela rara amizade e sinceridade.

Ao Daniel Takatori Ohara, pelos bons momentos, pela paciência e pelo excelente trabalho.

À Paula Asprino, Paola Carpinetti, Luis Felipe Campesato, Ana Paula Urlass, Camila Ramos, Juliana Quintanilha, e Fernanda Koyama pela amizade e colaboração. Além de termos momentos memoráveis, cada um de vocês foi fundamental para minha formação como pesquisador.

Ao Andrei Rozanski e à Ana Paula Urlass, pela paciência, pelo carinho e por me ensinarem mais do que poderia ensinar.

À Camila Olivato Navarro a quem estas poucas linhas não fazem jus ao quanto sou agradecido. Sem você, definitivamente, não teria conseguido.

Aos meus parentes, Roberto Parronchi Navarro, Cirena Cassarotti Navarro, ao meu irmão, Felipe Cassarotti Parronchi Navarro, por tudo, sempre.

Aos meus parentes mais recentes, Sandra Regina Moreira Olivato, Valentin Bráz Olivato, pelo apoio e carinho.

Ao meu cunhado Rafael Henrique Olivato, pela companhia, pelas inúmeras discussões científicas e pseudo-científicas.

Aos meus amigos de graduação, Alexandre Yukio Harano (Frank), Gabriel Marcondes (GG), Luiz Carlos Irber Junior (Gaúcho), Mario Junior (Gretchen), Diogo Kato (Xupeta), Carlos Eduardo Ki Lee (Kossa), Bruna Milaré e Raphael Nunes (Aphalapha), eu não seria quem sou sem a amizade de vocês.

À Universidade de São Paulo pelos fundamentos em ciências biológicas.

Ao Hospital Sírio-Libanês e ao Instituto Ludwig de Pesquisa sobre o Câncer, pela excelente infra-estrutura disponibilizada para a realização deste trabalho.

A CNPQ e à FAPESP pelo apoio financeiro direto e indireto.

RESUMO

Navarro, F.C.P. **A retrotransposição de mRNA como fator de variabilidade genética no genoma humano e de outros primatas**. 2014. 163p. Tese (Doutorado) - Programa de Pós-Graduação em Ciências Biológicas (Bioquímica). Instituto de Química, Universidade de São Paulo, São Paulo.

Duplicação genica é uma das principais forças levando a evolução dos genomas eucarioto. O impacto de duplicações gênicas/genômicas vem sendo investigado a muito tempo em humanos e outros primatas. Um segundo mecanismo de duplicação gênica, a retrotransposição baseada em RNA maduros, vem sendo menos estudada devido ao seu potencial menor de gerar cópias funcionais. No entanto, recentemente, publicações descreveram retrocópias funcionais em humanos, roedores e mosca de fruta.

Nesta tese, para investigar sobre retrocópias causando variabilidade genética no genoma de primatas, nós desenvolvemos a implementamos os métodos para detectar estas inserções. Utilizando nove genomas e transcriptomas publicamente disponíveis (sete primatas e dois roedores) nós confirmamos um número similar, porém, com origem independente, de retrocópias em primatas e roedores. Nós também encontramos um enriquecimento de retrocópias no genoma de Platyrrhini, possivelmente explicado pela expansão de L1PA7 e L1P3 nestes genomas. Posteriormente, nós analisamos a ortologia de retrocópias no genoma de primatas e encontramos 127 eventos específicos à linhagem humana. Nós também exploramos dados do projeto *1000 Genomes* para detectar retrocópias polimórficas (retroCNVs germinativos) e encontramos 17 eventos, presentes no genoma referência humano, mas ausentes em mais de um indivíduo. Similarmente, nós investigamos novas retroduplicações de mRNAs no genoma humano, detectando 21 eventos ausentes do genoma referência. Finalmente, investigamos a existência de retroCNVs somáticos e descrevemos sete possíveis retrocópias somáticas. Apesar de sua possível insignificância, nós encontramos que algumas retrocópias compartilhadas entre todos os primatas, espécie específicas, e polimórficas podem ser expressas *per* se ou como transcritos quiméricos com genes hospedeiros. Sobretudo, nós encontramos que retrocópias são um fator importante da variabilidade genética inter-espécie, intra-espécie e intra-indivíduo e podem estar influenciando a evolução de mamíferos ao criar reservatórios de duplicações potencialmente funcionais.

Palavras-chave: retrotransposição de mRNAs, retrocópia, polimorfismos humanos, evolução de primatas e variação somática.

ABSTRACT

Navarro, F.C.P. **The retrotransposition of mRNAs as a factor of genetic variability in the human and other primates genomes.** 2014. 163p. PhD Thesis - Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Gene duplication is a major driving force of evolution in eukaryotic genome. The impact of gene/genomic duplication has long been investigated in human and other primates. A second mechanism of gene duplication, retrotransposition, which is based on mature RNA, has been traditionally less studied due to their lower potential to generate functional copies. Recently, however, publications described functional retrocopies in humans, murines and drosophila. Here, to gain insights of the genetic variability arising from retrocopies on primate genomes, we developed and implemented the methods to detect these insertions. Using nine publicly available reference genomes and transcriptomes (seven primates and two rodents) we described a similar number independently arisen retrocopies in primates and rodents. We also found an enrichment of retrocopies in Platyrhinni genomes, putatively explained by the expansion of L1PA7 and L1P3 in these genomes. Next, we evaluated the orthology of retrocopies in primate genomes and found 127 events specific to human lineage. We also explored 1000 Genomes Project data to detect polymorphic events (germinative retroCNVs) on human populations and found 17 events, present on the reference genome, absent in more than one individual. Conversely, we also investigated new insertions of mRNA retroduplications in the human genome, detecting 21 events absent to the human reference genome. Finally, we evaluated the existence of somatic retroCNVs and described seven putative somatic retrocopies. Despite their putative insignificance, we found that some of these shared, specie-specific and polymorphic events may be expressed *per se* and as chimeric transcripts within host genes. Taken together, we found that retrocopies are a great factor of genetic variation interspecie, intraspecie e intraindividual and may be affecting mammal evolution by creating reservoirs of potentially functional duplications.

Keywords: mRNAs retrotransposition, retrocopy, human polymorphism, primate evolution and somatic variation.

Lista de Ilustrações e Tabelas

SUMÁRIO

Capítulo 1.
Introdução

“O universo (que outros chamam de Biblioteca) é composto de um número indefinido, e talvez infinito, de galerias hexagonais, com poços de ventilação no meio, cercados por balaustras baixíssimas”

Jorge Luis Borges - Ficções

1.1. Introdução geral

A variação de características entre espécies, populações, indivíduos e patologias são resultados da interação de diversos fatores, entre eles, as variações no material genético. Apesar do notável avanço científico em áreas como bioquímica, biologia molecular e computação, a extensão, redundância e complexidade do genoma humano dificultam a tradução da variação genética em variação fenotípica. A complexidade é tamanha, que conceitos centrais da biologia molecular e bioquímica, como a definição de função ou mesmo a definição de gene ainda são questões sem respostas consenso na comunidade científica (Gerstein et al., 2007 e Kellis et al., 2014). A disponibilização da sequência do genoma humano e o seu estudo, além do óbvio impacto no entendimento da biologia básica, influenciam e permeiam questões filosóficas como, por exemplo, o que nos define como seres humanos e qual a influência da variação genética sobre a natureza humana. Desta forma, o que nos difere está no âmago da discussão do que nos define como seres humanos.

O sequenciamento de genomas na década de 1990 e 2000, nos permitiu, pela primeira vez, observar a real extensão das variações em genomas de eucariotos. Enquanto, para nossa espécie e também para outros organismos, antes do sequenciamento e disponibilização dos genomas referência, as pesquisas eram focadas em pequenas regiões (variações em sítios de restrição ou microsatélites), o advento do sequenciamento de genomas completos permitiu que as análises estendessem-se a todas regiões codificadoras de proteínas, não codificadoras, reguladoras, enfim, por todo o genoma. Apesar de diversas evidências concretas sobre a atividade transcritos e DNA não codificadores apresentarem papéis

fundamentais no funcionamento celular (Esteller, 2011 e Mercer et al., 2009), variações de um único nucleotídeo em regiões codificadoras ainda são consideradas os principais atores da variabilidade fenotípica (1000 Genomes Project Consortium, 2010 e 1000 Genomes Project Consortium et al., 2012). Um exemplo prático é que a variação genética entre dois indivíduos humanos é estimada em 0.1%, a qual representa simplesmente o número de SNPs encontrados entre dois indivíduos. Além de transições e transversões, variações epigenéticas também assumiram um papel importante na última década. Outro tipo bastante estudado nos últimos anos foram as variações estruturais, as quais envolvem ganho ou perda de material genético e podem ser classificadas em deleções, inserções, inversões, duplicações e rearranjos intercromossomais (Sharp A. et al., 2006). Dentro desta classe de variação estão as variações de número de cópia, que podem envolver quaisquer regiões do genoma, entre elas regiões de genes codificadores de proteínas ou elementos repetitivos. Com o desenvolvimento de novas tecnologias e com o barateamento do sequenciamento de genomas completos, além de descrever uma vasta quantidade de variações pontuais e patologias genéticas (Mardis, 2011), verificamos que as variações estruturais são frequentes no genoma humano e podem estar associadas à diversas patologias e variações fenotípicas (Frazer et al., 2009). A extensão da variabilidade genética entre humanos teve grande avanço quando foram publicados os primeiros genomas completos de poucos indivíduos (Levy et al., 2007 e Wheeler et al., 2008). Muito mais recentemente, o aumento da acessibilidade e massificação dos métodos de sequenciamento de segunda geração permitiram o aumento da escala de genomas sequenciados e publicamente disponíveis, culminando em projetos de sequenciamento completo do genoma de mais de 2.500 indivíduos saudáveis (1000 Genomes Project Consortium et al., 2012)

ou 10.000 indivíduos britânicos (www.uk10k.org) que visam catalogar as variantes mais comuns entre indivíduos humanos. Neste ritmo, o notável avanço da compreensão das variações genéticas e genômica podem contribuir para o desenvolvimento de diversas áreas do conhecimento, tal como a medicina, biologia, computação e tem o potencial de transformar o nosso entendimento sobre a natureza e evolução humana e eclodir em uma revolução cultural nas próximos décadas.

1.2. Elementos transponíveis

Elementos transponíveis são sequências de DNA presentes na maioria dos genomas de eucariotos e capazes de moverem-se, ou copiarem-se, em um genoma hospedeiro como um parasita intracelular (Lynch, 2007). Elementos transponíveis são classificados com base em seus mecanismos de locomoção ou duplicação no genoma hospedeiro (Ostertag; Kazazian, H H, 2001a). Os transposons de DNA são sequências que codificam enzimas denominadas transposases (Craig, 1980). Resumidamente, a enzima transposase reconhece sequências sinalizadoras no DNA nuclear, que correspondem as extremidades dos transposons de DNA, e promove a excisão e re-inserção desta molécula em uma posição aleatória do genoma hospedeiro. Analogamente, este mecanismo é chamado de “recorta e cola” (Beck et al., 2011). Notavelmente, este mecanismo não gera duplicações dos elementos movimentados, portanto, são mais sujeitos a inativação por mutações nas regiões que codificam a transposase ou nos sinais de reconhecimento. Desta maneira, a relativa facilidade com que estes eventos são desativados faz com que sua representatividade no genoma humano seja relativamente pequena, correspondendo a, aproximadamente, 3% do genoma humano (Lander et al., 2001)

e do genoma de outros primatas. Retrotransposons, a segunda classe de elementos repetitivos, por definição, dependem da atuação de uma enzima com atividade de transcriptase reversa para realizar a sua movimentação no genoma hospedeiro (Ostertag; Kazazian, H H, 2001a). Resumidamente, o mecanismo de retroposição depende de uma molécula intermediária de RNA para promover a movimentação da sequência de DNA. Analogamente, estes mecanismos são conhecidos como “copia e cola”, de sorte que, a cada evento de retroposição a sequência movimentada é uma duplicação, ao menos parcial, de uma sequência parental.

Retrotransposons são classificados em autônomos, quando codificam a enzima para realizar sua retrotransposição, ou não autônomos, quando dependem de uma transcriptase reversa alheia para promover a sua retroposição. Em primatas, os elementos repetitivos não autônomos são principalmente representados pelos SINEs (*Small Insterspaced Nuclear Elements*), com destaque especial para o elemento Alu (Dewannieux, M; Heidmann, 2005). Estes elementos são sequências curtas de DNA, com cerca de 300 pares de bases não codificantes, compostos basicamente por um promotor interno de RNA polimerase 3 e sequências derivadas do transcrito do gene 7SL, que faz parte da maquinaria ribossomal (Ullu; Tschudi, 1984) (Figura 1). Essa combinação gerou uma sequência que, quando transcrita, apresenta grande afinidade pela maquinaria de transcriptase reversa e, em cerca de 65 milhões de anos (Batzer; Deininger, 2002), foi responsável pela colonização de aproximadamente 10% do genoma humano (Lander et al., 2001) e de outros primatas. Por se tratarem de elementos móveis não autônomos, a amplificação de Alus está diretamente correlacionada com a atividade dos retrotransposos autônomos no genoma hospedeiro (Zhang, Z. et al., 2003).

Os retrotransposons autônomos, que codificam as enzimas necessárias para a transcriptase reversa de seus transcritos, são classificados em duas categorias: com e sem repetições longas (LTR - do Inglês *Long Terminal Tepeats*) flanqueando o elemento transponível. Os retrotransposons LTR são assim chamados, pois, flanqueando a sequência de DNA que codifica as proteínas *gag*, *pol* e *env* (Figura 1), existem sequências não codificantes compostas pelas sequências U5-R na região a montante e U3-R a jusante. Estas sequências são utilizadas como alvos de t-RNAs, que atuam como *primers* durante o processo de transcriptase reversa no citoplasma. O resultado deste complexo processo é a duplicação das extremidades da sequência, gerando uma fita dupla de DNA contendo U3-R-U5-[RetrotransposonLTR]-U3-R-U5 (Figura 1) (Krebs et al., 2009). A sequência U3-R-U5 também é conhecida como LTR e apresenta diversas funções, entre elas, apresenta capacidade promotora para a RNA polimerase II. Os retrotransposons contendo LTRs, representados principalmente pelos retrovírus endógenos (ERVs), assemelham-se à infecções retrovirais ancestrais em células germinativas do hospedeiro (Havecker et al., 2004). No entanto, diferente dos retrovírus exógenos, estes elementos apresentam o gene responsável pela codificação do envoltório viral *(env)* comprometido (Magiorkinis et al., 2012). As inserções retrovirais e suas subsequentes amplificações correspondem a cerca de 8% do genoma humano (Lander et al., 2001). Estima-se que esta porcentagem também seja similar nos genomas dos outros primatas.

A segunda categoria de retrotransposons autônomos não apresentam repetições longas flanqueando a sua sequência e são chamados de retroposons ou retrotransposons não-LTR. Estes elementos são representados principalmente por

LINEs (*Long Insterspaced Nuclear Elements*) e compõem, aproximadamente, 20% do genoma humano (Lander et al., 2001) e de outros primatas.

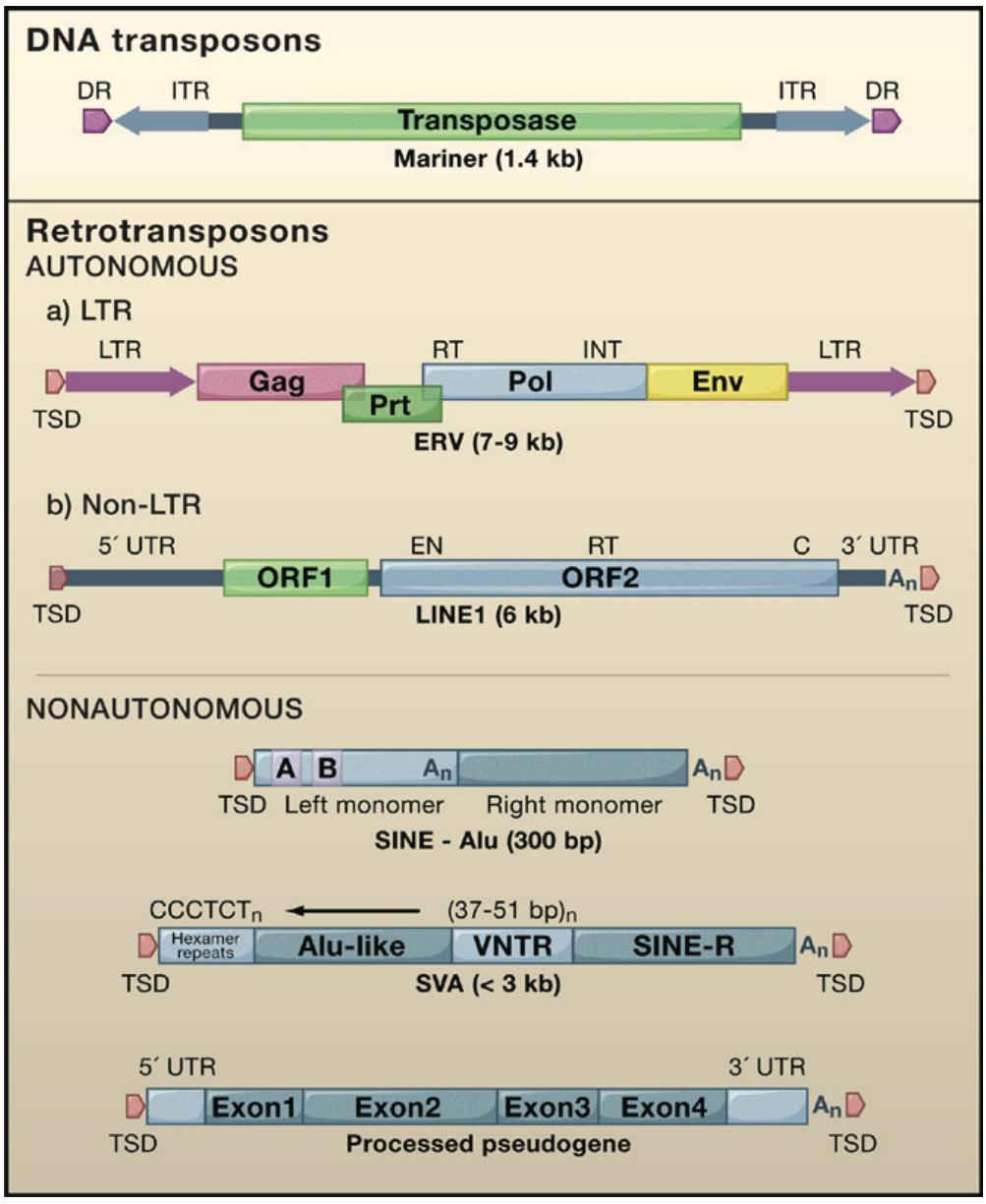


Figura 1. Classes de elementos repetitivos. Adaptado de (Goodier; Kazazian, Haig H, 2008)

1.3. Elementos LINE1

No genoma humano, retrotransposons não-LTR são principalmente representados por elementos LINE1 (L1). Estes elementos, quando íntegros, são

compostos por quatro regiões: i) Uma região não traduzidas a 5’ (5‘UTR), contendo um promotor interno; ii) Região não traduzida 3’ (3‘UTR), com sinal de poli(A); iii) Frequentemente apresentam um poli(A) em sua extremidade 3’; iv) Uma região codificadora policistrônica composta por dois quadros de leitura abertos (ORFs) (Figura 1) (Ostertag; Kazazian, H H, 2001a). O primeiro quadro de leitura (ORF1) codifica uma pequena enzima de 40kDa (Martin, 2006) que apresenta três domínios proteicos (Martin, 2010). Um domínio *coil-coil,* com pouca conservação entre as subfamílias de LINE1, um domínio de reconhecimento de RNA e um domínio c-terminal. A combinação destes domínios não é semelhante à nenhuma outra proteína descrita em genomas de eucariotos e apresenta atividade de ligação a DNA ou RNA, assim como atividade de chaperona (Martin, 2006). Apesar de ter suas funções descritas recentemente, o papel da enzima ORF1p na retroposição ainda é obscuro, porém, essencial para retroposição de LINEs (Martin et al., 2005) e dispensável para a retroposição de SINEs (Dewannieux, Marie et al., 2003). A segunda *ORF* codifica a enzima ORF2p com aproximadamente 150kDa, a qual apresenta dois domínios fundamentais para a retroposição. O primeiro domínio, com atividade de enzima de restrição AP (Feng et al., 1996), é responsável por criar quebras em fitas duplas de moléculas de DNA com sequência consenso fraca AAl TTTT. O segundo domínio, também fundamental para a retroposição, apresenta similaridade com o domínio de transcriptase reversa dos retrotransposons LTR, apesar de serem funcionalmente distintos (Xiong; Eickbush, 1990). Enquanto a transcriptase reversa de retrotransposons LTR ou retrovírus atua no citoplasma celular, utiliza tRNAs como *primer* e exige vários passos intermediários durante o complexo processo de síntese de DNA (Whitcomb; Hughes, 1992), a transcriptase reversa de elementos L1, atua no núcleo, utiliza DNA genômico como *primer* e

realiza a transcriptase reversa em um processo relativamente simples chamado transcriptase reversa com *primer* no alvo ou *target primed reverse transcription* (TPRT) (Cost et al., 2002).

A processividade da transcriptase reversa do L1, apesar de ser relativamente grande quando comparada a outras enzimas similares encontradas naturalmente, não é suficiente para gerar cópias completas de L1 (Piskareva; Schmatchenko, 2006). Segundo experimentos *in vitro,* a processividade desta enzima está próxima de 620 nucleotídeos (Piskareva; Schmatchenko, 2006), aproximadamente 10% do tamanho total de um L1 completo. Experimentos em linhagens celulares, que avaliam a correlação entre o número elementos retropostos e o tamanho do fragmento inserido, demonstraram que apenas 45% dos eventos de retroposição apresentam tamanho superior a três mil pares de bases (Farley et al., 2004). De fato, a maioria dos eventos de retroposição de elementos L1 são truncados na porção 5' (Lander et al., 2001) o que, invariavelmente, gera inativação da maioria das novas cópias.

1.4. Retroposição

A retroposição de elementos L1 tem como primeiro passo fundamental a sua transcrição. Em um L1 completo, os primeiros 670 pares de bases da extremidade 5' não traduzidos (5'UTR) apresentam atividade promotora. Esta região contem um promotor bidirecional interno (Speek, 2001), capaz de ligar-se a diversos fatores de transcrição, em especial ao codificado pelo gene YY1 (Becker et al., 1993). A transcrição se dá pela RNA polimerase II e é finalizada por um sinal poli(A) na região 3' não traduzida (3'UTR). A sequência de RNA polimerizada segue o fluxo normal dos transcritos codificadores de proteína, portanto, o cap 7-metilguanosina é inserido

no início do transcrito (Figura 3) e o sinal de poli(A) dispara enzimas poli-A-polimerases que sintetizam uma cauda de múltiplas adeninas no final dos transcritos L1 (Ostertag; Kazazian, H H, 2001a).

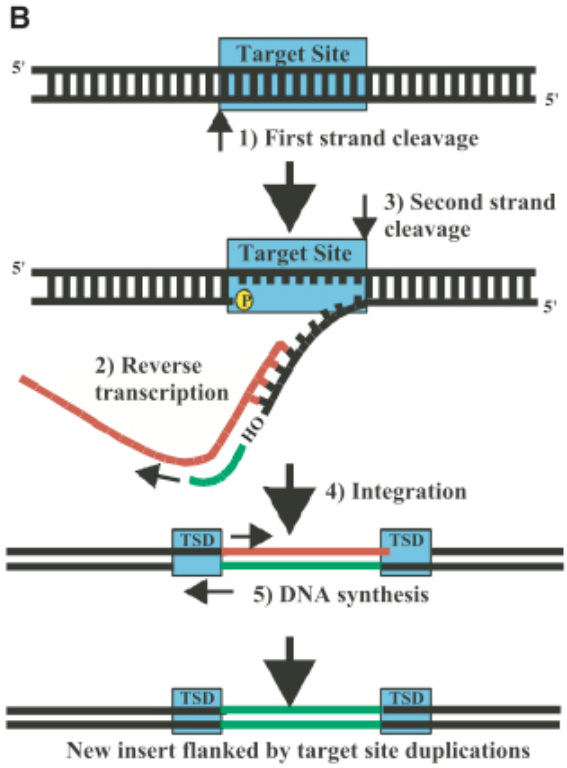


Figura 2. Transcriptase reversa com *primer* no alvo. Adaptado de (Kazazian, Haig H, 2004)

O transcrito maduro de elementos L1 é bicistrônico, portanto, é atípico quando comparado com transcritos maduros de genes codificadores de proteínas em eucariotos (Krebs et al., 2009). Entretanto, de uma forma não muito clara, ambas enzimas necessárias para a retroposição de LINEs, ORF1p e ORF2p, são traduzidas no citoplasma (Ostertag; Kazazian, H H, 2001a). Curiosamente, o RNA usado como molde para tradução, frequentemente se liga às enzimas que codificou, formando um complexo de ORF1p, ORF2p e transcrito-L1 em um processo

conhecido como “preferência em *cis”* (Wei et al., 2001). Este complexo de centenas de kDa é transportado para o núcleo da célula e, de maneira ainda não totalmente esclarecida (Ostertag; Kazazian, H H, 2001a), inicia-se o processo de incorporação do transcrito L1 no ‘genoma hospedeiro.

O domínio de endonuclease na ORF2p é responsável por reconhecer o padrão AAlTTT, de forma não específica (Feng et al., 1996), e catalisar a quebra de uma das fitas de DNA permitindo o anelamento do poli(A) na porção 3’ do transcrito com um poli(T) curto no ponto de inserção. Este processo é conhecido como transcriptase reversa com *primer* no alvo ou TPRT (*target primed reverse transcription*) (Luan et al., 1993) (Figura 2). O *primer* na região alvo permite o início da atividade de transcriptase reversa pela enzima ORF2p. Ao final da síntese da primeira fita de DNA, a ORF2p catalisa a segunda quebra de fita no ponto de inserção (Luan et al., 1993). Vias de reparo de DNA são ativadas pela presença de fita simples de DNA e, durante a síntese da segunda fita de DNA, oito a doze pares de bases de duplicação direta são geradas flanqueando o evento. Ao final da retroposição, a sequência de DNA parental que originou o transcrito sofre a duplicação em um local praticamente aleatório do genoma.

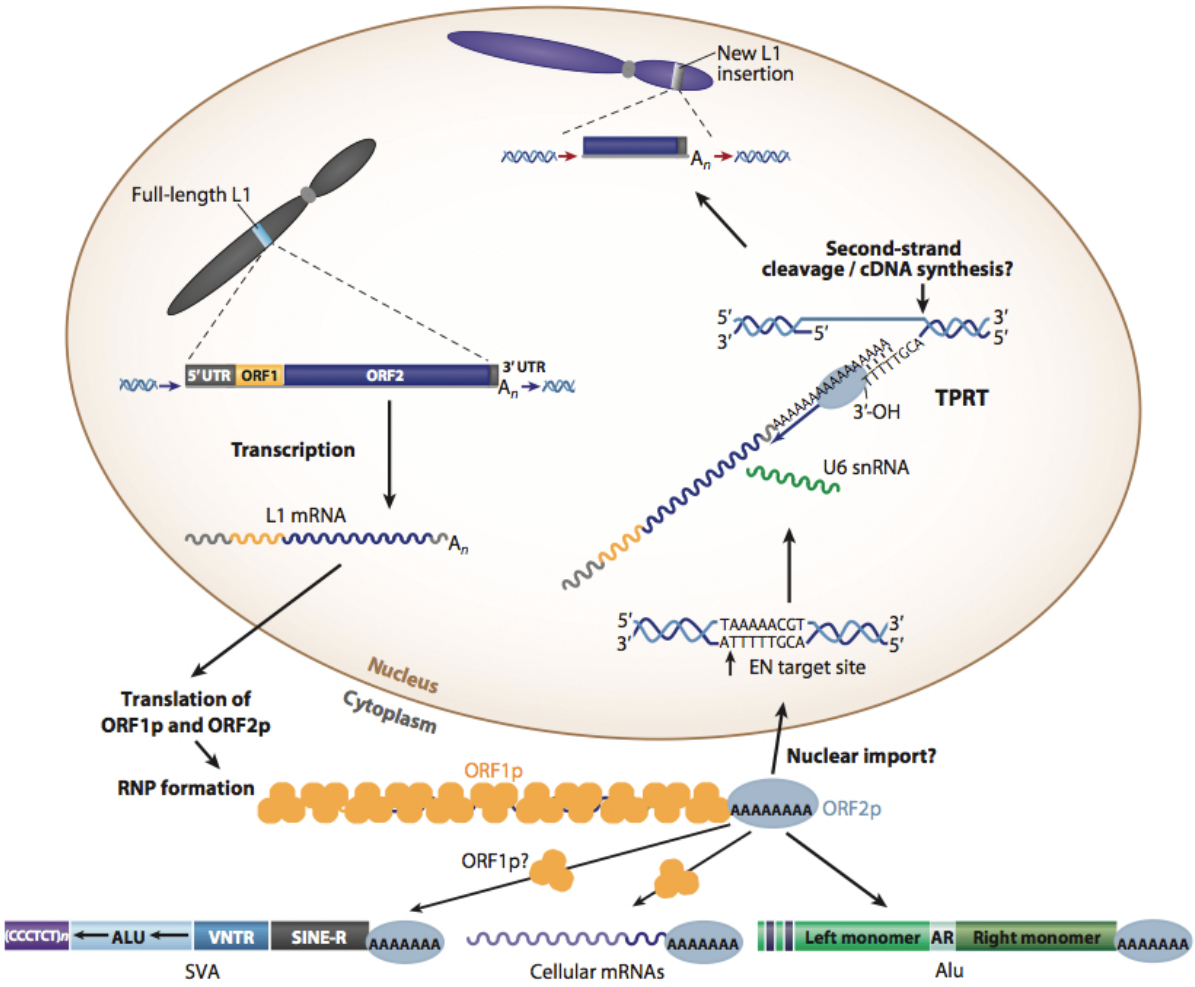


Figura 3. Etapas para a retroposição de elementos L1. Adaptado de (Beck et al., 2011).

1.5. Regulação dos eventos de retroposição

A retroposição de elementos transponíveis, sejam eles autônomos ou não autônomos, é considerada uma das maiores fontes de variabilidade genéticas em mamíferos (Kazazian, H H; Moran, J V, 1998). O resultado desta variabilidade genética é principalmente observado por modificações na arquitetura genômica, sejam elas, inserções, recombinações, deleções, modificações na expressão de genes ou alterações no perfil de *splicing*. Por chance, os efeitos da transposição são predominantemente neutros ou negativos ao hospedeiro (Lynch, 2007), logo, a

restrição da retroposição é fundamental para evitar níveis mutagênicos elevados.

Hipoteticamente, a restrição da retroposição pode acontecer em, pelo menos, quatro níveis: i) a repressão da expressão de elementos L1; ii) mecanismos de regulação pós-transcricionais; iii) regulação da tradução e, por fim; iv) a regulação da retroposição. A repressão da expressão de elementos L1 se dá principalmente por mecanismos de metilação de DNA (Walsh et al., 1998). Apesar das marcações de histonas e metilação de DNA serem perdidas durante a meiose (Smallwood; Kelsey, 2012), a (re)metilação de elementos L1 se dá de forma ativa, por intermédio de uma molécula de RNA (piRNA) e proteínas PIWI, MILI e MIWI2 (Aravin et al., 2007) que são quase que exclusivamente expressas em células germinativas. Mecanisticamente, estas proteínas associam-se a pequenos RNAs de 25 a 27 nucleotídeos formando um complexo ribonucleoprotéico que guiam as proteínas PIWi, MILI e MIWI2 que promovem a metilação de DNA (Lau et al., 2006).

Quando um elemento L1 consegue driblar a repressão em nível de DNA, existem mecanismos auxiliares que degradam seus RNAs mensageiros. Os mecanismos de regulação pós-transcricionais parecem resumir-se a RNAs não codificadores pequenos como miRNA, siRNA e piRNA (Lau et al., 2009). Recentemente, foi demonstrado que a mesma maquinaria responsável pela metilação do DNA de elementos repetitivos (PIWI e MIWI2) tem atividade de RNAse e a interrupção do domínio catalítico relacionado com RNAse na proteína MIWI provoca um aumento significativo de transcritos L1 na célula (Reuter et al., 2011). Este mecanismo deve ser especialmente importante quando células germinativas tem suas marcas de metilação removidas durante a meiose.

De maneira não muito clara, as proteínas APOBEC3, que inicialmente foram descritas como antiretrovirais, também estão relacionadas com a repressão de retroposição de elementos L1 (Muckenfuss et al., 2006). A família de genes APOBEC3 surgiu a partir de amplificações genômicas específicas de primatas (Muckenfuss et al., 2006). Os diferentes membros da família APOBEC3 reprimem a retroposição de elementos L1 com eficiência variável. Por exemplo, enquanto a super expressão de APOBEC3A diminui a integração de elementos L1 em 85% (APOBEC3A), APOBEC3F e APOBEC3G diminuem a integração em apenas 10% (Bogerd et al., 2006). Apesar destas atuarem como deaminases, promovendo a modificação da sequência de mRNAs retrovirais (Chiu; Greene, 2008), sabe-se que a restrição da retroposição não está diretamente relacionada com esta função, mas sim, com a interação entre as proteínas APOBEC3 e a proteína ORF1p de elementos L1 (Horn et al., 2013).

Finalmente, a regulação da integração de elementos repetitivos também pode ser realizada por proteínas endógenas como o dímero ERCC1 e XPF (Gasior et al., 2008). Gasior e colaboradores demostraram que a inativação destas enzimas aumenta significativamente a retroposição de elementos L1. Este dímero, que atua na via de reparo de DNA, tem função de reconhecimento e degradação de extremidades de fitas de DNA não pareadas (Houtsmuller et al., 1999). Portanto, ERCC1 e XPF atuam degradando as fitas simples de DNA no início da atividade da transcriptase reversa de elementos L1 (TPRT), quando há formação de fitas simples (Figura 2), impedindo os últimos passos da retroposição (Gasior et al., 2008).

1.6. Retroposição somática

A retroposição de elementos repetitivos autônomos e não autônomos em células germinativas é responsável por cerca de 45% dos nucleotídeos que compõem o genoma humano (Lander et al., 2001 e Venter et al., 2001) e de outros primatas. Em contraste, a retroposição de elementos repetitivos em células somáticas esta confinada ao indivíduo e, portanto, imunes a pressões seletivas. Apesar de, teoricamente, ser possível haver retroposição de elementos L1 em quaisquer células somáticas (Kubo et al., 2006), os primeiros eventos foram descritos em tecidos tumorais (Liu, J. et al., 1997 e Miki et al., 1992 e Morse et al., 1988). Posteriormente, foram identificadas retroposições somáticas em tecido neural sadio de camundongos utilizando reações de PCR quantitativas, demonstrando um aumento significativo no número de cópias de elementos L1 *in vivo* (Muotri et al., 2005). De forma similar, utilizando linhagens celulares humanas, o mesmo grupo demonstrou que linhagens celulares cerebrais de indivíduos saudáveis apresentavam mais retroposições somáticas quando comparado a linhagens celulares do fígado e do coração (Coufal et al., 2009). Posteriormente, questionando não só a variação no número de cópias, mas também o ponto de inserção de elementos L1 somática, Baillie e colaboradores utilizaram métodos de sequenciamento em larga escala para detectar a inserção de 7.743 inserções somáticas de elementos L1 no hipocampo de três pacientes (Baillie et al., 2011). A busca foi estendida para elementos não autônomos, onde foram encontrados 13.692 e 1.350 retroposições de Alus e SVAs, respectivamente. A atividade de elementos L1 no cérebro humano foi confirmada por um quarto trabalho que, utilizando sequenciamento de uma única célula (300 neurônios de três indivíduos saudáveis), estimou a existência de 0.6 inserções somáticas únicas por neurônios (Evrony et al., 2012).

Apesar da retroposição somática ter sido inicialmente descrita em tumores (Liu, J. et al., 1997 e Miki et al., 1992 e Morse et al., 1988), somente recentemente houve uma retomada do assunto para identificar e quantificar as retroposições somáticas envolvendo genomas tumorais. Três trabalhos foram pioneiros ao reportarem que a frequência da retroposição somática de retroelementos é maior do que se imaginava em tumores humanos e podem contribuir para a formação e progressão tumoral (Iskow et al., 2010 e Lee et al., 2012 e Solyom et al., 2012). Utilizando vinte amostras pareadas de tecido tumoral e normal de pulmão, dez amostras de tecido neural tumoral e com dez amostras leucócitos como controle, e sequenciadores de primeira e segunda geração Iskow e colaboradores identificaram 650 e 403 inserções distintas de elementos L1 e Alus, respectivamente (Iskow et al., 2010). Baseado no sequenciamento completo do genoma de 43 genomas tumorais de 5 tipos diferentes de câncer (colorretal, próstata, ovário, mieloma múltiplo e glioblastoma) foram identificados 194 eventos de retroposição somática (L1, Alu e ERVs) (Lee et al., 2012). Tumores colorretais apresentaram o maior número de retroposição entre os tumores analisados e, assim como os primeiros trabalhos identificando a retroposição de elementos L1, diversas inserções foram encontradas em regiões intragênicas ou próximas de oncogenes. Especula-se que o impacto da retroposição somática de elementos L1 está envolvido na modificação da arquitetura genômica dos tumores facilitando a recombinação, a modificação no perfil de *splicing* e/ou alterações nos níveis de expressão de genes próximos aos pontos de inserção (Lee et al., 2012).

Resultados similares foram obtidos ao analisar a retroposição de elementos L1 por sequenciamento de segunda geração de 16 genomas de amostras pareadas de tumores colorretais e tecido normal. Cento e sete inserções somáticas foram

identificadas, das quais, 35 tiveram o ponto de inserção identificados (Solyom et al., 2012). A maioria das retroposições são relativamente pequenas, truncados na região 5' e, novamente, várias inserções ocorrem em genes envolvidos na tumorigênese, como, por exemplo, CDH11 e PCM1 (Solyom et al., 2012). Muito mais recentemente, 200 pares de tecidos normais e tumorais, de 11 tecidos, foram sequenciados e analisados quanto a retroposição somática utilizando técnicas de sequenciamento de segunda geração (Helman et al., 2014). Neste trabalho, os autores analisaram 767 sequenciamentos de exoma de tecidos tumorais e encontraram, no total, 810 novas inserções de elementos L1.

1.7. Retroposição germinativa

A retroposição somática tem um impacto limitado sobre a espécie, pois está confinada ao tecido do indivíduo em que a retroposição aconteceu. Em contrapartida, quando há retroposição em células germinativas, existe a chance deste evento ser transmitido para gerações futuras. Uma vez que transmitido a um descendente, assim como variações pontuais, o evento de retroposição estará sobre a influência da seleção natural e da deriva genética podendo, com maior frequência, ser perdida, ou, alternativamente, alcançar a fixação em populações, espécies ou linhagens. Assumindo que a inserção de elementos repetitivos é praticamente aleatória, pode-se deduzir que, regiões sintênicas de indivíduos distintos que compartilhem uma mesma inserção (mesmo elemento repetitivo inserido em um mesmo ponto do genoma), são idênticas por descendência.

Sheen e colaboradores, exploram este conceito de inserção dimórfica (presente ou ausente) para explorar a genética de populações e genética forense

antes do genoma referência humano ser publicado e apontaram diversas vantagens para o uso de elementos L1 polimórficos quando comparados ao uso de polimorfismos de um único nucleotídeo. Entre eles: i) a genotipagem de presença ou ausência do elemento repetitivo pode ser feita rapidamente por um PCR; ii) dificilmente deleções acontecerão sobre o mesmo ponto do elemento repetitivo, diminuindo as chances de falsos positivos; iii) é possível analisar a inserção populacionalmente; iv) é possível expandir a análise e avaliar o genótipo de espécies próximas. Entretanto, ressaltou-se a dificuldade criada pelo fato da sequência em questão ser repetitiva e, portanto, poder gerar falsos positivos no processo de identificação e validação (Sheen et al., 2000). Diversos trabalhos seguiram o mesmo estilo de análise em escala reduzida (Badge et al., 2003 e Boissinot et al., 2004 e Myers et al., 2002 e Seleme et al., 2006), até que, em 2006, Wang e colaboradores criaram um banco de dados para armazenar eventos de retroposição polimórficos no genoma humano (Wang et al., 2006). O avanço destas análises permitiram que Witherspoon e colaboradores explorassem a estrutura populacional humana baseado em polimorfismos de presença ou ausência de elementos L1 e Alus (Witherspoon et al., 2006) e a aplicação destas variações em análises forense (Ray et al., 2007). O barateamento do sequenciamento de DNA permitiu que projetos expandissem o número de genomas sequenciados para um novo patamar. A identificação de inserções polimórficas de elementos L1 utilizando dados de sequenciamento de segunda geração de 25 genomas (Ewing; Kazazian, Haig H, 2010) permitiu uma das primeiras estimativas do número de inserções por indivíduos. Ewing e colaboradores, extrapolando o número de inserções encontradas e a frequência alélica destas inserções, estimaram que há uma nova inserção de elementos L1 a cada 140 nascimentos de humanos. O projeto *1000*

Genomes, por exemplo, possibilitou a análise em larga escala em diversas populações. Baseado na análise de apenas 185 indivíduos de 3 populações, identificou-se 7.830 polimorfismos de presença e ausência de elementos repetitivos, destes, 792 eventos correspondiam a inserções de elementos L1 (Stewart et al., 2011).

1.8. Retroposição em trans

Como já descrito anteriormente, o complexo ORF1p/ORF2p tende a se associar e promover a retroposição do transcrito usado como molde para sua tradução (Kulpa; Moran, John V, 2006). Entretanto, em raras situações, o complexo ORF1p/ORF2p não se associa em cis. Nestes raros eventos, há uma troca de template e o complexo associa-se a um transcrito qualquer presente no citoplasma (Mandal et al., 2013). Este evento é conhecido como “troca de molde” (Wei et al., 2001), nestes casos, o novo molde sofre a retroposição em um local aleatório no genoma. Diversos elementos não autônomos, como SINEs (Alus e SVAs), utilizam esta capacidade de troca de molde do complexo de retroposição para colonizar o genoma de primatas (Kazazian, H H; Moran, J V, 1998). Ainda mais raramente, o complexo ORF1p/ORF2p trocam seu molde por transcritos de genes codificadores de proteína. O processo de retroposição destes transcritos é conhecido como retrocópia ou retroduplicação de mRNAs (Kaessmann et al., 2009).

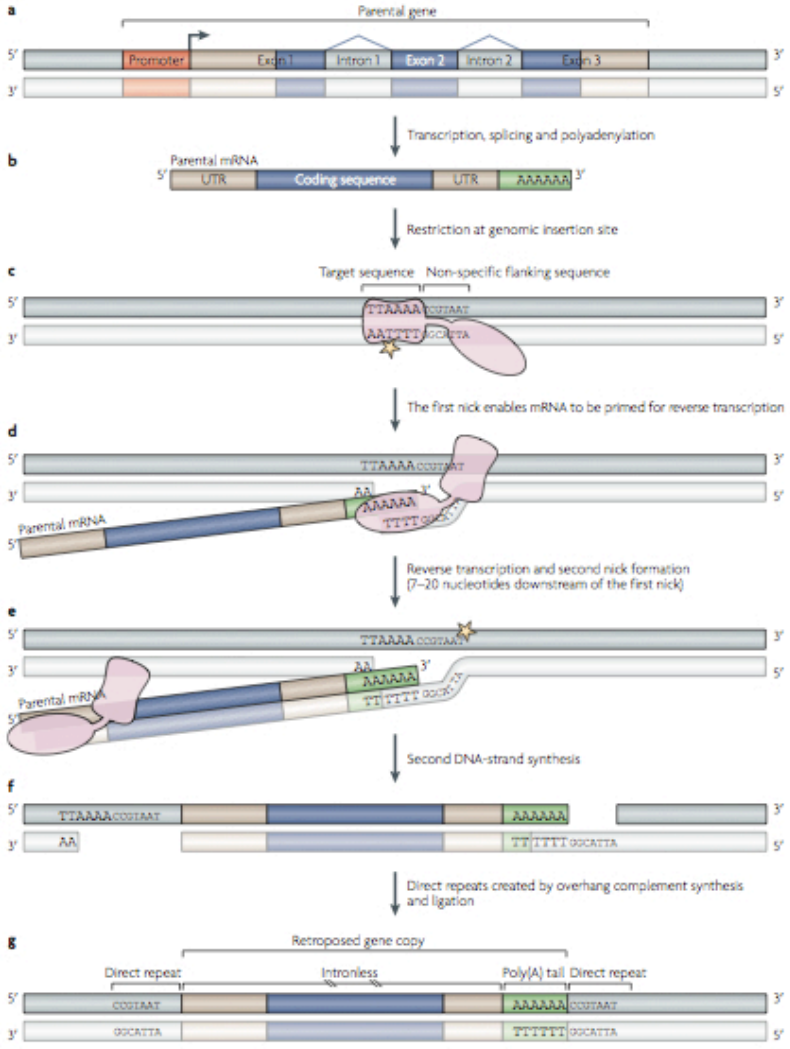


Figura 4. Processo de retroposição de um transcrito de genes codificadores de proteína.

No início dos anos 80, pouco tempo após a descrição dos introns (Berget et al., 1977 e Chow et al., 1977), diversos grupos de pesquisa descreveram famílias gênicas em genomas de eucariotos superiores. Inesperadamente, durante o estudo de genes como insulina (Lomedico et al., 1979) e globina (Vanin et al., 1980), diversos casos de duplicações sem introns foram reportados. Nos anos seguintes, outras duplicações sem introns, ou também chamados de pseudogenes processados, foram descritos (Wilde et al., 1982) e geraram surpresa na

comunidade científica (Vanin, 1985). Prontamente, hipóteses surgiram para explicar as características compartilhadas por estes eventos. A ausência de introns nos pseudogenes processados fez com que os pesquisadores da época hipotetizassem, corretamente, que um intermediário de RNA deveria estar envolvido no processo de duplicação. Nishioka e colaboradores sugeriram que a perda de introns de um gene poderia surgir a partir de um mecanismo de conversão do gene onde haveria a formação de um *heteroduplex* do gene e seu o RNA mensageiro e o DNA (Nishioka et al., 1980), gerando a excisão dos introns. Ueda e colaboradores, investigaram a presença de LTRs flanqueando um pseudogene processado em humanos, e propuseram que o surgimento de pseudogenes processados seria intermediado por transcriptases reversas de retrovírus endógenos (Ueda et al., 1982). Por fim, foi proposto que os pseudogenes processados seriam subprodutos da atuação maquinaria de *splicing* sobre moléculas de DNA (Vanin et al., 1980).

Apesar de ainda não estar totalmente elucidado, a retroposição de mRNAs maduros e consequente geração de pseudogenes processados teve diversos pontos chaves esclarecidos. Devido a suas principais características, como ausência de introns e presença de poli(A) na extremidade 3', foi postulado que o transcrito duplicado teria que ser transcrito, maturado e exportado para o citoplasma. A formação induzida de pseudogenes processados foi observada pela primeira vez, em células tumorais (HeLa), apenas uma década após o surgimento das primeiras hipóteses (Maestre et al., 1995), entretanto, somente cinco anos depois de observação, foi demonstrados que, de fato, retroposons L1 são capazes de gerar pseudogenes processados (Esnault et al., 2000). No mesmo período, também foi demonstrada a preferência da maquinaria de transcriptase reversa pelo mRNA que é molde de sua tradução (retroposição em *cis*), em detrimento da troca de RNA molde

(retroposição em *trans*) (Wei et al., 2001). Finalmente, Mandal e colaboradores demonstraram que transcritos de genes codificadores de proteínas encontram-se ligados à maquinaria de transcriptase reversa no citoplasma de linhagens celulares (Mandal et al., 2013). Portanto, postulam-se os seguintes passos para ciclo de retroposição de um mRNA: i) O gene parental é transcrito e o transcrito é processado, perdendo os introns, recebendo o CAP, poli(A) em sua extremidade 3' e, finalmente, o transcrito segue para o citoplasma. Neste ponto, não se sabe se o transcrito sofre tradução ou é imediatamente sequestrado pela maquinaria de transcriptase reversa - o que essencialmente não faz diferença para o evento de retroduplicação; ii) ao se ligar a maquinaria de transcriptase reversa, o complexo mRNA e L1-RNP voltam ao núcleo; iii) a maquinaria de transcriptase reversa, composta principalmente pela ORF2p dos elementos L1, gera uma quebra de uma das fitas e procede exatamente como se estivesse ligada a transcritos de elementos L1 (Cost et al., 2002). Curiosamente, algumas excessões foram detectadas no genoma humano, por exemplo, alguns transcritos são parcialmente processados (Zhang, Z. D. et al., 2008), abrindo a possibilidade destes serem capturados pela maquinaria da transcriptase reversa antes de serem exportados para o núcleo. Adicionalmente, alguns eventos apresentam um perfil de retroposição diferente do esperado se a transcriptase reversa acontecesse de forma totalmente linear. Alguns pseudogenes processados apresentam uma inversão na região 5' (Kojima; Okada, 2009), esta inversão é, provavelmente, causada por um segundo evento de *primming* que acontece durante a transcriptase reversa e foi nomeado como "*twin primming*" (Ostertag; Kazazian, H H, 2001b) o qual acontece quando há similaridade entre regiões do transcrito retrocopiado e o ponto de inserção. Finalmente, iv) se toda a reação de transcriptase reversa e consequente correção de erros endógena

for bem sucedida, haverá uma duplicação do gene parental criando um novo pseudogene processado (Kaessmann et al., 2009).

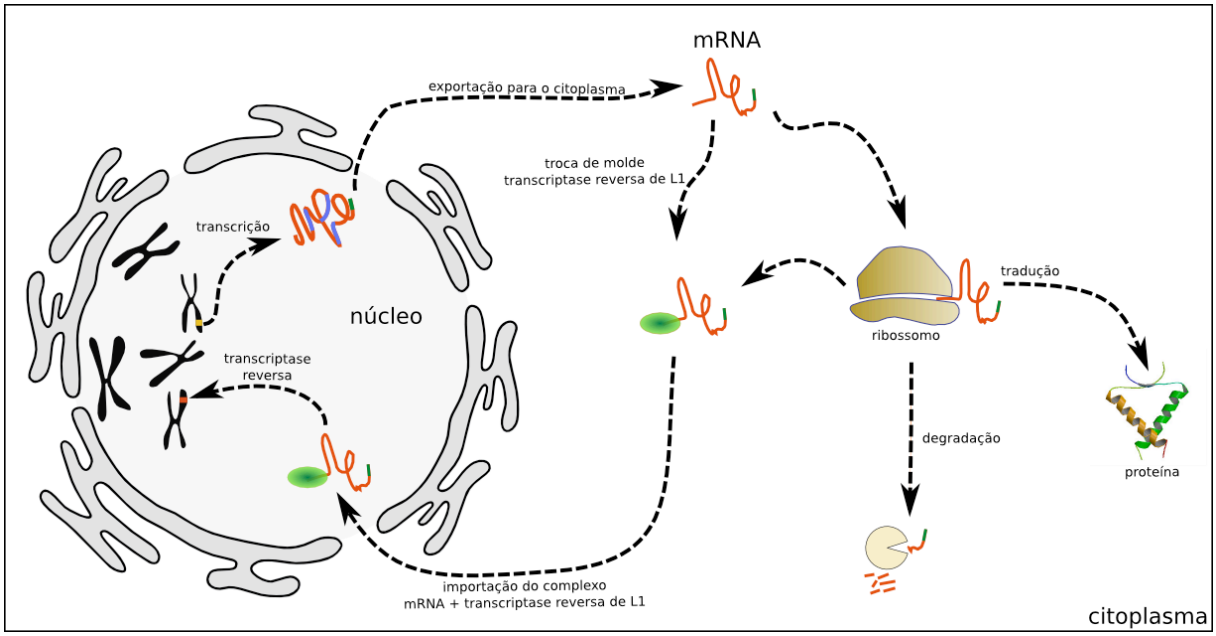


Figura 5. Etapas de retroposição de um transcrito de genes codificadores de proteína.

1.9. Nomenclatura

1.9.1. Retrocópias, pseudogenes processados e retrogenes

Diversos fatores influenciaram a nomenclatura de eventos de retroduplicação de mRNAs. Historicamente, quando os primeiros eventos foram descritos na década de 80, os poucos casos estudados apresentavam acumulo de mutações, perda da região promotora e enriquecimento da porção 3‘ do gene parental (Piskareva; Schmatchenko, 2006). Portanto, assumiu-se que estes eventos fossem “*dead-on-arrival*” e o termo pseudogene processado foi genericamente utilizado para descrever qualquer retroduplicação de mRNA. Quase uma década após o termo ser cunhando, em 1987, o termo “retropseudogene” foi utilizado por Srikantha e

colaboradores como sinônimo de pseudogene processado (Srikantha et al., 1987). Um ano depois, outro grupo utilizou o termo “retrogene”, também como sinônimo de pseudogene processado (Adra et al., 1988). Estes três termos, e variantes como “retrogenes processados” ou “pseudogenes retroprocessados”, foram utilizados indiscriminadamente até que, em 1991, Brousius, apesar de não ter evidência direta, publicou um trabalho de perspectiva (Brosius, 1991), sugerindo que alguns destes eventos poderiam ser expressos, codificar proteínas e, portanto, serem considerados como potencialmente funcionais.

A publicação do genoma referência humano e o desenvolvimento dos sequenciamentos de DNA em larga escala nos anos 2000 (Lander et al., 2001) trouxe alguma consistência na terminologia utilizada para descrever esta classe de eventos. Neste período os trabalhos passam a fazer a distinção mais clara entre retrogenes, como retrocópias codificadoras de proteínas; pseudogenes processados ou retropseudogenes como retrocópias com mutações ou indels que destroem as ORFs dos genes parentais; e retrocópias, um termo mais universal, utilizado para todos os eventos de retroduplicação de mRNAs, independente de sua classificação funcional. Entretanto, contra exemplos de trabalhos de grande impacto utilizando indiscriminadamente o termo pseudogene processado (Cooke et al., 2014) ainda são vistos na literatura. Nesta tese, o *locus* originado pela retroposição de transcritos maduros de genes codificadores de proteína será chamado de retrocópia ou retroduplicação de mRNAs. O termo pseudogene processado será utilizado em *loci* que apresentem mutações que destruam ORFs dos genes parentais. Finalmente, o termo retrogene será utilizado quando a retrocópia for, por si, codificante de proteína ou apresente função não codificadora já evidenciada na literatura.

1.9.2. Genes parentais e hospedeiros

Além da retrocópia em si, uma segunda entidade está sempre envolvida no processo de retroduplicação de mRNAs. Genes parentais (Figura 6A) são os genes que deram origem ao transcrito (Figura 6B) que sofreu retroposição. A definição do gene parental nem sempre é trivial. Algumas retrocópias estão fixadas no genoma humano e de outros primatas há dezenas de milhões de anos e acumularam mutações a ponto de dificultar a identificação dos seus respectivos genes parentais (Zhang, Z. et al., 2004). Finalmente, o ponto de inserção pode envolver uma terceira "entidade gênica". Caso a inserção aconteça dentro de um l*ocus* anotado como gene, seja em região intrônica ou exônica, este gene é chamado de gene hospedeiro (Figura 6D).

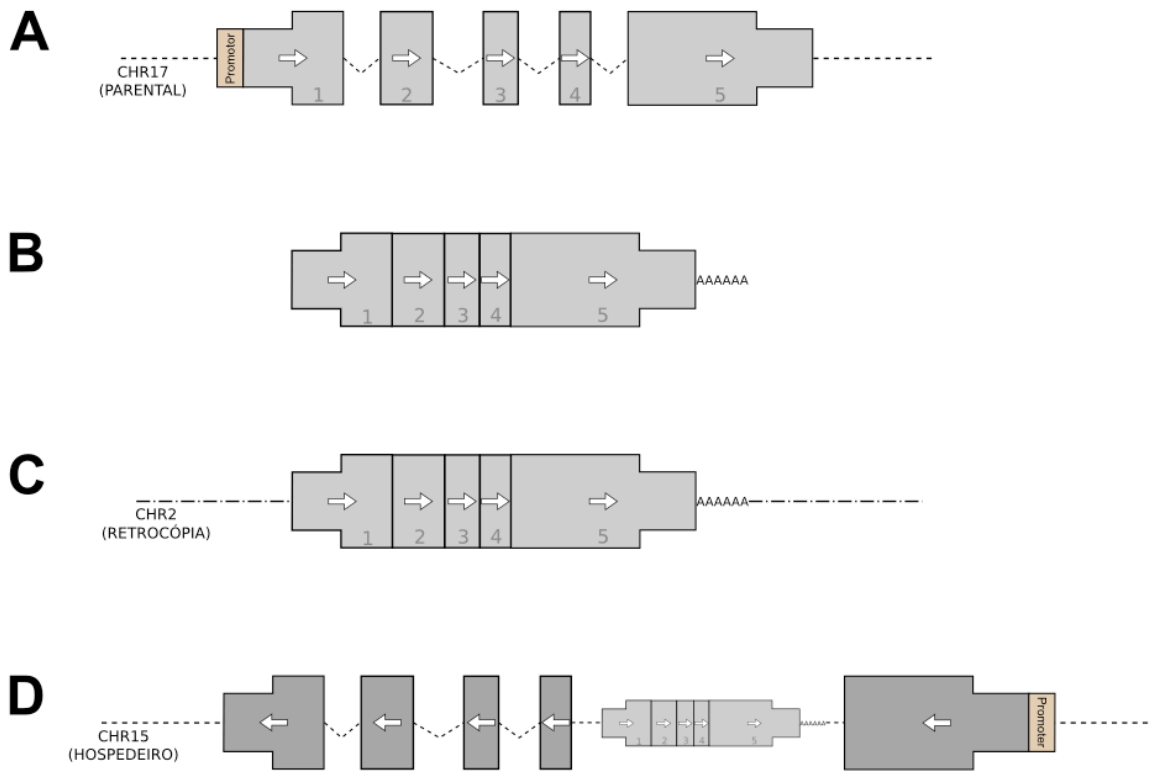


Figura 6. Entidades envolvidas no processo de retroduplicação de mRNAs. Diagramação hipotética de um evento de retroduplicação de mRNAs. **A)**

Representação de um gene parental hipotético no cromossomo 17; **B)** Transcrito maduro do gene parental; **C)** Retrocópia completa no cromossomo 2; **D)** Gene hospedeiro hipotético no cromossomo 15, com a retrocópia no primeiro intron.

1.10. Retrocópias no genoma humano

A análise quantitativa e qualitativa em larga escala de retrocópias e pseudogenes processados surge apenas após a publicação do genoma referência humano. Até então, os resultados quantitativos eram sempre restritos a um número pequeno de retrocópias e ocupavam papel secundário em discussões sobre número de genes. Alternativamente a descrição de eventos estava restrita a literatura de evolução de genes (Li et al., 1981). A ilustração perfeita para este cenário é a publicação do genoma referência por Venter e colaboradores (Venter et al., 2001). Neste trabalho, os autores descrevem superficialmente a existência de 2.909 *loci* anotados como pseudogenes processados, porém, nenhuma tabela ou figura é dedicada ao assunto (Venter et al., 2001). Apesar de ser uma das primeiras análises em larga escala, pouco se discute sobre os métodos para detecção ou implicações destes eventos no genoma humano. A publicação de Lander e colaboradores não cita a detecção de pseudogenes processados, apesar de discutir sobre a variabilidade gerada por elementos repetitivos L1 (Lander et al., 2001).

A primeira publicação específica sobre o assunto surge um ano após a publicação do genoma referência humano. Baseado somente no sequenciamento dos cromossomos 21 (Hattori et al., 2000) e 22 (Dunham et al., 1999), Harrison e colaboradores estimam a existência de 8.700 a 9.400 pseudogenes processados no genoma humano (Harrison et al., 2002). Adicionalmente, os autores descrevem o que viria a ser um dos principais métodos para detecção de pseudogenes

processados. Resumidamente, faz-se o alinhamento de sequências proteicas descritas e preditas no genoma humano, avalia-se o resultado, procurando por hiatos. O principal filtro verifica a ausência de hiatos maiores que 126 nucleotídeos entre exons do gene parental, e anota o *locus* como uma possível retrocópia. Filtros adicionais avaliam a presença de variações na região codificadora duplicada que, consequentemente, interrompam a codificação da proteína original do gene parental. Nestes casos o *locus* é anotados como pseudogene processado.

O incremento da qualidade das sequências do rascunho do genoma referência humano e aumento da capacidade computacional, permitiram que nos anos seguintes diversos trabalhos explorassem a descrição de pseudogenes processados de forma mais consistente. O mesmo grupo, limitando as análises a proteínas ribossomais, estimou a existência de 2.090 pseudogenes processados ribossomais no genoma humano (Zhang, Z. et al., 2002). Neste trabalho são descritas características gerais de pseudogenes processados como, por exemplo, percentagem do gene parental retrocopiado, distribuição da idade estimada das inserções, conteúdo GC, correlação positiva entre tamanho do cromossomo e número de pseudogenes, correlação da divergência de sequência entre pseudogenes processados, Alus e elementos L1. Ohshima e colaboradores também estimaram o número de pseudogenes processados (3.664) no genoma humano e encontraram um número de eventos similar ao inicialmente publicado por Venter e colaboradores (Ohshima et al., 2003), porém, mais importante que o número de eventos detectados, os autores exploraram em maior profundidade as análises qualitativas de pseudogenes processados. A partir de análises do número de substituições nas sequências dos pseudogenes processados, comparados com seus genes parentais, Ohshima estimou que a maioria destes eventos teria um pico de

surgimento há, aproximadamente, 40 milhões de anos e, portanto, a maioria destes eventos coincidia com o pico de atividade de subfamílias de LINEs (L1PA6, L1PA7 e L1PA8) específica de primatas.

No mesmo ano, Zhang e colaboradores, publicaram um dos marcos para a área de pseudogenes, que viria a ser a base de dados fundamental para a criação da ferramenta "*pseudogene.org"* (Zhang, Z. et al., 2003). Este trabalho confirmou as estimativas iniciais baseadas no cromossomo 21 e 22, descrevendo 7.819 pseudogenes processados no genoma humano (Zhang, Z. et al., 2003). Assim como Ohshima, análises qualitativas confirmaram que pseudogenes processados não apresentavam viés de inserção e, portanto, o número de pseudogenes processado apresenta uma correlação direta com número de bases do cromossomo analisado. Observações interessantes como a de que apenas 13% dos pseudogenes processados apresentam identidade superior a 90% (com média de 75%) indicam que, durante a evolução de primatas, a atividade de retroposição de retroelementos e retrocópias sofreu forte redução. Neste trabalho também estima-se a existência de 2.555 possíveis genes parentais e, pela primeira vez, observa-se que os genes parentais mais retrocopiados codificavam proteínas ribossomais. A distinção clara entre região retrotransposta e pseudogene processado surge de forma explicita e é quantificada no trabalho de Torrents e colaboradores. Os autores descrevem 10,511 retrocópias no genoma humano e, destas, 4.844 são pseudogenes processados (Torrents et al., 2003). Similares aos trabalhos de Ohshima e Zhang, Torrents e colaboradores também analisaram o número de mutações sinônimas e não sinônimas buscando por evidências de funcionalização destes eventos.

No ano seguinte, 2004, Zhang revisa seus resultados e publica um manuscrito comparando o número de pseudogenes processados em humanos e camundongos

(Zhang, Z. et al., 2004). Os valores foram atualizados, respectivamente, para 6.054 e 3.227 pseudogenes processados. A diferença no número de pseudogenes processados em camundongos foi inicialmente justificada pela maior frequência de mutações, inserções e deleções no genoma de camundongos quando comparado ao genoma humano, dificultado a detecção de pseudogenes mais antigos. Neste período, Emerson e colaboradores, utilizam o número de retrocópias como indicador da evolução mais rápida do cromossomo X. Partindo de um conjunto restrito de retrocópias no genoma humano (1.859 eventos), foram detectados 105 pares de genes parentais e retrocópias expressas que seriam potencialmente funcionais. Ao investigar os movimentos cromossomais e intercromossomais, verificaram um interessante viés de genes sendo exportados como retrocópias do cromossomo X para autossomos (Emerson et al., 2004). Adicionalmente, descreveram também que a maioria das cópias de genes do cromossomo X apresentavam expressão preferencial no testículo.

Quase um ano depois, Harrison e colaboradores, publicam uma análise em larga escala da transcrição de pseudogenes processados (Harrison et al., 2005). Neste manuscrito, que se baseia em dados de EST (*expressed sequence tags*), foram encontrados 233 pseudogenes processados transcritos e também são confirmados os vieses de exportação e expressão de pseudogenes processados do cromossomo X e no cromossomo X encontrados por Emerson e colaboradores. Este trabalho também faz, pela primeira vez, uma análise de eventos ortólogos entre humanos e camundongos. Apenas 11 dos 233 (5%) pseudogenes processados transcritos, tem um ortólogo correspondente no genoma de camundongos. Por representarem uma fração mínima do total de pseudogenes processados, os autores

discutem a inviabilidade de inferir função destes eventos de retroduplicação em escala evolutiva.

Com o aumento da importância dos pseudogenes processados, começam a surgir a primeiras publicações de bancos de dados de retrocópias ou pseudogenes processados. A primeira ferramenta do gênero, HOPPSIGEN, descreve 5.206 e 3.934 “retroelementos” em humanos e camundongos, respectivamente (Khelifi et al., 2005) e, pela primeira vez, uma ferramenta *web* é desenvolvida para facilitar a consulta de pseudogenes processados. Entretanto, a ferramenta tem o foco na disponiblização de dados brutos sobre retrocópias e não a usabilidade da ferramenta.

Na tentativa de discernir entre retrocópias funcionais (codificantes de proteínas), também chamados de retrogenes, e pseudogenes processados, Marques e colaboradores analisaram o número de mutações sinônimas contra mutações não sinônimas, entre retrocópias e seus genes parentais, para inferir funcionalização a partir de sinais de seleção (Marques et al., 2005). Este trabalho partiu de um conjunto relativamente restrito de eventos, 3.951 retrocópias. Destes eventos o grupo descreve 11 potenciais retrogenes, que são descritos como retrocópias com menos mutações não sinônimas que sinônimas e evidência de expressão. Este manuscrito, associado ao manuscrito de Harrison e colaboradores marcam o início do estudo de retrocópias como possíveis genes codificadores de proteínas. Um ano depois, Shemesh e colaboradores exploraram o conceito de fossilização de transcritos por eventos de retroposição de mRNAs maduros (Shemesh et al., 2006). Assumindo que retrocópias são majoritariamente duplicação de mRNAs maduros, é possível verificar se alguma retrocópia representa um transcrito não encontrado atualmente no transcriptoma humano. Neste mesmo ano,

Vinckenbosch e colaboradores, publicaram mais um manuscrito buscando por funcionalização de retrocópias no genoma humano. Partindo de um número ainda menor de eventos, 3.590 pseudogenes processados e dados de ESTs, são descritas 1.080 retrocópias expressas, sendo que 271 destes eventos são retrocópias intactas (com ORF parental funcional). Além de descreverem o maior número de retrocópias expressas até então, os autores exploram as possíveis formas de funcionalização de retrocópias. Nominalmente, são descritos os seguintes processos de funcionalização: i) Aquisição de promotores de genes e/ou elementos repetitivos; ii) geração de genes quiméricos; iii) e aquisição de novos exons. Neste mesmo trabalho, houve a confirmação do movimento de retrocópias do cromossomo X para autossomos e a primeira evidência de que retrocópias seriam frequentemente expressas em testículos, devido a uma diminuição nas restrições epigenéticas neste tecidos (Vinckenbosch et al., 2006). Portanto, neste período, havia um consenso de que cada vez mais retrocópias seriam anotadas como funcionais.

A primeira análise em larga escala e em múltiplos genomas surgiu em 2007, quando Yu e colaboradores, utilizaram o genoma de oito vertebrados (humano, chimpanzé, cachorro, vaca, rato, camundongo, galinha e baiacu) para detectar possíveis pseudogenes processados e retrogenes. Os valores encontrados foram abaixo da média da literatura até então. O genoma humano e de chimpanzé, por exemplo, apresentaram apenas 2.493 e 1.889 pseudogenes processados respectivamente (Yu et al., 2007).

Até 2007, todos os trabalhos baseavam-se em sequências de proteínas para predizer retrocópias nos genomas estudados. Sakai e colaborares, foram os primeiros a aplicar métodos similares aos desenvolvidos anteriormente, porém baseados na sequência de transcritos para detectar retrocópias (Sakai et al., 2007).

O número de possíveis retrocópias em humanos e camundongos foi, respectivamente, de 7.348 e 6.188, equiparando o número de retrocópias descritas nestes dois organismos. Sem analisar a ortologia das retrocópias em ambas espécies, os autores estimaram o número de retrocópias compartilhadas baseado no número de substituições sinônimas em retrocópias e comparadas com seu respectivo gene parental. Os autores afirmaram que ao menos 80% das retrocópias teriam surgido após a divergência entre humanos e camundongos, isto é, de maneira específica a cada linhagem. Além disso, como mamíferos apresentam um número muito maior de retrocópias que, por exemplo, galinha, concluiu-se que pseudogenes processados poderiam contribuir para a evolução de mamíferos (Sakai et al., 2007). Na mesma linha de Sakai e colaboradores, Baertsch e colaboradores desenvolveram um método baseado em transcritos para detectar possíveis retrocópias no genoma humano. Por meio do BLASTZ (Schwartz et al., 2003), este trabalho descreveu 12.801 retrocópias (sem distinção de entre funcional ou não funcional) de genes codificadores de proteínas com múltiplos exons (Baertsch et al., 2008). Assim como o trabalho de Vinckenbosch e colaboradores, este trabalho explora os possíveis impactos de retrocópias no genoma hospedeiro. São descritos em maior profundidade 15 eventos de retrocópias expressas (de um total de 766 retrocópias expressas) e seus impactos sobre genes hospedeiros ou genes próximos ao evento.

Focando apenas nos genes relacionados com a via glicolítica, Liu e colaboradores, descreveram todos os eventos de pseudogenes processados e não processados em nove organismos, entre eles, humanos, chimpanzés, camundongos e ratos (Liu, Y.-J. et al., 2009). Este trabalho, pela primeira vez, fez uso de regiões sintênicas para avaliar o número de pseudogenes processados ortólogos entre

organismos. Foram descritos 64 pseudogenes compartilhados entre primatas (humanos e chimpanzés) e 135 compartilhados entre roedores (camundongo e ratos), como descrito pelos trabalhos anteriores, a maioria das retrocópias destes organismos surgiram após a divergência entre roedores e primatas e, portanto, apenas quatro destes eventos são compartilhados entre roedores e primatas.

Na mesma linha, Balasubramanian e colaboradores, publicaram um manuscrito comparando pseudogenes processados e não processados de genes ribossomais no genoma de quatro primatas (Balasubramanian et al., 2009). Similarmente, humanos e chimpanzés compartilham 70.36% dos pseudogenes relacionados com genes ribossomais, enquanto apenas 13.86% do genes relacionados com proteínas ribossomais são compartilhados entre roedores (ratos e camundongos, enquanto, apenas 0.6% dos pseudogenes processados são compartilhados entre primatas e roedores.

Neste período, de 2009 a 2012, diversos manuscritos foram publicados procurando formas de selecionar e diferenciar retrocópias não funcionais (pseudogenes processados) e retrocópias funcionais (Khachane; Harrison, 2009). Porém, uma nova tendência passou a existir a partir do trabalho de Khachane e colaboradores. Com o surgimento e estabelecimento do potencial funcional de RNA não codificadores, ficou claro que, devido a alta similaridade entre retrocópias e seus genes parentais, as retrocópias, quando transcritas, poderiam atuar como reguladores diretos (RNAi endógeno) ou indiretos (sequestrando miRNA, por exemplo) auxiliando na regulação de seus genes parentais. Mais recentemente, um marco para retrocópias surgiu com a publicação de Poliseno e colaboradores, explorando um par, retrocópia e gene parental, que corregulam-se pós transcricionalmente por compartilhar sítios alvos de miRNA (Poliseno et al., 2010).

Após este período no final da década de 2000, onde diversos grupos estabeleceram o conjunto de retrocópias em diversos organismos, diferenciaram retrocópias funcionais de não funcionais, houve uma mudança de foco, onde diversos trabalhos passaram a estudar retrocópias específicas, descrevendo suas possíveis funções e seus possíveis impactos fisiológicos e patológicos (Ehsani et al., 2011 e McEntee et al., 2011 e Tay et al., 2011 e Zhang, J. et al., 2012).

Capítulo 2.
Objetivos

“Cada exemplar é único, insubstituível, mas há sempre
várias centenas de milhares de fac-símiles imperfeitos:
de obras que não diferem entre si a não ser por uma
letra ou por uma vírgula”

Jorge Luis Borges - Ficções

2.1. Objetivos gerais

Este doutorado tem como objetivo estudar variações estruturais que contribuam para a variabilidade genética baseando-se em dados originais e públicos de sequenciamento de DNA genômico (gDNA) e de sequências transcritas (cDNA) geradas, principalmente, por sequenciadores de segunda geração. Sob a perspectiva da genômica, iremos investigar a contribuição das retrocópias para a evolução de primatas, seu impacto em diferentes populações e para o desenvolvimento de patologias como o câncer. Como objetivo secundário, iremos desenvolver os métodos computacionais necessários para a detecção destas variações estruturais genômicas em eucariotos e disponibilizar os resultados publicamente.

2.2. Objetivos específicos

Variações genéticas entre indivíduos, espécies, linhagens e patologias podem ser classificadas em um vasto espectro que vão da substituição pontual de nucleotídeos à variação no número e composição de cromossomos. O advento da tecnologia de sequenciamento de DNA em larga escala permitiu comparar genomas em alta resolução descrevendo, por exemplo, variações pontuais presentes na população humana (1000 Genomes Project Consortium, 2010 e International HapMap Consortium, 2003), pequenas inserções e deleções (1000 Genomes Project Consortium et al., 2012) até a caracterização de variações estruturais envolvendo milhões de nucleotídeos. Este projeto tem como objetivo investigar o repertório de retrocópias no genoma de primatas e roedores e descrever a abrangência de um novo tipo de variação estrutural, o polimorfismo de presença e ausência de

retrocópias, ou retroCNVs com origem germinativa ou somática. Os objetivos detalhados deste projeto são:

1. Descrever, catalogar e disponibilizar as retrocópias no genoma de primatas (humanos, chimpanzés, gorilas, orangotangos, rhesus, saguis e macaco esquilo) e roedores (camundongos e ratos).
2. Investigar o perfil das retroposições de transcritos de genes codificadores de proteína e entender o impacto destes *loci* no genoma de primatas e roedores.
3. Comparar a ortologia das retrocópias descritas em humanos com retrocópias em outros primatas, a fim de entender como estas variações comportam-se em escala evolutiva.
4. Investigar o potencial polimórfico germinativo das retrocópias e descrever a variação alélica destes eventos na populacional humana.
5. Investigar a ocorrência de retrocópias somáticas em genomas tumorais e descrever eventos potencialmente relacionados com o desenvolvimento de tumores.

Capítulo 3. **M**ateriais e **M**étodos

“A escrita metódica me distrai da presente condição dos homens.

A certeza de que tudo está escrito nos anula ou faz de nós fantasmas”

Jorge Luis Borges - Ficções

3.1. Dados primários

3.1.1. Genomas referência.

Sequências do genoma referência de sete primatas (*Homo sapiens* - hg19, *Pan troglodytes* - panTro3, *Gorilla gorilla* - gorGor3, *Pongo abelii* - ponAbe2, *Rhesus macaque* rheMac2, *Callithrix jacchus* - calJac3 e *Saimiri boliviensis* - saiBol1.0) e dois roedores (*Mus musculus* - mm9 e *Rattus norvegicus* - rn4) foram obtidos do UCSC Genome Browser (http://hgdownload.cse.ucsc.edu). Sequências e coordenadas de transcritos codificadores de proteína foram obtidos a partir do RefSeq (Pruitt et al., 2013) (humano, camundongo e rato: versão 49; chimpanzé: versão 50; orangotango, saguí, rhesus: versão 51; macaco esquilo: versão 61). Devido a inexistência do transcriptoma de gorila no RefSeq no período em que as análises foram realizadas, coordenadas e sequências de transcritos codificadores de proteína para este organismo foram obtidos a partir do ENSEMBL (http:// www.ensembl.org).

3.1.2. Dados de expressão.

Sequências públicas do transcriptoma de seis tecidos (cérebro, cerebelo, testículo, fígado, rim e coração de cinco primatas (humano, chimpanzé, gorila, orangotango e rhesus), oriundas de trabalhos previamente publicados, foram obtidas pela plataforma *Sequence Read Archive* (SRA), em especial, dois trabalhos serviram como base para as análises de expressão: SRP007412 (Brawand et al., 2012) e (Scally et al., 2012), cujo sequências foram obtidas após contato por email.

3.1.3. Genomas individuais.

As análises de polimorfismo de presença e ausência de retrocópias no genoma referência basearam-se em alinhamentos disponibilizados publicamente pelo projeto *1000 Genomes* (1000 Genomes Project Consortium, 2010). Para as

retrocópias específicas de humanos presentes no genoma referência utilizamos alinhamentos com sequências pareadas de 946 indivíduos. No total, utilizamos dados de aproximadamente 20.000 sequenciamentos da plataforma Illumina (*Illumina Genome Analyzer* e *Illumina Genome Analyzer II*) com fragmentos, em média, de 230.6 nucleotídeos. As análises piloto de detecção de retrocópias polimórficas ausentes no genoma referência humano basearam-se em um número mais restrito de indivíduos. Selecionamos os 20 indivíduos com maior cobertura de sequência em todos os indivíduos do projeto *1000 Genomes* (Tabela 1). Seis destes indivíduos compõem os dois trios (mãe, pai e filho) e seus sequenciamentos também foram utilizados para confirmar a genotipagem do pseudogene processado DHFRP1.

Tabela 1. Número de bases sequenciadas e cobertura de cada genoma.

Amostra	Número de bases sequenciadas	Vezes de cobertura
NA12878	181.137	58,43x
NA12892	142.265	45,89x
NA19240	126.464	40,79x
NA12891	122.661	39,57x
NA19239	78.071	25,18x
NA19238	55.386	17,87x
AAC	28.216	9,10x
NA07346	25.148	8,11x
NA07347	24.433	7,88x
NA12045	24.272	7,83x
SJS	21.631	6,98x
NA11918	20.625	6,65x
NA11881	14.049	4,53x

Amostra	Número de bases sequenciadas	Vezes de cobertura
NA11894	12.637	4,08x
NA11931	11.717	3,78x
NA12287	11.560	3,73x
NA12043	10.576	3,41x

A revisão das análises de retrocópias ausentes no genoma referência, utilizando computação na nuvem, basearam-se em cem vezes mais indivíduos que a análise piloto. No total utilizamos o sequenciamento do genoma de 2.535 indivíduos para detectar novas inserções de retrocópias no genoma humano. Assim como na versão piloto, a maioria dos genomas foram sequenciados com aproximadamente três vezes de cobertura e com tamanho de fragmento, em média, próximo de 200 nucleotídeos.

Adicionalmente, também detectamos retrocópias polimórficas (retroCNVs) em dados de sequenciamento da plataforma SOLiD. A equipe do Instituto Ludwig responsável pelo sequenciamento construiu bibliotecas pareadas do genoma nuclear de dois indivíduos saudáveis de amostras doadas pelo Hospital Alemão Oswaldo Cruz. O sequenciamento destas bibliotecas foi realizado utilizando a plataforma de sequenciamento SOLiD 3.0. As leituras geradas foram alinhadas com o Bioscope v3.1 com parâmetros padrão (Tabela 1 - AAC e SJS).

A aluna de doutorado Paola de Avelar Carpinetti construiu bibliotecas pareadas (*mate-pair*) do material genômico de seis biópsias de tumores de cólon. Identificadas como AAS, CMCA, LIM, MM, MDS e SKE (Tabela 2). Adicionalmente, também foram construídas bibliotecas pareadas de amostras de sangue de três destes pacientes (Tabela2 - CMCA_normal, MM_normal e MDS_normal). O DNA

genômico destes indivíduos foi sequenciado nas plataformas SOLiD 4 e SOLiD 5500. Algumas amostras foram sequenciadas múltiplas vezes, portanto, o número de leituras geradas para cada indivíduo varia entre ~300 milhões e 1.7 bilhões de leituras (Tabela 2), com coberturas físicas variando entre aproximadamente 14 vezes para a amostra de sangue do paciente MDS e 100 vezes para a amostra tumoral do mesmo paciente. Leituras foram alinhadas contra o genoma referência GRCh37 com o alinhador Bioscope v3.1 e parâmetros padrão.

Tabela 2. *Compilação quantitativa das amostras sequenciadas.*

Amostras	Leituras	Leituras mapeadas	Cobertura	Cobertura física
AAS	393.756.912	322.877.464	4,07x	~19x
CMCA	1.266.261.844	982.159.574	8,65x	~69x
CMCA_normal	728.574.754	522.915.575	6,86x	~25x
LIM	385.789.584	305.832.709	3,70x	~18x
MDS	1.766.771.058	1.295.756.169	14,83x	~110x
MDS_normal	305.101.394	238.532.077	2,91x	~14x
MM	808.736.644	645.449.661	8,08x	~50x
MM_normal	682.517.714	600.827.214	8,55x	~30x
SKE	398.436.826	323.562.661	4,07x	~20x

3.2. Detecção de retrocópias no genoma referência

Para identificar retrocópias em genomas referência publicamente disponíveis, nós utilizamos a estratégia desenvolvida para a construção de um banco de dados de retrocópias (chamado por nós de RCPedia) (Navarro; Galante, 2013) e para identificar retroCNVs germinativos presentes no genoma referência (Schrider et al., 2013). Para tal, utilizamos as sequências dos transcritos maduros de genes

codificadores de proteínas obtidas de bancos de sequências públicos, por exemplo, para os primatas presentes no RefSeq (humano, chimpanzé, orangotango, rhesus, sagui e macaco esquilo) utilizamos todas as sequências de genes com identificadores iniciados com NM_ (transcritos codificadores validados) e XM_ (transcritos codificadores preditos). De forma similar, para gorilas, que estavam ausentes do RefSeq quando as análises foram realizadas, utilizamos todos os transcritos de genes codificadores de proteínas do ENSEMBL referentes a gorila. Baseado na principal característica de mRNA retrotranspostos pela maquinaria de transcriptase reversa dos elementos repetitivos LINEs, a ausência de introns, desenvolvemos as ferramentas necessárias para interpretar os resultados de alinhamentos da sequência de transcritos de genes codificadores de proteínas no genoma referência e selecionar os *loci* potencialmente originados por retrotransposição. Para tal, alinhamos as sequências de transcritos no genoma referência utilizando a ferramenta BLAT (parâmetros:-mask=lower; -tileSize=12; -minIdentity=75; -minScore=100). Posteriormente, selecionamos os alinhamos com identidade superior a 75% e, pelo menos, 50% ou 120 nucleotídeos alinhados no genoma referência. Alinhamentos contendo iatos longos (maiores que 15 mil pares de bases) foram excluídos das análises posteriores. Enquanto esse último filtro remove os casos com introns mais óbvios, ele também permite a presença de eventos com inserções de elementos repetitivos com LINES (~6.200 pares de bases), SINEs (~400 pares de bases) e retrovírus endógenos (~9.000 pares de bases). Os alinhamentos restantes são filtrados considerando: i) quais exons do gene parental estão presentes no alinhamento; ii) se há alguma junção de exons que encontram-se separados no gene parental; iii) e qual a porcentagem de cada exon foi alinhada. Finalmente, selecionamos os alinhamentos cujo, pelos menos, dois

exons adjacentes estão presentes, com limite inferior de 50 pares de bases alinhados, e são, portanto, selecionando possíveis ausências de introns de genes codificadores de proteínas no genoma referência. Como cada gene codificador de proteína pode ter mais de um transcrito, realizamos o agrupamento dos alinhamentos por coordenada genômica. Neste processo, todos os possíveis transcritos parentais de uma retrocópia são comparados para que o melhor alinhamento defina o gene parental mais provável. Esta comparação é feita baseando-se em: i) identidade da sequência analisada com o transcrito alinhado; ii) e sobreposição da sequência analisada com o transcrito alinhado. Por fim, são executadas dois filtros adicionais: i) verificamos as coordenadas genômicas dos genes parentais e suas possíveis retrocópias para eliminar artefatos de alinhamentos no próprio gene parental; ii) verificamos se a região anotada como retrocópia contem um segundo *locus* com regiões flanqueantes similares ao da retrocópia indicando, assim, uma duplicação genômica da região retrocopiada. Em ambos os casos, caso encontremos alinhamentos sobre o próprio gene parental ou evidência de duplicação genômica, removemos o *locus* das análises posteriores (Figura 7).

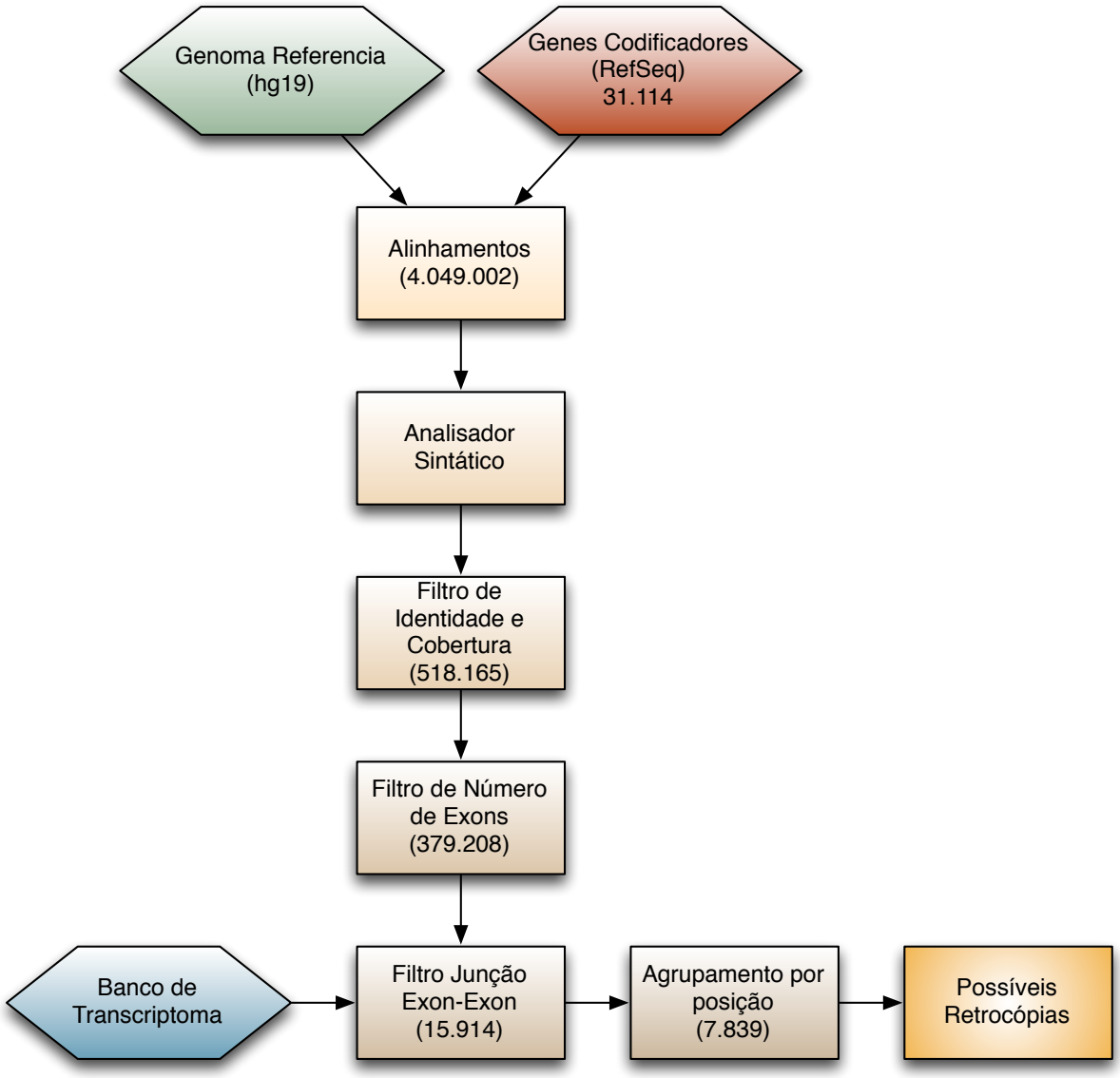


Figura 7. Fluxograma do *pipeline* de detecção de retrocópias no genoma humano.

3.3. Análise de contexto genômico

Para cada *locus* detectado e anotado como retrocópia, analisamos as características do contexto da inserção. Todos os eventos foram classificados quanto ao seu contexto genomico em: i) intragênicos ou intergênicos, baseado em coordenadas de transcritos codificadores e não-codificadores de proteínas; ii) centromérico ou telomérico, baseado nas coordenadas das respectivas regiões do UCSC genome browser; iii) proximidade a um poli(A) (eventos com distância menor

de 15 mil nucleotídeos de um sítio de poli(A)) e iv) proximidade aos sítios de início de transcrição, ambas coordenadas dos parâmetros iii e iv foram obtidas através de disponibilizados pelo GENCODE v15.

3.4. Caracterização das famílias de LINE1s em genomas referência

A fim de melhor entender a composição de elementos repetitivos nos genomas de primatas e roedores, utilizamos a ferramenta Repeat Masker que é especializada na detecção e caracterização de elementos repetitivos. Os genomas referência de todos os organismos analisados foram mascarados com os parâmetros padrões do Repeat Masker. Eventos foram agrupados pelo identificador fornecido pela ferramenta. O número total, proporção de sub-famílias e distribuição do comprimento dos eventos detectados foram gerados a partir do processamento da saída padrão do Repeat Masker.

3.5. Detecção de retrocópias ortólogas em genomas de eucariotos

Para entender o perfil de inserção e fixação de retrocópias no genoma de primatas e roedores, desenvolvemos uma estratégia para detectar eventos ortólogos baseado no compartilhamento de retrocópias em regiões sintênicas. Os organismos analisados por este trabalho tem tempo de divergência relativamente pequeno (Perez et al., 2013 e Steiper; Young, 2006), menos de 120 milhões de anos, o que garante uma identidade relativamente alta entre regiões sintênicas, mesmo que sejam sob seleção neutra. Para cada retrocópia detectada, definimos regiões flanqueantes como blocos compostos por três mil pares de bases (a montante e a jusante) do evento de retrocópia. Estes blocos são compostos por sub-blocos de, ao menos, 150 pares de bases de sequências não repetitivas. Para garantir que

nenhum fragmento das regiões flanqueantes fosse composto pelo evento detectado, utilizamos uma margem de segurança de cinco mil pares de bases a partir das extremidades da retrocópia. Cada uma destas regiões flanqueantes e suas respectivas retrocópias foram alinhados no genoma referência dos organismos a serem comparados utilizando o alinhador BLAT (parâmetros: -mask=lower; -tileSize=12; -minScore=50; -minIdentity=0). As retroduplicações com evidência de alinhamento da retrocópia dentro ou proxímo dos dois melhores alinhamentos da região flanqueante no genoma a ser testado foram classificadas como ortólogas entre ambos os organismos.

3.6. Análise de Ka/Ks

Para todos os genes com pelo menos um evento de retroduplicação, a sequência da região codificadora foi extraída utilizando informações de coordenadas definidas como CDS pelo RefSeq. Após remover as sequências repetitivas das retrocópias, utilizamos a ferramenta CLUSTALW2 (Larkin et al., 2007) para realizar os alinhamentos múltiplos entre cada sequência da retrocópia e seu respectivo gene parental. Posteriormente, baseado nas coordenadas das regiões codificadoras, removemos sequências referentes a regiões não traduzidas, adicionalmente, iatos nas regiões das retrocópias foram completados com sequências do gene parental e iatos nos genes parentais foram removidos das análises posteriores. Finalmente, utilizamos a biblioteca BioPerl (Stajich et al., 2002) para calcular os valores de Ka (substituições não sinônimas) e Ks (substituições sinônimas) de cada alinhamento múltiplo. Somente sequências com pelo menos uma mutação sinônima e uma mutação não sinônima foram consideradas em todas as análises de Ka/Ks.

3.7. Expressão de genes parentais

Para avaliar a expressão de genes parentais em tecidos germinativos utilizamos dados públicos de *microarray* de diversos tecidos de indivíduos saudáveis. Calculamos a expressão média de transcritos representados na plataforma *ABI Human Genome Survey Microarray Version 2*. Focamos somente as análises posteriores em três amostras de testículo e três amostras de ovário. Após normalizar os dados brutos de expressão utilizando o algoritmo MAS5, utilizamos os testes de Kolmogorov-Smirnov para comparar a distribuição do nível de expressão de genes com pelo menos um caso de retroduplicação contra a distribuição de dez mil grupos aleatórios de 2.570 genes sem nenhum evento de retroduplicação detectado.

3.8. Identificação de retrocópias expressas

Desenvolvemos duas estratégias distintas para detectar a expressão de *loci* anotados como retrocópias: i) para retrocópias intragênicas, buscamos por evidência de transcritos de expressão quimérica, ou seja, com alinhamentos reportando a "fusão" entre o gene hospedeiro e sua(s) retrocópia(s); ou ii) para todas as retrocópias, incluindo as intragênicas, nós buscamos por alinhamentos que fossem confiáveis e que evidenciassem a expressão da retrocópia.

Para detectar a transcrição quimérica entre a retrocópia e o gene hospedeiro, leituras de seis tecidos de cinco primatas (humano, chimpanzé, gorila, orangotango e rhesus) (Brawand et al., 2012) foram alinhados em seus respectivos genomas referência utilizando a ferramenta gsnap (Wu; Nacu, 2010) (parâmetros: -mask=lower; -tileSize=12; -minScore=50; -minIdentity=0). Posteriormente, selecionamos os alinhamentos com iatos onde uma extremidade foi alinhada sobre

regiões exônicas do gene hospedeiro e outra extremidade no *locus* anotado como retrocópia. Por fim, selecionamos os alinhamentos cujo iatos eram flanqueados pelos sítios canônicos de *splicing* (GT-AG), qualidade de alinhamento superior a 40 (escala Phred) e, ao menos, cinco leituras suportando o mesmo evento quimérico.

O alinhamento de leituras de transcritos parentais pode ser frequentemente alinhado em regiões genômicas com ausência de introns (por exemplo, em retrocópias), portanto, para avaliar a expressão de retrocópias, faz-se necessário o desenvolvimento de uma sequência de filtros para evitar falsos positivos. Para evitar tais falsos alinhamentos, primeiramente, criamos um “cromossomo” composto pelas sequências de transcritos maduros de todos os genes parentais do respectivo organismo e realinhamos todo o sequenciamento de transcriptoma utilizando a ferramenta bowtie2 (Langmead; Salzberg, 2012) (parâmetros: -mask=lower; -tileSize=12; -minScore=50; -minIdentity=0). Ao adicionar um cromossomo composto por todos os transcritos maduros de genes codificadores de proteínas, esperamos que leituras referentes ao gene parental, em especial as leituras sobre junções exon-exon, sejam “fisgadas” pelo cromossomo adicional e diminuam o número de alinhamentos falsos positivos sobre as retrocópias. Somente alinhamentos únicos e qualidade de alinhamento superior a 40 (escala Phred) foram selecionados para quantificar a expressão de *loci* anotados como retrocópia.

3.9. Interface *web*

Para construir o sítio RCPedia, utilizamos um conjunto de ferramentas e plataformas de software livre. Baseado nos resultados de detecção de retrocópias presentes no genoma humano, criamos uma estrutura de banco de dados mysql contendo as informações de genes, transcritos e retrocópias. As informações de

transcritos e genes foram obtidos a partir do RefSeq. Para as retrocópias, um conjunto de pós-processamentos foram realizados para disponibilizar aos usuários da ferramenta um conjunto de informações que poderiam auxiliar o melhor entendimento de eventos de um gene específico ou de um evento de interesse. Por exemplo, foram realizados alinhamentos múltiplos de todas as retrocópias e seus transcritos parentais, análise de conservação, detecção de possíveis repetições diretas flanqueando a inserção, expressão do *loci* anotado como retrocópia e uma análise de contexto genômico. Todas essas informações foram compiladas e carregadas em um banco MySQL (www.mysql.com). Para o desenvolvimento da interface, utilizamos a plataforma cakePHP (www.cakephp.org), que possibilita a criação uma interface gráfica ao fornecer a estrutura do banco de dados. Modificações foram realizadas no âmbito de interface e navegação, por exemplo, utilizando bibliotecas jQuery (jquery.com), incluímos a possibilidade de omitir ou mostrar parte da interface, de forma que fosse permitido ao usuário personalizar a interface conforme seu interesse. Para facilitar a busca por eventos de interesse, também desenvolvemos um sistema de busca universal, que interpreta os termos de entrada em um único campo de busca e direciona a pesquisa para as funções mais adequada. Uma vez construído o esqueleto da ferramenta, expandimos os dados disponíveis para todos os primatas analisados, permitindo ao usuário comparar retrocópias ortólogas ou específicas de humanos, chimpanzés, gorilas, orangotangos, rhesus e saguis.

3.10. Identificação de retroCNVs presentes no genoma referência

Utilizando as informações de retrocópias presentes no genoma referência desenvolvemos um *pipeline* para verificar quais das 7,831 retrocópias presentes no

genoma referência humano (hg19/GRCh37) teriam evidência de ausência em indivíduos sequenciados pelo projeto *1000 Genomes* (1000 Genomes Project Consortium, 2010). A estratégia baseia-se em baixar todos os alinhamentos de leituras pareadas em regiões contendo uma retrocópia e, utilizando o *samtools* (Li, H. et al., 2009) e o ftp do projeto *1000 Genomes*, selecionar todas as leituras com perfil anormal. Para tal, removemos das análises posteriores todos os alinhamentos de leituras pareadas e com tamanho de fragmento estimado menor que a mediana dos tamanhos estimados mais dois desvios padrão. Também eliminamos das análises posteriores quaisquer pares cuja orientação é diferente da esperada ou estão em cromossomos diferentes. Finalmente, selecionamos todos os alinhamentos cujos pares flanqueiam as coordenadas das retrocópias e têm distância maior que a distância padrão dos pares somado ao tamanho da retrocópia analisada. Após selecionarmos as evidências com, pelo menos, cinco pares de leituras suportando a ausência da retrocópia no indivíduo sequenciado, inspecionamos manualmente cada agrupamento e região a fim de evitar falsos positivos (Figura 8A).

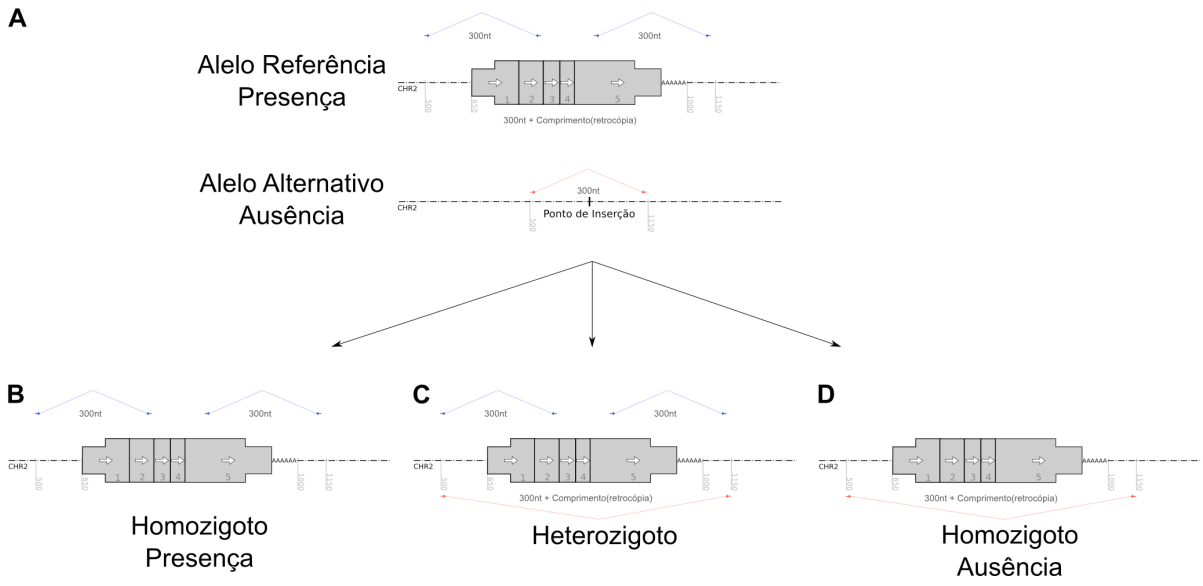


Figura 8. Diagrama com perfil de alinhamento de alinhamentos reportando ausência ou presença de retrocópias presentes no genoma referência. **A)** Representação livre dos alelos referência (com retrocópia) e alelos alternativos (sem retrocópia) e perfil de alinhamentos hipotético em cada um dos casos. **B)** Representação de alinhamentos no genoma referência para indivíduos homozigoto presença. **C)** Heterozigoto. **D)** Homozigoto ausência.

3.11. Identificação de retroCNVs ausentes no genoma referência

A priori, qualquer "nova" sequência não representada no genoma referência requer uma análise extensiva de sequências não alinhadas ou de todos os alinhamentos do genoma sequenciado para que sua coordenada seja definida. Este processo é muito exigente do ponto de vista computacional. Portanto, para avaliarmos a existência de retrocópias não representadas no genoma referência humano, avaliamos vinte indivíduos do projeto *1000 Genomes* com maior cobertura de sequência. O *pipeline* desenvolvido baseia-se em avaliar todos os alinhamentos de leituras pareadas gerados pelo sequenciamento do genoma de cada indivíduo. Para detectar sequências não representadas pelo genoma referência, removemos todos os alinhamentos onde a distância dos pares é similar à distância da biblioteca construída, feito isso, agrupamos por coordenada todos os alinhamentos com perfil anormal. Para tal, definimos janelas de 8011 nucleotídeos e selecionamos todos os agrupamentos (*clusters*) com mais de 3 leituras suportando a anormalidade. Posteriormente, filtramos todos os agrupamentos por posição genômica, e selecionamos todos possíveis agrupamentos correspondentes a possíveis genes parentais ao selecionar agrupamentos cujo pelo menos um dos lados do agrupamento sobrepõe uma região anotada como exônica de um gene codificador

de proteína. Finalmente, verificamos se todos os alinhamentos agrupados na região do possível gene parental estão sobre exons, e consultamos mais uma vez todos os alinhamentos do indivíduo para verificar se existem leituras adicionais em outros exons que reportam a mesma retroposição. Definido o gene parental, implementamos filtros adicionais para eliminar possíveis agrupamentos sobre retrocópias presentes no genoma referência ou sobre elementos repetitivos. Finalmente, realizamos uma triagem manual para eliminar possíveis falsos positivos (Figura 9).

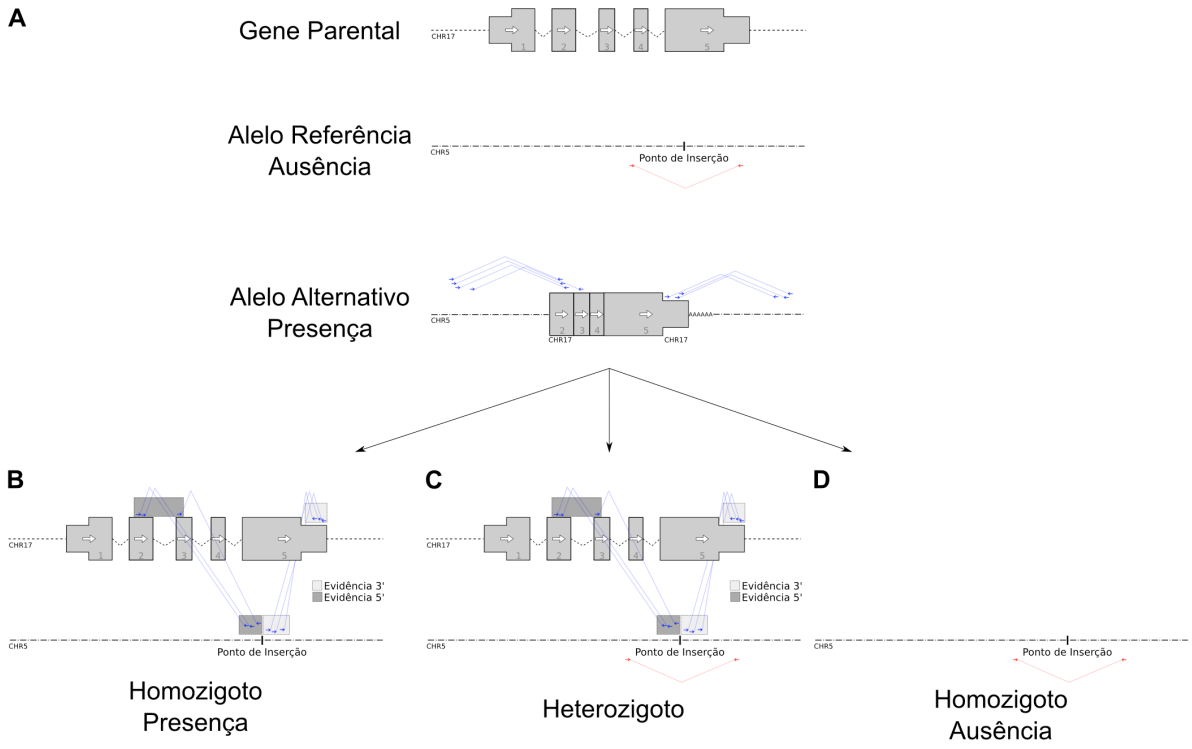


Figura 9. Diagrama com perfil de alinhamento reportando ausência ou presença de retrocópias ausentes no genoma referência. **A)** Representação livre dos alelos referência (sem retrocópia) e alelos alternativos (com retrocópia) e perfil de alinhamentos hipotético em cada um dos casos. **B)** Representação de alinhamentos

no genoma referência para indivíduos homozigoto presença. **C)** Heterozigoto. **D)** Homozigoto ausência.

3.12. Genotipagem dos retroCNVs

Para avaliar a frequência alélica de cada um dos retroCNVs germinativos detectados, analisamos as leituras alinhadas sobre o ponto de inserção e sobre as retrocópias. Para os casos de retroCNVs presentes no genoma referência a análise é mais simples: Contamos o número de alinhamentos que sobrepõem a borda da retrocópia e consideramos estas leituras como evidência de presença da retrocópia, de maneira similar, contamos o número de pares que flanqueiam a sequência da retrocópia, mas não a sobrepõe e consideramos estas leituras como evidência de ausência. Indivíduos com ambas evidências são classificados como heterozigotos (Figura 8C); apenas evidência de presença ou apenas evidência de ausência, como homozigoto presença (Figura 8B) e homozigoto ausência (Figura 8D), respectivamente. De maneira similar, os retroCNVs germinativos ausentes no genoma referência também podem ser genotipados, porém, existem algumas peculiaridades na genotipagem destes eventos. Primeiro, para definir a evidência de ausência precisamos definir, sem muita margem de erro, o ponto de inserção. Após a validação e sequenciamentos realizada pelo Dr. Raphael Bessa Parmagiani, pudemos definir com precisão de nucleotídeos o ponto exato onde a retrocópia foi inserida. Portanto, alinhamentos pareados que flanqueiam o ponto de inserção sem sobrepo-lo foram considerados evidência de ausência do retroCNV e alinhamentos no gene parental pareados com o ponto de inserção foram considerados evidência de presença. Assim como os retroCNVs germinativos presentes no genoma

referência, classificamos cada uma das amostras como heterozigoto (Figura 9C), homozigoto presença (Figura 9B) e homozigoto ausência (Figura 9D).

3.13. Identificação de retroCNVs somáticos

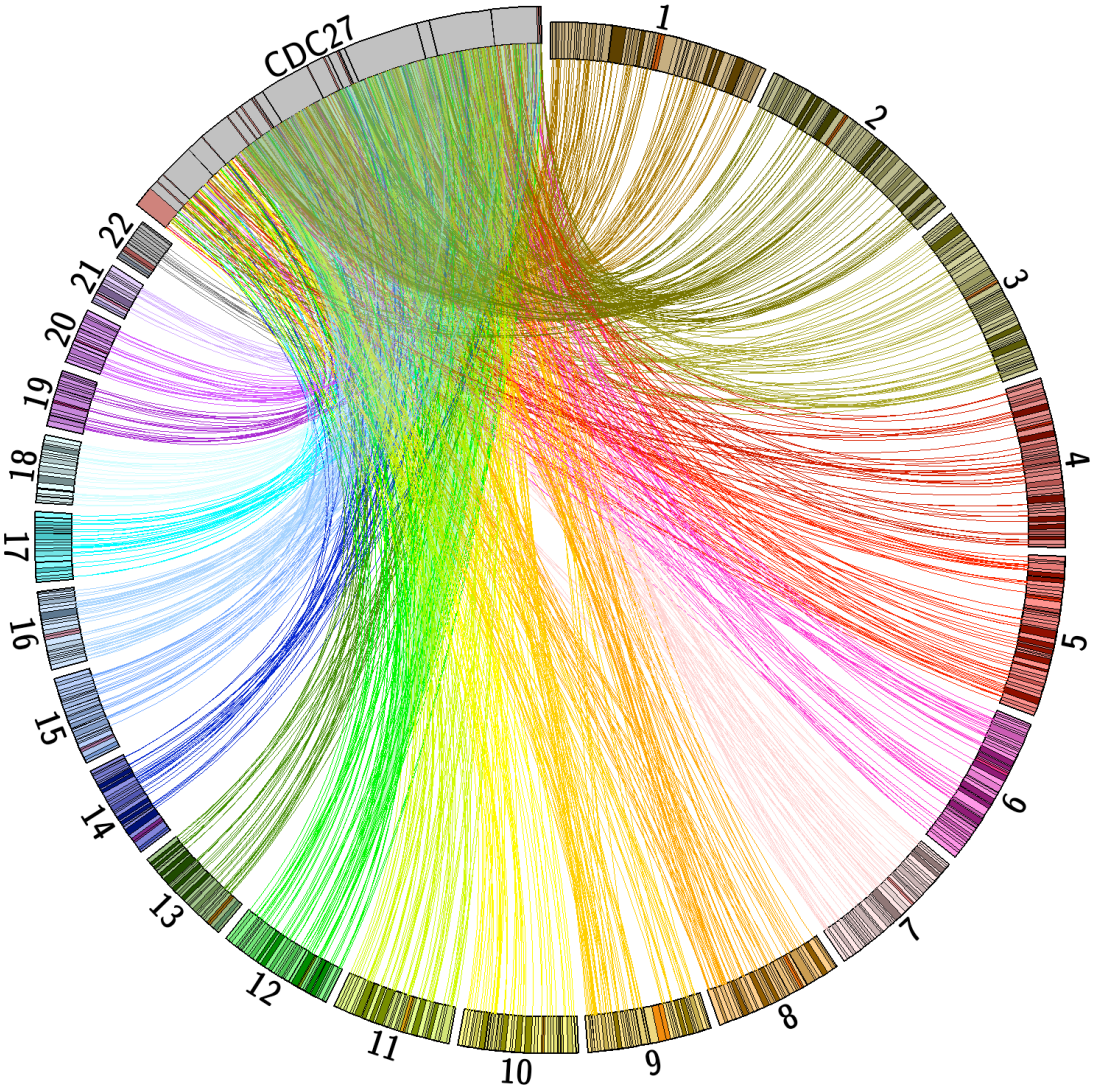
Similar à primeira versão do *pipeline* de detecção de retroCNVs germinativos, a detecção de retroCNVs somáticos se dá pela análise de todas as sequências de cada genoma sequenciado. Como retroCVNs somáticos não são recorrentes, não faz sentido avaliar todas as amostras como um único conjunto de leituras. Portanto, utilizamos uma estratégia muito similar à desenvolvida inicialmente para os retroCNVs germinativos, no entanto, melhoramos alguns filtros. Por exemplo, o filtro de alinhamento em regiões exônicas, o qual passamos a utilizar os dados do GENCODE de genes codificadores e não codificadores de proteínas.

Para estudar a frequência e impacto dos retroCNVs somáticos em tumores de cólon, focamos nossos esforços em aplicar os métodos descritos abaixo em sequenciamentos originais da plataforma SOLiD 4 e SOLiD 5500 feitos pelo nosso grupo no Instituto Ludwig. Inicialmente, o *pipeline* seleciona todos os alinhamento de pares que estão fora do perfil esperado. Nominalmente, selecionamos todos os alinhamentos cujos pares estão a uma distância maior que 10.000 pares de bases ou em cromossomos diferentes. Posteriormente, selecionamos todos os pares “anormais” de cada possível gene parental (coordenadas obtidas do transcritos do GENCODE v16) (Representação gráfica do *pipeline* para um possível gene parental (CDC27) - Figura 10A). Estes pares anormais são então agrupados baseado em coordenadas de ponto de inserção e são selecionados somente os agrupamentos com suporte maior que três leituras (Figura 10B). Entretanto, apesar de suportados por três alinhamentos confiáveis, a alta taxa de sequências repetitivas do genoma

humano pode favorecer o surgimento de artefatos de alinhamento. Para evitar possíveis falsos positivos, desenvolvemos filtros adicionais para garantir que o número de falsos positivos nos possíveis pontos de inserção fossem minimizados e os alinhamentos nos genes parentais suportassem eventos de retroduplicação. Inicialmente, para diminuir consideravelmente o número de *loci* analisados e assim diminuir o custo computacional do nosso *pipeline*, eliminamos todos os agrupamentos que apresentam extremidades em introns, estes agrupamentos não devem ter como origem retroduplicações, pois, estas seriam estritamente exônicas. De maneira similar, removemos todos os agrupamentos em que o ponto de inserção com mais de 90% de sobreposição com um elemento repetitivo, eliminando a possibilidade da retroposição haver acontecido no sentido contrário do esperado (por exemplo, um LINE1 sendo retrocopiado na região 3' de um gene) (Figura 10C). Uma vez eliminados inúmeros eventos e diminuído, consideravelmente, o custo computacional dos eventos restantes, voltamos à analisar cada alinhamento do agrupamento individualmente. Nesta etapa, verificamos se pelo menos 90% dos alinhamentos estão sobre exons (evidenciando ausência de introns). Antes de enviar estes eventos para a validação por PCR são necessários dois filtros adicionais. Apesar de termos eliminado os falsos positivos mais óbvios, sabemos que parte dos agrupamentos restantes podem ser artefatos de retrocópias presentes no genoma referência, portanto, como último passo de remoção de falsos positivos alinhamos a região potencialmente parental contra a região de inserção. Caso haja um alinhamento minimamente confiável, eliminamos o evento das análises posteriores (Figura 10D). Finalmente, sabemos que algumas retrocópias não estão presentes no genoma referência (Schrider et al., 2013), apresentando um perfil polimórfico em diversas populações humanas. Para eliminar retroCNVs germinativos fazemos uma

busca em todos os alinhamentos da região de interesse em 2.535 indivíduos do projeto *1000 Genomes* e analisamos se alguns indivíduos apresentam evidência da mesma retroduplicação em questão. Se encontrarmos cinco ou mais leituras evidenciando a mesma inserção, removemos o evento das análises posteriores. Este processo é repetido para cerca de 28.000 potenciais genes parentais.

A)



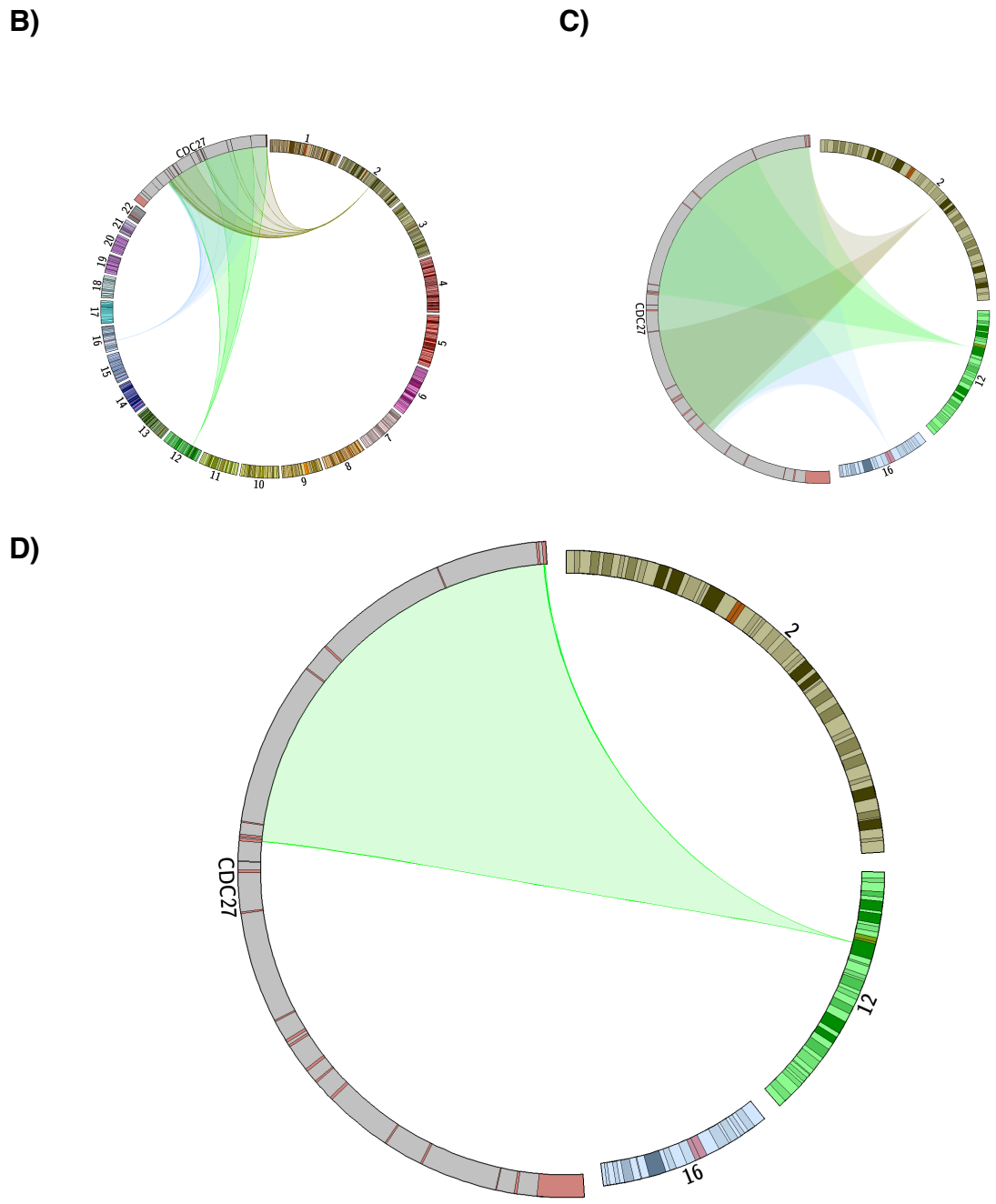


Figura 10. Representação gráfica, baseado na ferramenta circos, dos sucessivos filtros do *pipeline* de detecção de retroCNVs somáticos. **A)** Representação de todos as leituras pareadas de um gene parental que sofreu retrocópia (CDC27). Barra mais externa, em cinza e vermelho, representa a sequência do gene parental, sendo que os segmentos vermelhos são regiões exônicas. Outras barras coloridas representam os cromossomos autossomos [1-22]. Ligações internas representam

pares anormais coloridos pela cor do cromossomo destino. **B)** Agrupamento de leituras usando suporte mínimo de 3 leituras suportando uma região de ~1Kb. **C)** Destaque para os agrupamentos gerados. Estão representados somente os cromossomos com pelo menos um agrupamento. **D)** Agrupamentos selecionados após os filtros de região exônica, região repetitiva, similaridade de sequências e potencial polimórfico.

Capítulo 4.
Resultados

“Em alguma prateleira de algum hexágono
(pensaram os homens) deve existir um livro
que seja a chave e o compêndio perfeito de
todos os demais”

Jorge Luis Borges - Ficções

4.1. Retrocópias no genoma humano

A publicação do rascunho do genoma humano em 2001 (Lander et al., 2001 e Venter et al., 2001), abriu caminho para uma série de estudos para atribuir, entre outras características, o sentido, a origem e sinais de seleção aos 3 bilhões de nucleotídeos que compõem o genoma da nossa espécie. Nós descrevemos e anotamos sequências do genoma referência humano (hg19), que possivelmente surgiram pela retroposição de transcritos maduros de genes codificadores de proteína. Baseado na implementação dos métodos descritos acima, encontramos 7.831 eventos de retroduplicação de 2.570 genes parentais e codificares de proteínas (Tabela 3). A fim de avaliarmos os métodos desenvolvidos para a anotação de retrocópias, comparamos quantitativamente e qualitativamente características dos *loci* anotados por nós contra resultados previamente descritos na literatura.

Tabela 3. Número de retrocópias e genes parentais no genoma humano.

Organismo	Número de retrocópias	Número de genes parentais
Humano	7.831	2.570

Caso a distribuição do número de retrocópias por gene parental fosse uniforme, esperaríamos encontrar ~3 retrocópias por gene parental. Entretanto, esta distribuição é similar a uma Poisson e a maioria dos genes parentais (1.516 ou 58,98%) tem apenas uma retroduplicação. No outro extremo deste espectro, temos apenas uma centena de genes (119 ou 4,63%) com mais de dez retroduplicações (Figura 11).

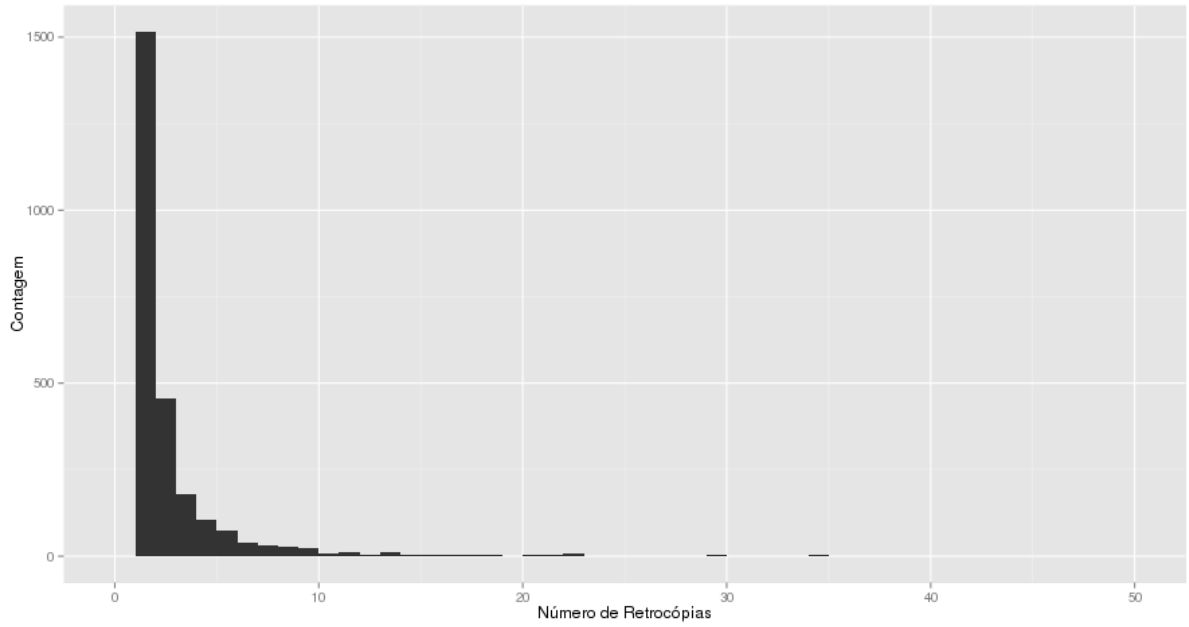


Figura 11. Distribuição do número de retrocópias para cada gene parental no genoma humano.

Compilamos a lista de genes mais retrocopiados e selecionamos os dez genes com o maior número de retrocópias (Tabela 4). Seis dos genes parentais mais retrocopiados tem função relacionada com proteínas da porção maior (RPL) ou menor (RPS) de ribossomos (Zhang, Z. et al., 2002). Os quatro genes parentais restantes apresentam função de manutenção do funcionamento celular básico, nominalmente, as proteínas codificadas pelos genes, PPIA, HNRNPA1, KRT18 e HMGN2 estão relacionadas, respectivamente, com Pepitidil Prolil Isomerase, hnRNPs, queratina e ligação ao DNA em nucleossomos. Este enriquecimento para genes relacionados a funções celulares básicas nos fez avaliar os enriquecimentos funcionais de todos genes retrocopiados (Zhang, Z. et al., 2004). Encontramos que grande parte (616 ou 41%) dos genes com pelo menos uma retrocópia estão em listas de genes de manutenção celular, os *housekeeping genes* (She et al., 2009).

Tabela 4. Genes parentais com maior número de retrocópias no genoma humano.

Gene Parental	Número de retrocópias
RPL21	148
PPIA	89
HNRNPA1	88
RPL23A	71
RPSA	70
RPL7A	67
KRT18	65
RPL31	62
HMGN2	59
RPL17	58

Ainda interessados em confirmar possíveis vieses na retroposição de transcritos de genes codificadores de proteínas, utilizamos dados públicos de expressão em células germinativas (testículo e ovário) para avaliar se o nível de expressão de genes com pelo menos um evento de retroduplicação é maior que os genes sem nenhuma retroduplicação. Encontramos que genes retroduplicados apresentam, em conjunto, uma expressão significativamente maior que a expressão de genes sem retroduplicações (p-valor=2,2x10-16, teste de Kolmogorov-Smirnov, Figura 12). Confirmando a correlação direta entre a expressão gênica em células germinativas e a chance do transcrito ser retrocopiado.

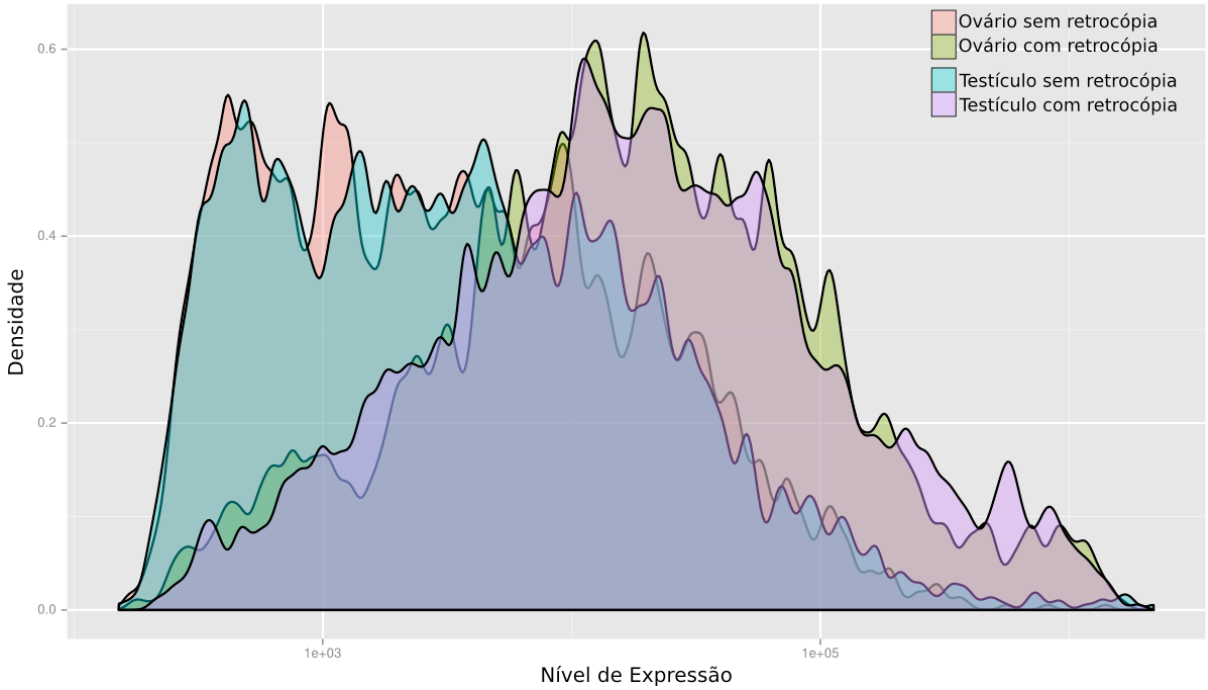


Figura 12. Distribuição do nível de expressão de genes com pelo menos uma retrocópia (roxo e verde) e genes sem nenhuma retrocópia (rosa e azul) em células germinativas.

Para investigar a aleatoriedade do ponto de inserção de retrocópias no genoma humano, buscamos por possíveis vieses de cromossomos aceptores de eventos de retroposição. Nossos resultados mostraram que os *loci* anotados como retrocópias estão distribuídos de forma aleatória pelos cromossomos autossomos (Figura 13A), indicando uma forte correlação entre o número de retrocópias e o comprimento de cada cromossomo humano (Figura 13B, correlação de Spearman 0.93, p-valor=7.28×10^{-12}).

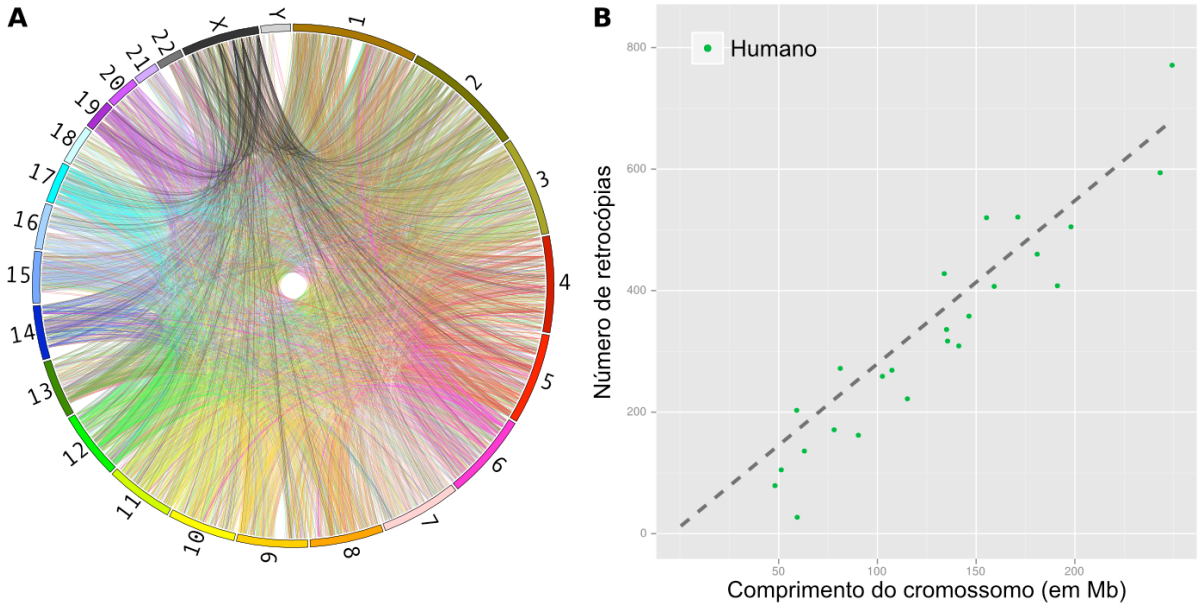


Figura 13. Retrocópias detectadas no genoma humano. **A)** Representação gráfica dos movimentos de sequências de genes codificadores de proteína por meio da retroposição. O anel mais externo representa os cromossomos humanos, as linhas internas representam os movimentos de sequências coloridos pelo cromossomo do gene parental. **B)** Correlação positiva entre tamanho do cromossomo e número de retrocópias no genoma humano.

Estes resultados indicam que na escala de cromossomos, não há nenhuma forma de seleção dirigindo a concentração de retrocópias em um autossomo específico, ou região específica. No entanto, também estávamos interessados em confirmar um possível viés para a importação e exportação de retrocópias de cromossomos sexuais, como já descrito por outros grupos (Emerson et al., 2004). De maneira muito similar a Emerson e colaboradores, porém em uma escala maior, encontramos um enriquecimento de retrocópias movidas do e para o cromossomo X que será melhor explorado. Para melhor avaliar o impacto destes eventos, contamos o número de eventos presentes em regiões intergênicas e intragênicas (na mesma

fita do gene hospedeiro, ou na fita oposta ao gene hospedeiro) (Figura 14). Dado que aproximadamente 40% do genoma humano é composto por introns e exons, utilizamos todos os transcritos presentes no RefSeq para verificar qual a porcentagem de retrocópias em regiões intragênicas e intergênicas. Em uma análise simplista, caso as retrocópias fossem neutras, esperaríamos que estas estivessem distribuídas de maneira independente de contexto (40% das retrocópias em regiões intragênicas). Encontramos uma sub-representação (aproximadamente 30%) de eventos em região intragênica (P<0.0002, chi quadrado), sugerindo uma seleção negativa para inserções de retrocópias em regiões transcritas.

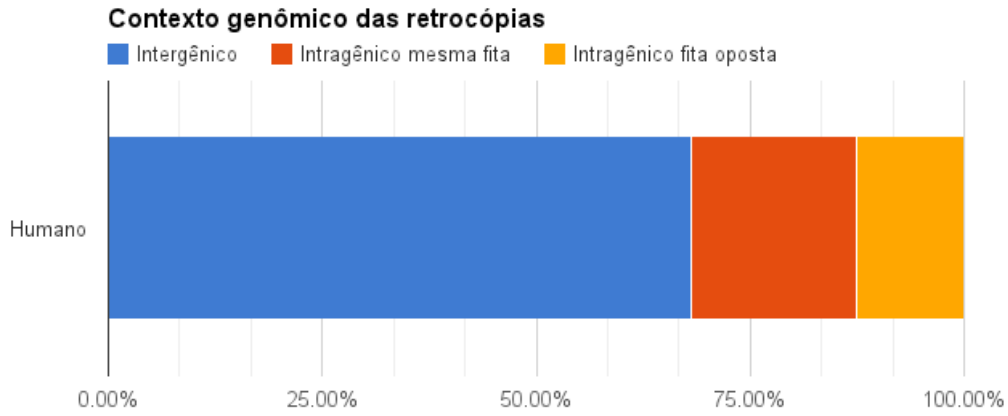


Figura 14. Porcentagem de retrocópias em regiões intergênicas e intragênicas.

4.2. Comparação entre RCPedia e bancos públicos.

Ao investigarmos a literatura de retrocópias e pseudogenes processados percebemos que não há um consenso sobre o número de retrocópias no genoma humano. Nossos resultados apresentam um número maior de eventos quando comparado com os primeiros métodos de detecção retrocópias (Khelifi et al., 2005 e Marques et al., 2005 e Ohshima et al., 2003 e Venter et al., 2001). Porém, quando comparamos estes resultados com versões mais recentes de banco de

pseudogenes processados (Baertsch et al., 2008 e Karro et al., 2007 e Pei et al., 2012), encontramos um número menor de eventos. Para verificar quão representativo o nosso catalogo de retrocópias é, comparamos nossas anotações com resultados de dois outros bancos de pseudogenes. Nominalmente, o banco pseudogene.org (Karro et al., 2007) e o GENCODE (Pei et al., 2012). O projeto GENCODE não foi concebido para catalogar pseudogenes no genoma humano, portanto, os dados disponibilizados pelo projeto não contem, por exemplo, informações sobre os possíveis genes parentais, dificultando a análise qualitativa dos *loci* anotados como retrocópias. Entretanto, baseado nas coordenadas genômicas dos eventos descritos, podemos analisar quantitativamente o número de *loci* concordantes entre os três métodos.

Inicialmente, encontramos que o projeto GENCODE e o banco pseudogene.org reportam 10.455 e 8.215 pseudogenes processados, respectivamente. Ambos os bancos apresentam concordância relativamente baixa, apresentando apenas 61% (aproximadamente 6.300) dos eventos presentes do GENCODE também estavam no banco *pseudogene.org*. Aproximadamente 10% das retrocópias descritas pelo nosso *pipeline* não foram detectadas por nenhum destes bancos. Dado que o GENCODE é a atual referência para anotação de genes e transcritos (Pei et al., 2012), analisamos a qualidade dos nossos dados em comparação com este banco. Seis mil setecentos e oitenta e oito retrocópias são compartilhados por ambos os bancos (86.88% das nossa retrocópias). Desta maneira, restaram 3.667 pseudogenes processados específicos do GENCODE (potencialmente falsos negativos em nossos resultados ou falsos positivos no GENCODE) e 1.043 eventos específicos em nossos resultados (potenciais falsos positivos em nossos resultados ou falsos negativos no GENCODE).

Dado que estes números de eventos específicos de cada método são relativamente altos, decidimos analisar manualmente um conjunto aleatório de candidatos para melhor entender as características de cada método. Nós analisamos 30 (~1%) e 20 (~2%) eventos encontrados exclusivamente no GENCODE e RCPedia, respectivamente. Para cada evento, nós alinhamos manualmente a sequência anotada como pseudogene processado no genoma referência humano (hg19) utilizando o alinhador BLAT (Kent, W J, 2002) disponibilizado pelo UCSC Genome Browser (Kent, W James et al., 2002). Entre esses pseudogenes processados, nós encontramos seis eventos que alinharam sobre genes parentais com um único exon, portanto, sem sobrepor com junções *exon-exon* e todos eles foram anotados manualmente pelo consórcio HAVANA (Tabela 5, eventos #25, #26, #27, #28 #29, e #30). Nós também encontramos eventos cujo alinhamento no genoma referência não nos permitiu identificar o gene parental que deu origem à retrocópia (Tabela 5, eventos #21, #22, #23, e #24). Provavelmente, estes eventos não foram descritos por nosso método porque nós nos baseamos, principalmente, na presença de pelo menos uma junção exon-exon e um gene parental para anotar retrocópias.

Além disso, outros fatores também poderiam explicar porque os eventos remanescentes não foram encontrados por nossos métodos. Por exemplo, nós encontramos sete eventos (Tabela 5, eventos #1, #7-#13) que tiveram origem após uma duplicação genômica de um *loci* com uma retrocópia. Nós também observamos pelo menos um caso classificado como retroduplicação de um gene sem introns (Tabela 5, evento #19), o que não está presente em nossos dados pela dificuldade de distinguir esse tipo de evento de uma outra forma de duplicação gênica. Também encontramos três eventos que parecem ser falsos positivos no GENCODE. Primeiro,

o ENSG00000257721.1 que é uma duplicação genômica contendo apensa um exon e os introns flanqueantes de seu possível gene parental, claramente, uma duplicação genômica sem envolver retroposição. Segundo, nós encontramos um *locus* multi-exônico (ENSG00000152117.13) com tamanho de 47kb sem evidência de retroposição. Terceiro, nós encontramos um LRT anotado como pseudogene processado (ENSG00000258073.1). Finalmente, nós também observamos três sequências exportadas do DNA mitocondrial (NumTS) no conjunto de pseudogenes processados no GENCODE v16. Já que pseudogenes processados são, por definição, resultado da atividade de uma transcriptase reversa (Kaessmann et al., 2009), e a maioria das sequências de mitocôndria exportadas para o DNA nuclear não se dão pela transcriptase reversa (Hazkani-Covo et al., 2003), estes eventos não foram incluídos em nossos resultados.

Tabela 5. Conjunto aleatório de pseudogenes processados (retrocópias) encontrados exclusivamente no GENCODE v16.

ID	Chr	Início	Transcrito Parental	Comprimento	Anotação Manual
1	chr14	19336524	ENSG00000257721.1	144	Duplicação genômica
2	chr2	132250386	ENSG00000152117.13	27608	Pseudogene não processado
3	chr19	58175648	ENSG00000269097.1	759	2 exons – Retrocópia antiga
4	chr16	31176969	ENSG00000263343.1	279	2 exons – Retrocópia antiga
5	chr2	131185304	ENSG00000230646.1	1494	3 exons – Retrocópia antiga
6	chr3	20049344	ENSG00000230697.1	395	Nenhum alinhamento
7	chr16	70113032	ENSG00000241183.1	495	Duplicação genômica
8	chr9	41776064	ENSG00000269692.1	1370	Duplicação genômica
9	chr15	82664459	ENSG00000237550.4	84325	Duplicação genômica
10	chr21	15148407	ENSG00000173231.6	1180	Duplicação genômica
11	chr22	16122720	ENSG00000215270.3	1048	Duplicação genômica
12	chr11	89498052	ENSG00000255170.2	254	Duplicação genômica
13	chr12	8559429	ENSG00000256136.1	362	Duplicação genômica
14	chr22	36568982	ENSG00000231576.1	1014	NumTs
15	chr9	42779843	ENSG00000225433.2	155	NumTs
16	chrX	102061669	ENSG00000229794.2	1083	NumTs
17	chr12	85333303	ENSG00000258073.1	144	Elemento repetitivo. LTR
18	chr16	34375269	ENSG00000260449.1	510	Elemento repetitivo. Satélite (SST1)

19	chr8	43139769	ENSG00000253707.1	180	Parental com um único exon
20	chrX	51453887	ENSG00000223591.4	485	Duplicação em tandem
21	chr12	34315397	ENSG00000256986.1	506	Gene parental indefinido
22	chr17	21476800	ENSG00000265363.1	210	Gene parental indefinido
23	chr11	50249920	ENSG00000255001.1	184	Gene parental indefinido
24	chr14	74005925	ENSG00000258408.1	560	Gene parental indefinido
25	chr8	13210910	ENSG00000253257.1	129	Sem junção de exons
26	chr4	29909281	ENSG00000249564.1	132	Sem junção de exons
27	chr9	41796924	ENSG00000231511.2	471	Sem junção de exons
28	chr12	25593809	ENSG00000255988.1	177	Sem junção de exons
29	chr2	75825197	ENSG00000230477.1	488	Sem junção de exons
30	chrX	27865705	ENSG00000232834.1	351	Sem junção de exons

Como evidenciado pela análise manual, acreditamos que, de maneira geral, a maioria dos eventos exclusivos do GENCODE (v16) foram excluídos de nossas análises devido aos parâmetros e filtros escolhidos. A maior parte desta diferença encontrada pode ser decorrente do GENCODE utilizar sequências de proteínas enquanto nós baseamos nosso *pipeline* em sequências de transcritos. Está claro que os métodos baseados em sequências de proteína são capazes de identificar eventos mais antigos, devido a sua maior sensibilidade durante o passo de alinhamento das sequências proteicas. Entretanto, os métodos baseados nas sequências de transcritos podem detectar eventos envolvendo somente regiões não codificadoras (ex: regiões 3‘UTRs) ou mesmo transcritos de genes não codificadores (Baertsch et al., 2008).

Para avaliarmos os nossos possíveis falsos positivos, nós analisamos manualmente 20 eventos aleatórios exclusivos da RCPedia. Inicialmente, observamos que dois eventos também são anotados pelo GENCODE, porém como genes codificadores de proteínas (Tabela 6, eventos #6 e #16). No entanto, estes *loci* claramente originaram-se pela atividade de transcriptase reversa, pois, são cópias sem introns de genes facilmente identificáveis como parentais. Portanto, há

forte indício de serem retrocópias. Ao não fazermos a distinção entre retrocópias funcionais (retrogenes) e não funcionais (pseudogenes processados) incluímos em nossos resultados eventos que são sabidamente anotados como genes.

Tabela 6. Conjunto aleatório de 20 possíveis retrocópias presente exclusivamente em nossos resultados.

ID	Chr	Início	Transcrito Parental	Comprimento
1	chr6	35038627	NM_001016	199
2	chr2	8897224	NM_001177	1256
3	chr2	74104255	NM_022494	1735
4	chr6	64190037	NM_021121	1794
5	chr7	44947961	NM_005274	399
6	chrX	56590436	NM_013438	2820
7	chr7	138913182	NM_001071775	800
8	chr1	185301590	NM_022818	814
9	chr22	22457789	NM_001085411	1302
10	chr17	63996465	NM_005796	843
11	chr20	11585629	NM_024674	4139
12	chr9	92324648	NM_021104	421
13	chr11	11202851	NM_004965	1111
14	chr5	94107897	NM_007209	210
15	chr2	65860969	NM_015933	160
16	chr2	70315029	NM_001128912	1249
17	chr17	63996465	NM_005796	843
18	chr11	56098383	NM_016255	1632
19	chr8	74743365	NM_002925	356
20	chr12	25070653	NM_001344	613

Todos os *loci* remanescentes estão ausentes do GENCODE. Conseguimos especular o motivo da ausência para três eventos, que apresentam uma grande proporção de elementos repetitivos em sua sequência e, portanto, podemos justificar pelo fato do nosso *pipeline* ser mais leniente com iatos de alinhamentos. Apesar da ausência de falsos negativos neste conjunto aleatório de retrocópias específicas do nosso *pipeline*, nós não acreditamos que os 1.026 *loci* específicos dos nossos resultados sejam todos verdadeiros positivos. No entanto, podemos estimar que a

taxa de falsos positivos entre os candidatos específicos à RCPedia é menor que 5% (1/20).

4.3. RCPedia

Apesar de alguns trabalhos disponibilizarem bancos de dados ou dados brutos sobre pseudogenes processados (Karro et al., 2007 e Pei et al., 2012) e retrocópias (Khelifi et al., 2005) no genoma humano, os autores destas ferramentas não se preocuparam em disponibilizar uma interface informativa, intuitiva e de fácil consulta para não especialistas nas áreas de retroposição ou pseudogenes. Com intuito de ressaltar a relevância de retrocópias na evolução de primatas e facilitar o acesso destas informações desenvolvemos uma ferramenta *web*, a RCPedia, do Inglês RetroCoPy encyclopEDIA. A ferramenta compila a maioria das informações geradas sobre retrocópias, genes parentais, pontos de inserção, expressão e compartilhamento de retrocópias em primatas.

Iniciamos o desenvolvimento da ferramenta para disponibilizar os resultados obtidos no genoma humano e compilamos todas as informações obtidas para cada retrocópia. Desenvolvemos uma perspectiva que é subdividida em blocos de informações que agrupam informações relacionadas sobre retrocópias (Figura 15). Neste exemplo, apresentamos as informações da retrocópia GABARAPL3, uma retrocópia do gene GABARAPL1. Características como identidade (94.18%), sobreposição com a transcrito parental (97.14%), possíveis repetições diretas flanqueando o evento, coordenada no genoma humano, fita de inserção e possível transcrito que deu origem a retrocópia fazem parte deste primeiro bloco de informação. Para visualizar graficamente o contexto genômico da retrocópia, integramos nesta perspectiva a *API sequence viewer*, desenvolvida pelo NCBI.

Neste bloco, é possível visualizar possíveis genes hospedeiros ou genes próximos da retrocópia, bem como anotações alternativas da região. O bloco de informação sobre o gene parental agrupa informações básicas como nome oficial, nome completo, nomes alternativos, coordenada, fita e um sumário de sua função fornecida pelo RefSeq. O bloco de ortologia representa o compartilhamento das retrocópias entre espécies de primatas (maiores detalhes nas sessões subsequentes), caso o evento tenha um ortólogo em outro organismo uma figura representando o organismo apresentará um tom mais escuro e, ao clicar na espécie, o usuário é redirecionado para uma página com informações da retrocópia na espécie de interesse. O bloco de informações sobre transcrição compila os resultados de expressão da retrocópia em seis tecidos, disponibilizando o número de leituras encontradas sobre o *loci* anotado como retrocópia. Finalmente, os dois últimos blocos compilam informações sobre o alinhamento múltiplo e sequências da retrocópia e o seu gene parental.

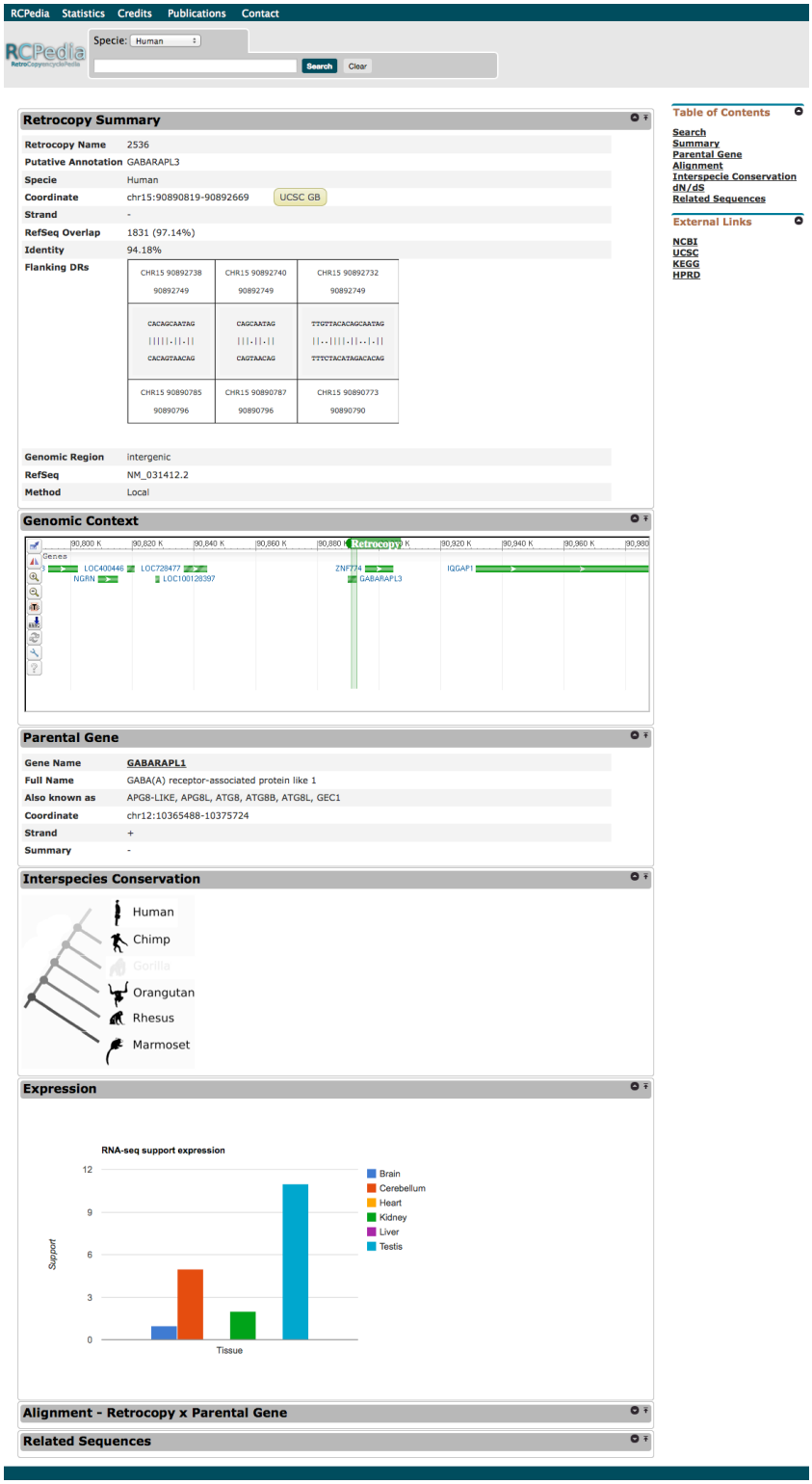


Figura 15. Dados segundo a perspectiva da retrocópia. Neste exemplo, são apresentados os dados de uma retrocópia do gene GABARAPL1.

Os usuários também tem acesso a uma segunda perspectiva que agrupa as informações sobre genes parentais (Figura 16). Neste exemplo, são compiladas as informações sobre o gene DHFR, o qual contem seis retroduplicações no genoma humano. O primeiro bloco disponibiliza informações gerais como nome oficial, nome completo, nomes alternativos, coordenadas, fita e um sumário simplificado de sua função. O segundo bloco representa os movimentos por meio de retroposições representadas pela ferramenta Circos (Krzywinski et al., 2009). O anel mais externos representam os cromossomos e as linhas internas os movimentos, coloridas pela cor do cromossomo de origem. Neste exemplo, o gene parental DHFR está no cromossomo cinco, portanto, as ligações internas (movimentos) terão a cor vermelha (cor do cromossomo parental). O bloco de informações sobre retrocópias compila todos os eventos de retroduplicação que tem como gene parental o gene de interesse. Finalmente, transcritos, sequências relacionadas e alinhamentos múltiplos estão, respectivamente, compilados nos três últimos blocos.

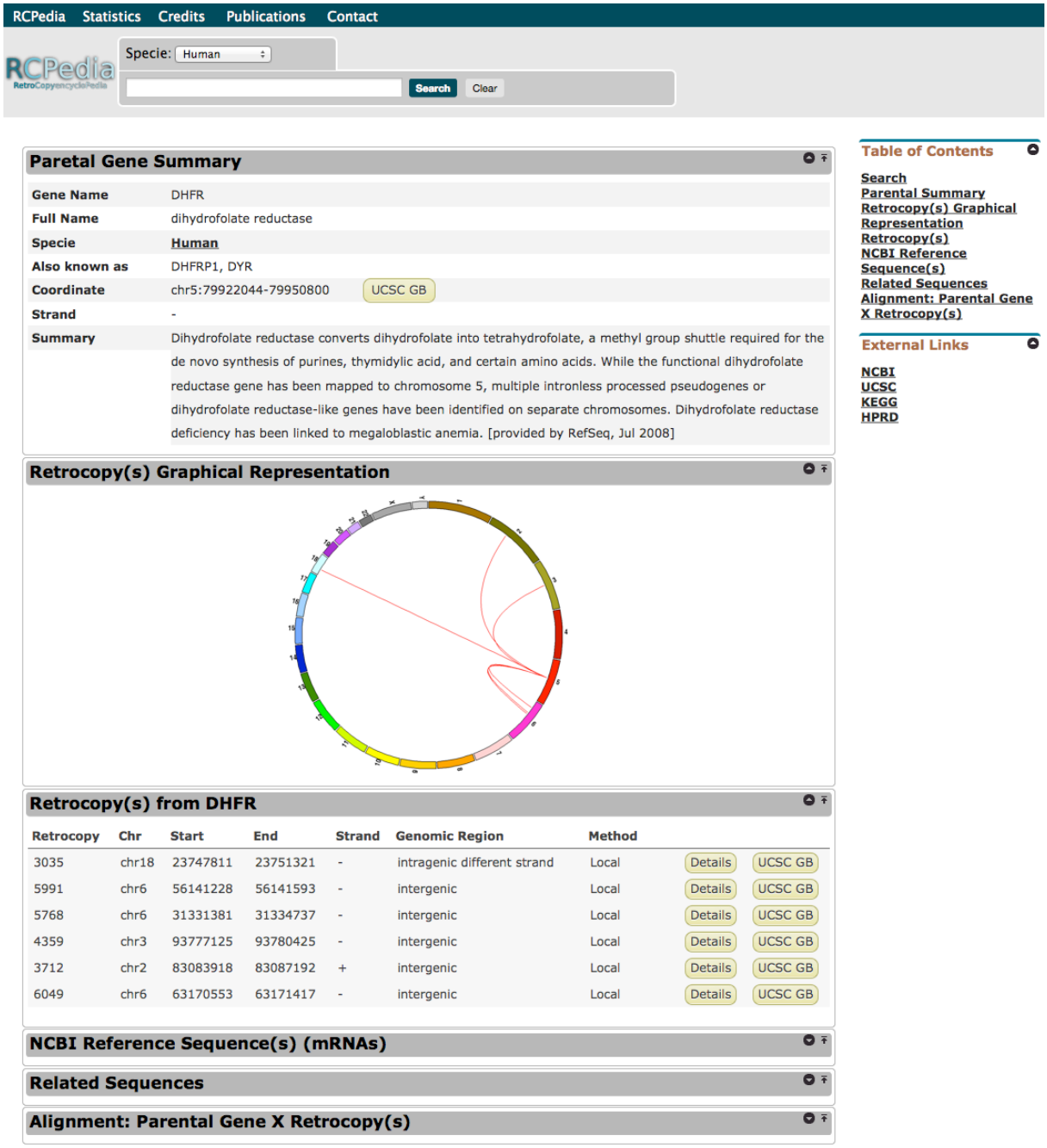


Figura 16. Dados organizados segundo a perspectiva do gene parental DHFR humano.

A principal forma de navegação pela RCPedia é a busca por termos. Nós desenvolvemos um campo de pesquisa que aceita diversos tipos de entrada e, hierarquicamente, busca possíveis retrocópias relacionados com o termos inseridos. De maneira geral, se a entrada estiver no formato de coordenada, retornamos todas

as retrocópias dentro das coordenadas fornecidas, caso contrário, a ferramenta relacionará o termo de busca com o código de retrocópias, nome de gene parental, nome do gene hospedeiro e descrição de gene. Na Figura 17 mostramos, por exemplo, uma busca pelo termo DHFR, o nome oficial de um gene que, segundo a nossa ferramenta, apresenta seis retrocópias no genoma humano. Outro exemplo interessante é a busca por termos mais genéricos e potencialmente relacionados a diversos genes. Por exemplo, se o termo "*kinase*" for usado como entrada, verificaremos que não há retrocópias ou nomes oficiais de genes parentais idênticos ao termo de busca e, portanto, retornamos todas as retrocópias de genes com o termo "*kinase*" na descrição ou nome completo. Neste exemplo, são retornadas 355 retrocópias, que podem ser ordenadas por qualquer um dos campos que descrevem o evento, facilitando a busca por eventos de interesse.

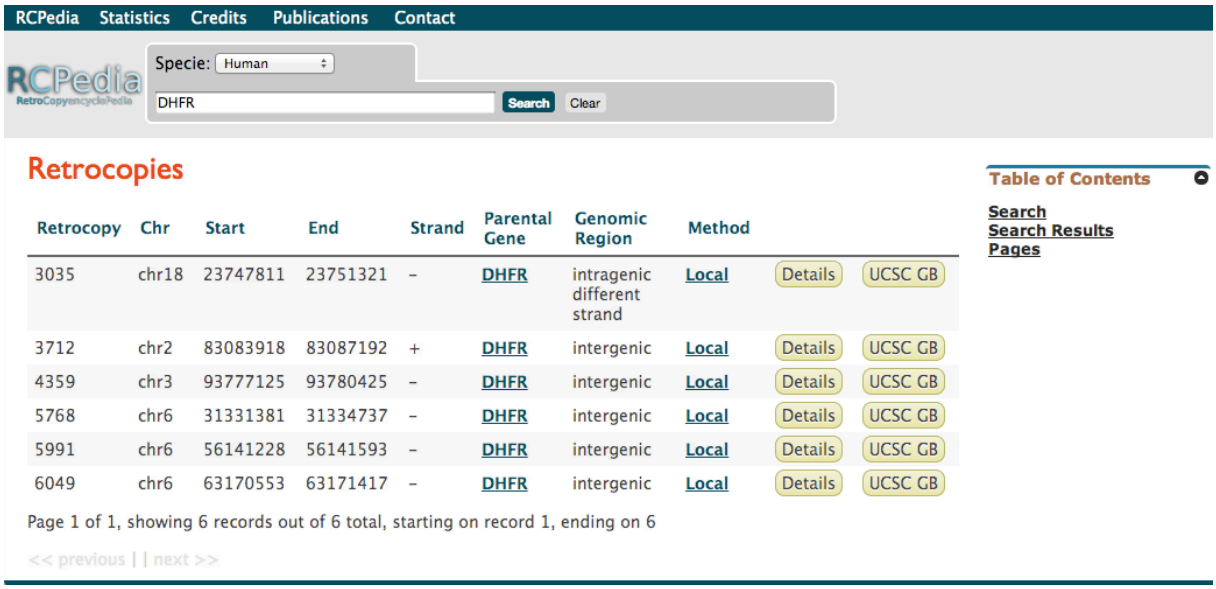


Figura 17. Busca por retrocópias do gene DHFR.

4.4. Detecção de retrocópias no genoma de primatas.

Visto que a detecção de retrocópias segundo nossos métodos e implementações estão de acordo com a literatura de retrocópias em humanos, expandimos nossa detecção e análise de retroduplicações de genes codificadores de proteínas para seis primatas, nominalmente, analisamos os genomas referência de chimpanzés, gorilas, orangotangos, rhesus, saguis e macacos esquilos. A escolha destes organismos baseou-se na qualidade de sequenciamento e montagem destes genomas. Inicialmente, avaliamos as características gerais destes genomas (Tabela 7).

Tabela 7. Composição geral dos genomas de primatas.

Organismo	Tamanho do Genoma	Numero de Genes	Número de Transcritos	Porcentagem do genoma composta por LINEs/SINEs
Humano	2,86Gb	19.364	32.201	22,32% / 13,89%
Chimpanzé	2,83Gb	20.998	33.616	22,23% / 13,66%
Gorila	2,92Gb	20.371	26.821	20,35% / 11,35%
Orangotango	2,94Gb	23.284	28.671	23,31% / 13,72%
Rhesus	2,93Gb	21.018	28.446	18,86% / 12,54%
Sagui	2,80Gb	18.739	23.275	21,34% / 13,45%
Macaco Esquilo	2,61Gb	23.577	25.608	18,95% / 13,01%

Os genomas de primatas são notavelmente similares, todos apresentam quantidades similares de material genético (~2.8Gb), com cerca de 20.000 genes codificadores de proteínas e 30.000 transcritos, com exceção do sagui, que, provavelmente pela baixa quantidade de seu transcriptoma, apresenta uma quantidade menor de genes e transcritos anotados. Baseado nas anotações do Repeat Masker, todos os primatas também apresentam uma porcentagem similar de

elementos repetitivos. Cerca de 20% dos nucleotídeos de seus genomas são LINEs e aproximadamente 13% são SINEs.

Após executar o *pipeline* de detecção de retrocópias no genoma referência destes organismos, também encontramos um número muito similar (~7.500) de retrocópias entre humanos, chimpanzés, gorilas, orangotangos e rhesus (Tabela 8), em contraste, encontramos aproximadamente dez mil retrocópias no genoma referência de saguis e macacos esquilos. Devido à baixa qualidade de montagem do genoma referência de macaco esquilo, avaliamos este organismo separadamente.

Tabela 8. Número de retrocópias e genes parentais no genoma de primatas.

Organismo	Número de retrocópias	Número de genes parentais
Chimpanzé	7.512	2.561
Gorila	7.709	2.669
Orangotango	6.873	2.439
Rhesus	7.502	2.453
Sagui	10.465	3.067

Para entender melhor o número elevado de retrocópias no genoma referência de saguis, investigamos a composição de elementos repetitivos no genoma referência de todos os primatas analisados. Apesar da composição de elementos repetitivos (LINEs e SINEs) ser muito similar entre todos os primatas, encontramos variações quando analisamos o número de elementos em cada subfamília de elementos LINE1. Enquanto o genoma referência de humanos, chimpanzés, gorilas, orangotangos e rhesus apresentam uma composição muito similar de subfamílias (Figura 18A), saguis tem uma porcentagem elevada de elementos da subfamília L1PA7 e L1P3. Esta subfamília corresponde a aproximadamente 30% e 5% dos

elementos L1 mais abundantes em saguis e apenas 5% e 1% dos elementos L1 mais abundantes em primatas do velho mundo. Como era de se esperar, saguis também apresentam um número absoluto elevado destes elementos, sugerindo que a subfamília L1PA7 e L1P3 esteve potencialmente ativa e, portanto, codificando a maquinaria de transcriptase reversa após a divergência de primatas do novo e velho mundo.

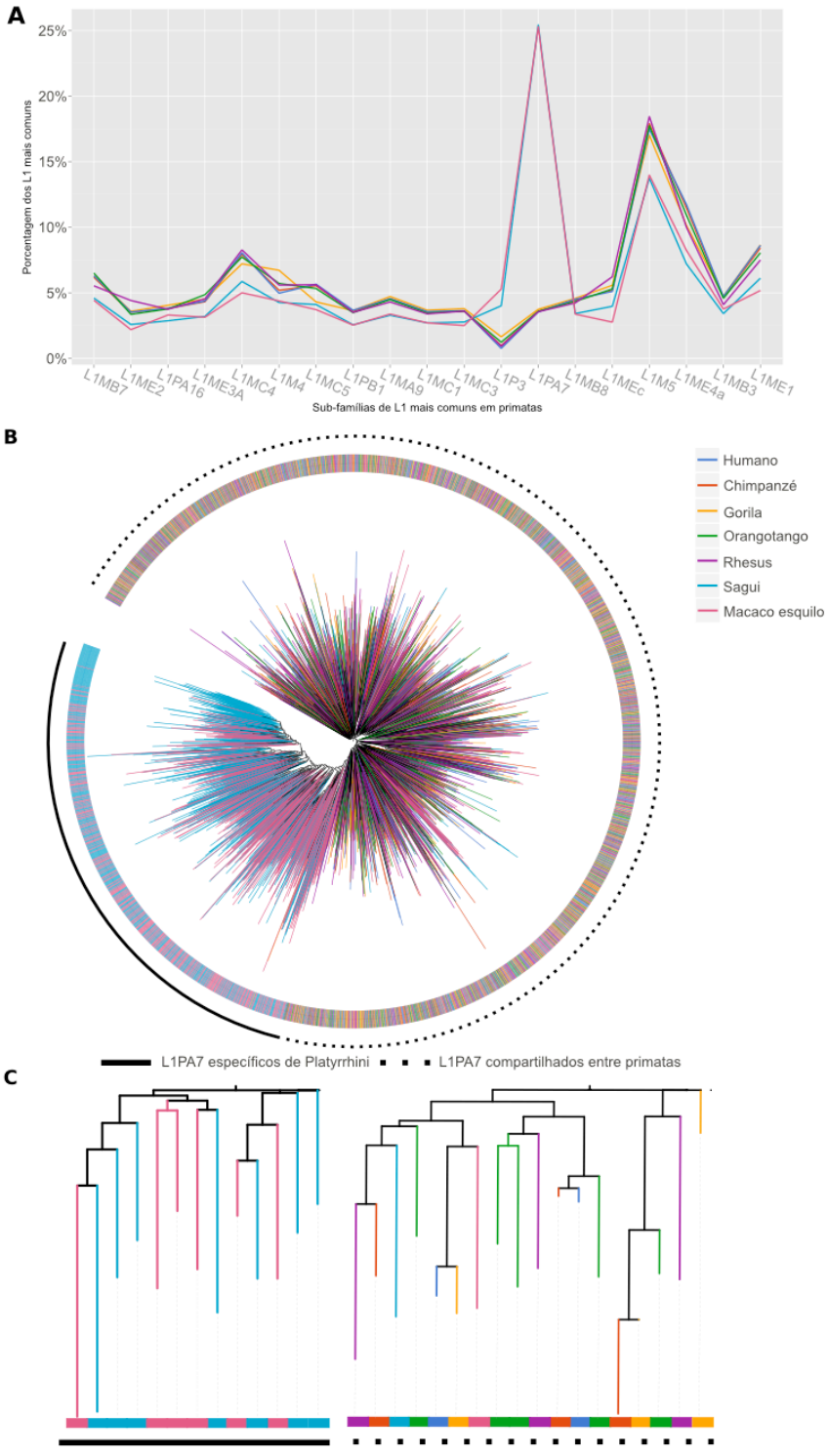


Figura 18. Representatividade de sub-famílias L1 nos genomas de humanos e outros primatas. **A)** Frequência de cada sub-família considerando as sub-famílias L1 mais frequentes em primatas **B)** Árvore filogenética resultante do alinhamento múltiplo das sequências de ORF2 de elementos L1PA7 no genoma de primatas

C) Fragmento da árvore filogenética (B) demonstrando elementos específicos de primatas do novo mundo (esquerda) e comuns a todos os primatas (direita).

Para entender em maior profundidade a discrepância entre o número de retrocópias em genomas de primatas do novo mundo e outros primatas, incluímos uma segunda espécie de primatas do novo mundo em nossas análises. *Saimiris bolivensis*, ou macaco esquilo, que apresenta aproximadamente 25 milhões de anos de divergência de saguis (Perez et al., 2013 e Steiper; Young, 2006). Para nossa surpresa, também encontramos um número elevado de retrocópias (9.320) no genoma deste primata. Ao avaliarmos o conteúdo de elementos repetitivos, verificamos que apesar de saguis e macacos esquilos apresentarem uma diferença no número absoluto de elementos L1 (provavelmente devido a qualidade da montagem do genoma de macaco esquilo), há uma grande semelhança no perfil de subfamílias L1 em ambas espécies (Figura 18A).

Para que haja maior atividade de elementos L1 em um genoma hospedeiro, elementos L1 tem de fugir dos mecanismos que restringem sua atividade. Hipoteticamente, esta fuga pode acontecer de duas formas: i) Inativação dos mecanismos de restrição, ou ii) mutação na sequência de um elemento L1 funcional, que diminua a eficiência de sua restrição. Para investigar a segunda hipótese, executamos um alinhamento múltiplo de todas as sequências da ORF2 dos elementos L1 anotados como L1PA7 de todos os primatas analisados. Ao investigarmos a árvore filogenética gerada pelo CLUSTALW2 (Larkin et al., 2007) (Figura 18B), verificamos que os ramos dividem-se em dois grupos. O primeiro conjunto, agrupa elementos L1PA7 similares em todos os primatas (18B borda externa colorida) e, provavelmente, agrupa eventos homólogos com origem anterior

à divergência de primatas do velho e do novo mundo. O segundo conjunto, predominantemente azul claro e rosa, agrupa elementos similares entre saguis e macacos esquilos e distintos dos elementos em outros primatas (18B borda externa azul e rosa). Portanto, o segundo agrupamento deve conter expansões específicas de espécies de primatas do novo mundo, talvez subfamílias novas anotadas erroneamente como L1PA7 pelo *Repeat Masker*. Indiretamente, a maior atividade de elementos L1 nestas espécies (Boissinot; Roos; et al., 2004b) indicam uma possível expansão que pode justificar o maior número de retrocópias nestes organismos.

Além de investigar possíveis causas para o aumento de retrocópias, nós avaliamos as características gerais de retrocópias nos genomas de primatas. Verificamos que outros primatas também apresentam i) forte correlação entre o comprimento do cromossomo e o número de retrocópias fixadas no cromossomo (Tabela 9), ii) sub-representação de eventos intragênicos (Figura 14), iii) super-representação de retroduplicações importados para e exportados do cromossomo X (Tabela S1) e, por fim, iv) um conjunto similar de genes com grande número de retroduplicações.

Tabela 9. Correlação entre número de retrocópias e comprimento do cromossomo.

Organismo	Correlação de Spearman
Chimpanzé	0.897865
Gorila	0.9321739
Orangotango	0.8946154
Rhesus	0.9175607
Sagui	0.9608696

Entre os genes mais retrocopiados, identificamos que o gene RPL21 é, consistente, o gene com mais retroduplicações em todos os primatas. Encontramos 148 retrocópias no genoma humano, 161, 117, 141, 130 e 181 em chimpanzé, gorila, orangotango, rhesus e sagui, respectivamente (Tabela 10). Assim como em humanos, encontramos um enriquecimento de retrocópias de genes relacionados com a porção maior e menor de ribossomos e funções básicas para o funcionamento celular.

Tabela 10. Genes parentais com maior número de retrocópias no genoma de primatas não humanos.

Chimpanzé		**Gorila**		**Orangotango**		**Rhesus**		**Sagui**	
RPL21	161	RPL21	117	RPL21	141	RPL21	130	RPL21	181
HNRNPA1	89	PPIA	73	ATP1A2	89	HNRNPA1	90	RPL29	134
RPL23A	77	RPL7A	69	RPSA	70	RPL23A	81	PPIA	127
RPSA	69	RPL31	64	RPL7A	67	RPL7A	74	RPL23A	117
KRT18	67	KRT18	63	RPL23A	66	KRT18	72	KRT18	104
RPL31	65	RPL23A	59	RPL39	57	RPL7	57	PRL1	92
MBOAT1	64	RPL7	50	RPL12	56	KRT8	57	KRT8	92
RPL7A	62	RPS3A	47	RPL36A	55	RPSA	54	RPS2	85
RPL7	59	RPL39	47	HMGN2	54	PPIA	54	RPL31	83
RPS26	55	HMGB1	46	KRT18	52	RPL39	52	RPSA	81

O grande número de cópias de um mesmo gene parental nos fez cogitar se seria possível definir quais retrocópias seriam o mesmo evento de retroposição compartilhado entre vários primatas e quais eventos seriam espécie específicos. Como uma análise piloto, realizamos um alinhamento múltiplo de todas as retrocópias do gene RPL21 de todos os primatas (Figura 19). De fato, enquanto algumas sequência retrocópias agruparam-se entre diferentes espécies, indicando

uma possível ancestralidade comum, outras sequências agruparam-se em grupos de uma única espécie, sugerindo que parte destas retrocópias aconteceram após a divergência entre os organismos analisados e, potencialmente, correspondem a eventos espécie específicos.

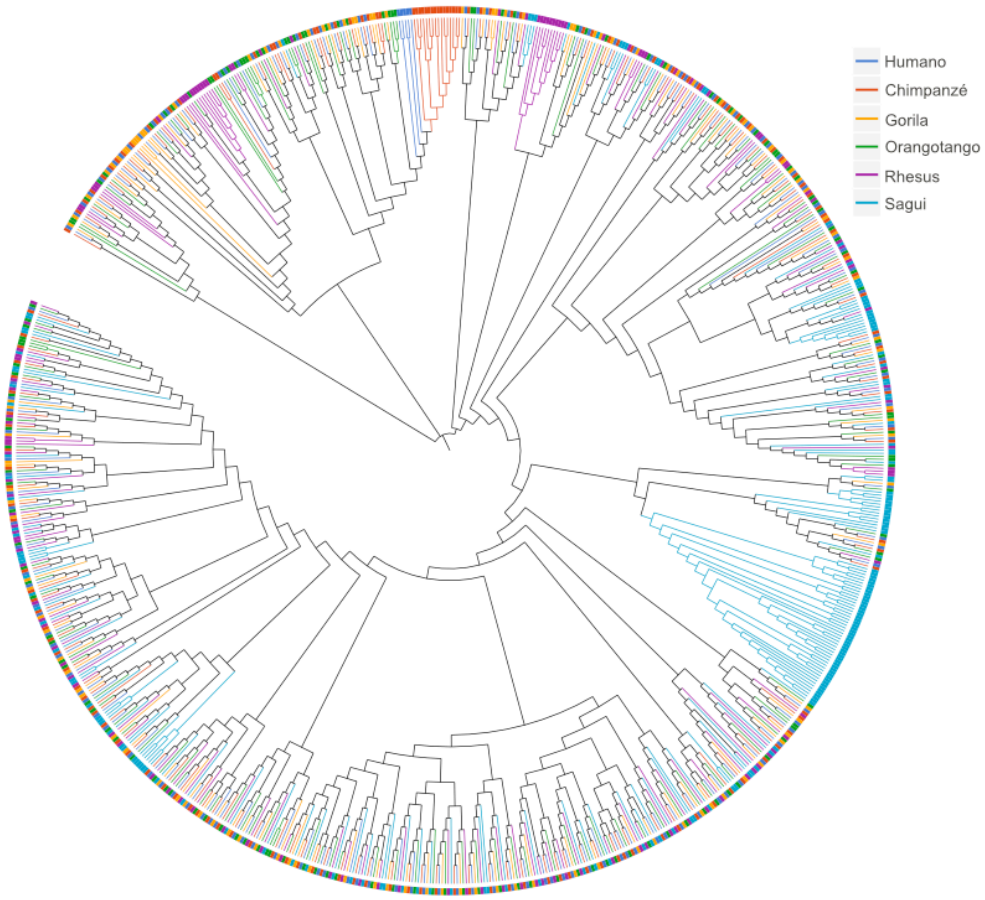


Figura 19. Árvore filogenética resultante do alinhamento múltiplo de todas as retrocópias do gene RPL21 do genoma de seis primatas.

4.5. Detecção de retrocópias ortólogas no genoma de roedores.

Esta análise nos fez questionar se seria possível implementar um método para avaliar a ancestralidade de cada uma das retrocópias anotadas nos genomas referências de primatas. Baseado no método que analisa retrocópias e regiões

sintênicas em outros organismos, inferimos se a inserção do evento aconteceu antes ou depois da divergência entre organismos par a par. Inicialmente, confirmamos que a maioria dos eventos presentes em humanos também estão presentes em todos os outros primatas analisados (4.168 ou 50.50% - Figura 20). Portanto, cada uma destas retrocópias é um evento independente de retroposição em uma célula germinativa em um ancestral comum, que hoje, encontra-se em todos os primatas.

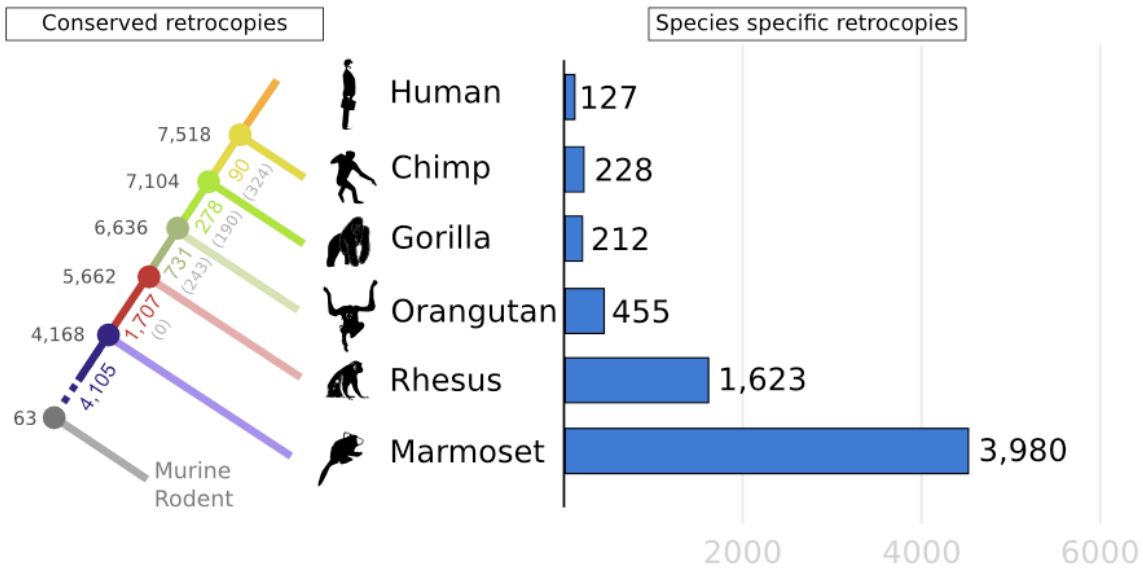


Figura 20. Número de retrocópias compartilhadas e retrocópias espécie específicas analisadas. Números em cinza escuro representam o número de retrocópias humanas compartilhadas entre todos os organismos até a respectiva altura da árvore filogenética. Números em cinza claro representam o número de retrocópias humanas compartilhados entre humano e o organismo em questão (sem a necessidade que esteja nos organismos intermediários). Número coloridos representam o número de retrocópias humanas originados em cada período em questão.

A fim de confirmar a explosão de elementos repetitivos no genoma de primatas e o não compartilhamento de retrocópias entre primatas e camundongos (Ohshima et al., 2003), expandimos a detecção e análise de compartilhamento de retrocópias em genomas de camundongos e ratos. Inicialmente, descrevemos e anotamos retrocópias presentes em ambos genomas referência. Encontramos, respectivamente, 7.109 e 7.364 eventos de retroduplicação de genes codificadores de proteínas (Tabela 11) em camundongos e ratos.

Tabela 11. Número de retrocópias e genes parentais no genoma de roedores.

Organismo	Número de retrocópias	Número de genes parentais
Camundongo	7.109	2.205
Rato	7.364	2.114

Apesar do número de retrocópias em primatas e roedores ser similar (entre 7 e 7.5 mil), encontramos apenas 63 eventos, menos de 1% do total de retrocópias de cada espécie, compartilhados entre ambas linhagens (Tabela 12). Confirmando que o surgimento das retrocópias ocorreu de maneira independente e após a divergência de primatas e roedores. Curiosamente, 51 das 63 retrocópias compartilhadas entre primatas e roedores apresentam evidência de transcrição por transcritos do RefSeq. Quarenta e cinco eventos (71.42%) são codificadores de proteína, 4 transcritos não codificantes e dois eventos fazem parte da região exônica de um gene hospedeiro, totalizando 80% das retrocópias compartilhadas entre primatas e roedores com evidência de transcrição.

Tabela 12. Retrocópias compartilhadas entre primatas e roedores.

Gene Parental	Humano	Chimpanzé	Gorila	Orangotango	Rhesus	Sagui	Camundongo	Rato
FAM133B	X		X	X	X	X	X	X
RPL23A	X	X	X		X	X		X
RPL29	X	X	X	X	X		X	
ACTG1	X	X	X	X	X	X		X
GAPDH	X	X	X	X	X	X		X
H3F3A	X	X	X	X	X	X		X
HMGB1	X	X	X	X	X	X		X
HSPA8	X	X	X	X	X	X		X
SARNP	X	X	X	X	X	X		X
RPS2	X	X	X	X	X	X		X
PJA2	X	X	X	X	X	X	X	
RPL21	X	X	X	X	X	X	X	
OXCT1	X	X	X	X	X	X	X	X
TCEAL6	X	X	X	X	X	X	X	X
CHM	X	X	X	X	X	X	X	X
ATXN7L3	X	X	X	X	X	X	X	X
SMEK2	X	X	X	X	X	X	X	X
CNBP	X	X	X	X	X	X	X	X
PCBP2	X	X	X	X	X	X	X	X
PABPC4	X	X	X	X	X	X	X	X
TMEM151B	X	X	X	X	X	X	X	X
LDHAL6A	X	X	X	X	X	X	X	X
TKTL1	X	X	X	X	X	X	X	X
KLHL13	X	X	X	X	X	X	X	X
PDHA1	X	X	X	X	X	X	X	X
GK	X	X	X	X	X	X	X	X
CSTF2	X	X	X	X	X	X	X	X
FBL	X	X	X	X	X	X	X	X
ACTG2	X	X	X	X	X	X	X	X
ATP6V1E1	X	X	X	X	X	X	X	X
KPNB1	X	X	X	X	X	X	X	X
IPO5	X	X	X	X	X	X	X	X
PRPS1	X	X	X	X	X	X	X	X
UBA52	X	X	X	X	X	X	X	X
YY1	X	X	X	X	X	X	X	X
NAA10	X	X	X	X	X	X	X	X
CRK	X	X	X	X	X	X	X	X
FER	X	X	X	X	X	X	X	X
HNRNPH1	X	X	X	X	X	X	X	X
HNRNPH1	X	X	X	X	X	X	X	X
ACTR3	X	X	X	X	X	X	X	X
TUBA3C	X	X	X	X	X	X	X	X
RPL10	X	X	X	X	X	X	X	X
HSPA8	X	X	X	X	X	X	X	X
HSPA8	X	X	X	X	X	X	X	X
EPN1	X	X	X	X	X	X	X	X
MKRN1	X	X	X	X	X	X	X	X
SLC25A15	X	X	X	X	X	X	X	X
USP22	X	X	X	X	X	X	X	X
DCAF8	X	X	X	X	X	X	X	X
TAF9B	X	X	X	X	X	X	X	X

Gene Parental	Humano	Chimpanzé	Gorila	Orangotango	Rhesus	Sagui	Camundongo	Rato
RRAGB	X	X	X	X	X	X	X	X
LPCAT2	X	X	X	X	X	X	X	X
MFF	X	X	X	X	X	X	X	X
NKAP	X	X	X	X	X	X	X	X
DDI2	X	X	X	X	X	X	X	X
PAPOLA	X	X	X	X	X	X	X	X
WDR5	X	X	X	X	X	X	X	X
DNAJB6	X	X	X	X	X	X	X	X
KCNJ14	X	X	X	X	X	X	X	X
CHSY3	X	X	X	X	X	X	X	X
MORF4L1	X	X	X	X	X	X	X	X
GPR153	X	X	X	X	X	X	X	X

Visto que há um viés de movimentação de retrocópias de e para o cromossomo X, verificamos os movimentos de retrocópias compartilhadas entre primatas e roedores considerando o cromossomo de origem e cromossomo de inserção. Enquanto, por chance, é esperado que duas retrocópias fossem exportadas do cromossomo X e três retrocópias importadas para o cromossomo X, nós observamos 14 (p-valor=0.0032) e 13 (p-valor=0.016) retrocópias respectivamente. Portanto, aproximadamente 50% de todas as retroduplicações compartilhadas entre primatas e roedores estão relacionadas com o cromossomo X enquanto apenas 10% de todas as retrocópias no genoma humano envolvem este cromossomo. Diversos genes interessantes estão presentes na Tabela 12, por exemplo, nós identificamos que o gene PAPOLB, uma retrocópia do gene PAPOLA, tem expressão específica em testículo e codifica uma enzima que catalisa a polimerase de poli(A). Interessantemente, *knock-out* de PAPOLB, resulta na infertilidade causada pela prisão da espermatogênese (Kashiwabara et al., 2002). Adicionalmente, nós também encontramos genes sem função descrita, por exemplo, a retroduplicação do gene TMEM151B, que gerou o gene TMEM151A, codifica uma

proteína com dois domínios transmembranares e apresenta expressão específica em cérebro e cerebelo.

4.6. Detecção de retrocópias ortólogas no genoma de primatas.

Humanos e chimpanzés tem seu último ancestral comum a aproximadamente seis milhões de anos (Perez et al., 2013 e Steiper; Young, 2006), portanto, como era de se esperar, ambas espécies compartilham (mesmas retrocópias em um mesmo ponto de inserção) grande parte, 7.518 (96%, tomando como base as retrocópias de humanos), das retrocópias. Se a taxa de retroposição e fixação de retrocópias fosse constante durante a evolução de primatas, esperaríamos encontrar um número de retrocópias proporcional ao tempo de divergência do último ancestral comum comparado ao tempo total de divergência de primatas. Por exemplo, a divergência de humanos e chimpanzés (aproximadamente seis milhões de anos) corresponde a aproximadamente 14% dos 42 milhões de anos que separam humanos de primatas do novo mundo (Perez et al., 2013 e Steiper; Young, 2006). Portanto, se a taxa de retroposição de mRNAs e fixação fosse constante, esperaríamos que aproximadamente 14% das retrocópias em humanos (1.055) tivessem surgido após a divergência de humanos e chimpanzés, ou seja, fossem humano específicas. Entretanto, encontramos apenas 127 (1.67%) retrocópias específicas do genoma humano, indicando uma grande diminuição na taxa de retroposição e fixação de retrocópias no genoma humano. Notavelmente, a maioria dos eventos específicos de humanos (74%) apresentaram somente uma retrocópia por gene parental (Tabela S1). Há treze genes parentais que apresentaram mais de uma duplicação após a separação entre humanos e chimpanzés. Oito destes eventos são genes relacionados a proteínas do ribossomo (RPL22, RPL23A, RPL3, RPS28, RPL21,

RPL41, RPS26 e PSMC1) e os genes parentais restantes (AK4, CKS1B, PGAM1, RNF145 e RAP1GDS1) são genes relacionados com o funcionamento celular básico. Apesar de ser praticamente impossível inferir se há e qual a função destas retrocópias específicas dado o pequeno tempo de divergência destas espécies, alguns trabalhos estão endereçando questões relacionadas a alguns destes *loci*. Por exemplo, o *locus* chr15:35375427-35377509 é uma retroduplicação do gene NANOG e é anotado como um gene codificador de proteína, o NANOGP8, um "oncoretrogene" específico da espécie humana (Fairbanks et al., 2012).

Humanos, chimpanzés e gorilas apresentam apenas 127, 228 e 212 eventos sem ortólogo em outros primatas (Figura 20). Apesar de considerarmos esta lista de retrocópias espécie específicas com poucos falsos positivos, imaginamos que para os outros primatas como orangotango, rhesus e sagui haja uma maior porcentagem de eventos compartilhados com outros primatas não analisados devido a falta do sequenciamento de seus genomas. Este fato deve ser mais crítico para as 3.980 retrocópias específicas de sagui, pois o organismo mais próximo (macaco esquilo) tem pelo menos 25 milhões de anos de divergência, o equivalente ao tempo de divergência entre humanos e rhesus (Perez et al., 2013 e Steiper; Young, 2006). Mesmo assim, a fim de entender melhor o perfil de retrocópias em primatas do novo mundo, analisamos as retrocópias no genoma referência de saguis e avaliamos qual o número de retrocópias conservadas no genoma de macacos esquilo. Inicialmente avaliamos potenciais falsos negativos pelo genoma referência de macaco esquilo não estar montado em cromossomos, e, sim, em *contigs* e *scafolds* relativamente pequenos. Dos 10.465 *loci* com uma retrocópia no genoma de sagui, encontramos 10.188 regiões sintênicas equivalentes no genoma de macaco esquilo (97.35%).

Portanto, a porcentagem de falsos negativos pela ausência de *contigs* equivalentes deve estar na ordem de 3%.

Das 10.188 bordas presentes em macaco esquilo, 6.134 apresentam uma retrocópia similar à retrocópia de saguis. Portanto, aproximadamente 60% das retrocópias em saguis surgiram antes da divergência de macaco esquilo. Apesar de saguis e macacos esquilos compartilharem um número grande de retrocópias, quando avaliamos a porcentagem de retrocópias compartilhadas, observamos que o valor é relativamente similar a humanos e rhesus (52%) que também estão separados a aproximadamente 25 milhões de anos. Portanto, se considerarmos as retrocópias compartilhadas por todos os primatas e retrocópias compartilhadas entre primatas do novo mundo, cerca de 4.000 retrocópias (Figura 20) surgiram nos últimos 25 milhões de anos em saguis. Para testar esta hipótese, verificamos a distribuição da identidade dos genes parentais suas respectivas retrocópias no genoma de sagui, subdividindo as retrocópias entre compartilhados entre saguis e macacos esquilo e sagui específicos. Como esperado, as retrocópias compartilhadas entre saguis e macacos esquilo tem mediana da identidade igual a 90.54%, enquanto que, as retrocópias específicas de sagui apresentam uma mediana de 95.46%, próximo ao esperado dado os 25 milhões de anos de divergência.

A fim de estimarmos o taxa de origem e fixação de retrocópias durante a evolução de primatas, nós fizemos uma estimativa do número médio de retrocópias originadas e cada período da evolução de primatas (Tabela 13). No geral, nós encontramos um decaimento contínuo na origem e fixação de retrocópias de primatas. No início da ordem dos primatas (entre 42 e 30 milhões de anos atrás), encontramos uma média de aproximadamente 142 retrocópias por milhão de anos (1707/12). Nos dois próximos períodos encontramos um forte decaimento na taxa de

criação e fixação de retrocópias até que na linhagem de humanos, chimpanzés e gorilas, há um novo pico de 45 retrocópias por milhão de anos. Em contraste, a linhagem de humanos apresenta a menor taxa de origem e fixação de retrocópias (21 retrocópias por milhão de anos). Em primatas do novo mundo, a taxa de origem e fixação é a mais alta e similar ao período anterior a divergência de primatas, com 152 retrocópias por milhões de anos.

Tabela 13. Estimativa da taxa de origem e fixação de retrocópias em primatas.

Período	Número de retrocópias	Tempo de divergência	Retrocópias por milhões de anos (média)
0 – 6 ma	127	6 ma	~21
6 – 8 ma	90	2 ma	~45
8 – 18 ma	278	10 ma	~28
18 – 30 ma	731	12 ma	~61
30 – 42 ma	1.707	12 ma	~142
0 – 42 ma	6.397	42 ma	~152
42 – 90 ma	4.105	48 ma	~85

ma: milhão de ano

4.7. Retrocópias polimórficas germinativas.

A identificação de 127 retrocópias específicas em humanos implica na retroposição de genes codificadores de proteína após a divergência entre humanos e chimpanzés. Podemos admitir que a maioria destas 127 retrocópias surgiu em ancestrais humanos como alelos raros e aumentaram de frequência com ou sem influência de seleção natural há, no máximo, cinco milhões de anos atrás. Visto que retrocópias não surgiram simultaneamente, é razoável imaginar que algumas destas retrocópias sejam mais antigas e que, portanto, já estejam fixadas na espécie *Homo sapiens*. Por outro lado, devem existir outras retrocópias com origem mais recentes

e que ainda não alcançaram a fixação e, portanto, podem ser encontradas como polimórficas (presentes ou ausentes) na população humana.

O polimorfismo de presença e ausência de retrocópias (ou pseudogenes processados) foi descrito pela primeira vez no final da década de oitenta. Anagnou e colaboradores localizaram o pseudogene processado DHFRP1 no genoma humano e descreveram evidências de presença e ausência do *locus* em amostras de noventa indivíduos. Neste trabalho também se investiga a frequência de alelos com este pseudogene processado em cinco populações (Anagnou et al., 1988). Como uma análise piloto para investigar a existência de retrocópias polimórficas, também avaliamos o polimorfismo e genotipagem da retrocópia do gene DHFR (chr18:23,747,811-23,751,321) utilizando dados públicos de sequenciamento de genoma completo. O método consiste, resumidamente, em verificar se há algum alinhamento pareado evidenciando a ausência de *loci* anotados como retrocópias em pelo menos dois indivíduos na população humana. A fim de investigar a frequência alélica destes eventos, também desenvolvemos métodos para genotipar indivíduos do projeto *1000 Genomes* e comparamos os resultados encontrados em 1988 (Tabela 14 e 15).

Tabela 14. Frequência alélica da presença de DHFRP1 em subpopulações humana encontrados no estudo de Anagnou e colaboradores.

Grupo Racial	Porcentagem dos cromossomos com DHFRP1
Mediterrâneos	94.7%
Indianos asiáticos	77.5%
Chineses	67.6%
Asiáticos	57.1%
Americanos Negros	32.5%

Tabela 15. Frequência alélica da presença de DHFRP1 em subpopulações humana encontrados em nossos resultados.

Grupo Racial	Porcentagem dos cromossomos com DHFRP1
Europeus	90.70%
Americanos	81.87%
Asiáticos	74.28%
Africanos	43.10%

Apesar de aumentarmos consideravelmente o número de indivíduos analisados, os resultados encontrados para a retrocópia DHFRP1 são similares aos encontrados por Anagnou e colaboradores. Aproximadamente 90% dos indivíduos com ancestralidade europeia apresentam alelos com a retrocópia DHFRP1. Na outra ponta do espectro de frequência alélica, a minoria dos cromossomos (43.1%) de indivíduos com ancestralidade africana contem a mesma retrocópia do gene DHFR.

A fim de verificarmos quão confiáveis os resultados de genotipagem seriam, realizamos três experimentos. Inicialmente, verificamos a genotipagem de dois trios (mãe, pai e filho) presentes no projeto *1000 Genomes*. Todos os indivíduos foram genotipados como homozigotos presença, portanto, não encontramos desvio que violasse as leis de herança mendeliana. Além disso, utilizamos dados públicos de análises de variação de número de cópia de indivíduos do projeto HapMap (International HapMap Consortium, 2003). A sobreposição de indivíduos do projeto *1000 Genomes* e indivíduos analisados pelo projeto HapMap nos permitiu comparar a genotipagem por dados de sequenciamento (nosso *pipeline*) com os resultados de análise variação de número de cópias por técnicas de *array* de hibridização (Conrad et al., 2010). Ao comparar os resultados de Conrad e colaboradores, encontramos que 100%, 85% e 98% dos indivíduos homozigotos presença, heterozigotos e homozigotos ausência, respectivamente, apresentaram o mesmo genótipo em

ambos estudos. Finalmente, para ganharmos mais confiança nestes resultados, nós validamos experimentalmente os genótipos por PCRs com *primers* flanqueando a retrocópia DHFRP1 e validamos 94.4% dos genótipos identificados pelos métodos *in silico*, sendo que, 100% dos indivíduos homozigotos e 85.3% dos indivíduos heterozigotos foram genotipados corretamente.

Além da frequência alélica da retrocópia DHFRP1, avaliamos também a distribuição de genótipos em cada uma das subpopulações analisadas. A Figura 21, demostra que mais de 75% dos europeus (CEU, FIN, GBR e TSI) analisados são homozigotos para a presença da retrocópia enquanto africanos sub-saharianos (LWK e YRI) apresentam menos de 25% dos indivíduos como homozigotos para a presença da retrocópia. Curiosamente, indivíduos com ancestralidade africana residentes dos Estados Unidos (ASW) apresentam uma frenquência maior de homozigotos presença (40%), similar à frequência de homozigotos em asiáticos e americanos.

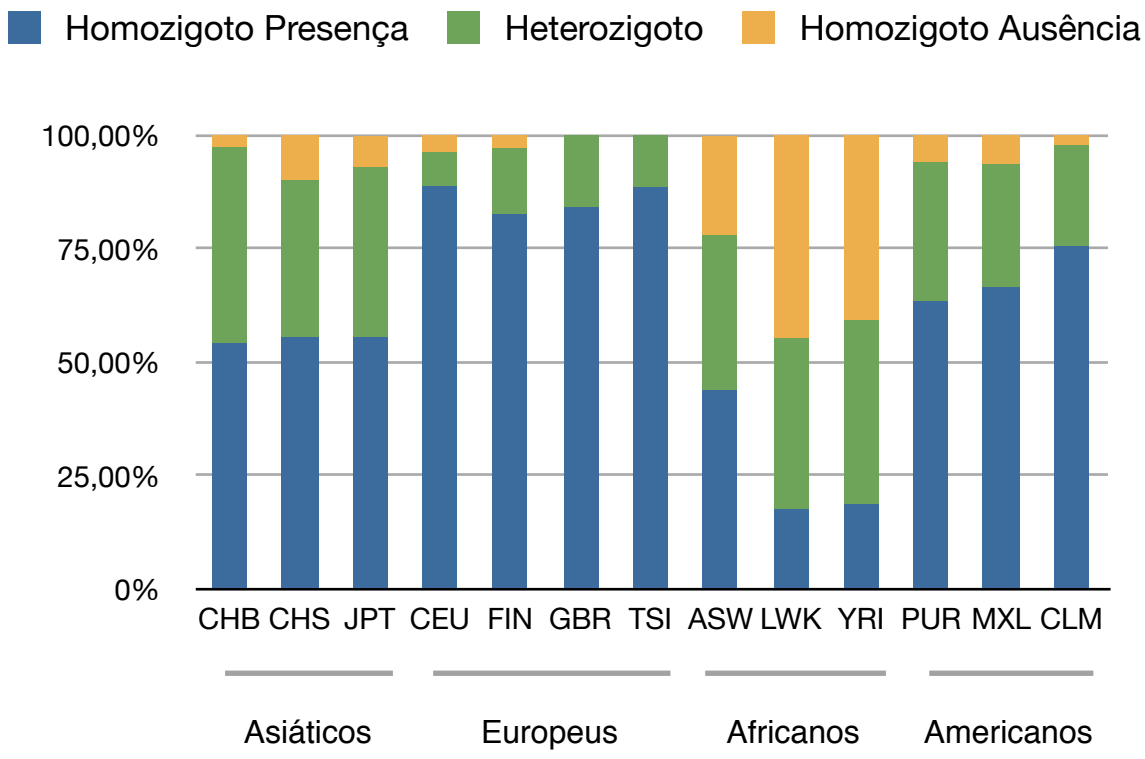


Figura 21. Porcentagem dos genótipos encontrados para a presença da retrocópia DHFRP1 em diversas populações humanas.

Visto que a análise de genotipagem e frequência alélica eram confiáveis e traziam resultados interessantes, expandimos a busca de evidência de ausência em indivíduos do projeto *1000 Genomes* para todas as retrocópias específicas da espécie humana. Dos 127 eventos específicos de humanos, detectamos evidência de ausência para 17 destes (incluindo DHFRP1), dos quais, 10 foram validados experimentalmente em nosso laboratório. Os eventos não validados, em geral, assim não o foram porque a região de inserção não permitia a validação ou por não termos DNA dos indivíduos com evidencia de ausência (Tabela 16). Estes eventos foram chamados de retroCNVs. Retro por serem originados da retroposição de transcritos de genes codificadores de proteínas, e CNV, do Inglês variação de número de cópia (*copy number variation*), para destacar o polimorfismo de presença e ausência e a possibilidade do indivíduo poder apresentar zero (homozigoto ausência), uma (heterozigoto) ou duas (homozigoto presença) cópias da retrocópia.

Tabela 16. Retrocópias presentes no genoma referência humano com ausência de evidência em indivíduos do projeto *1.000 Genomes*.

Nome do gene parental	Cromossomo da inserção	Início da inserção	Final da inserção	Fita de inserção	Contexto	Gene hospedeiro	Fita do hospedeiro
CKS1B	chr5	61807580	61808309	-	Intrônico	IPO11	+
DHFR	chr18	23747811	23751321	-	Intrônico	PSMA8	+
FAM103A1	chr6	166998987	167000150	-	Intrônico	RPS6KA2	-
FAM133B	chr5	60670885	60672859	+	Intrônico	ZSWIM6	+
GCSH	chr1	168024597	168025731	-	Intrônico	DCAF6	+
GNG10	chr11	10292761	10293730	-	Intrônico	SBF2	-
ITGB1	chr19	14732345	14733056	-	Intrônico	EMR3	-
RPL13A	chr10	98510023	98510680	+	Intergênico	-	NA
RPL18A	chr12	104659052	104659669	+	Intrônico	TXNRD1	+
RPL21	chr16	9250199	9250778	-	Intergênico	-	NA
RPL29	chr6	118320091	118320745	+	Intrônico	SLC35F1	+
RPL3	chr14	99439148	99439638	-/+	Intergênico	-	NA

Nome do gene parental	Cromossomo da inserção	Início da inserção	Final da inserção	Fita de inserção	Contexto	Gene hospedeiro	Fita do hospedeiro
RPL36A	chr11	16996261	16996591	-	Intrônico	PLEKHA7	-
RPS26	chr17	43685906	43686369	+	Intergênico	-	NA
RPS26	chr4	114135112	114135576	-	Intrônico	ANK2	+
SDHC	chr17	1760573	1761755	-	Intrônico	RPA1	+
SMS	chr1	160864679	160866356	+	Intergênico	-	NA

Algumas características notáveis podem ser observadas neste conjunto de retroCNVs do genoma humano. Em média, 73.6% do transcrito parental é retrocopiado, o que, em média, resulta em 1051 pares de base. O maior retroCNV contendo 3.510 e o menor 398 pares de bases. As leituras pareadas do projeto *1000 Genomes* tem bibliotecas com fragmentos de aproximadamente 200 nucleotídeos, portanto, essa característica impede a detecção de retroCNVs pequenos, provavelmente, enriquecendo esta análises com falsos negativos de tamanho menor que 400 pares de bases.

Como era esperado, devido a origem recente, os retroCNVs apresentam uma alta identidade com o transcrito que lhe deu origem. Em média, esta identidade é de 99.40%, sendo o caso mais divergente apresentando 98.47% de identidade (RPL29) e, os menos divergentes, com três retroCNVs idênticos aos transcrito do gene parental (RPS26, RPL36A e ITGB1). Aleatoriamente, esperaríamos que cerca de 40% das retrocópias estivessem em regiões intragênicas, no entanto, encontramos uma super-representação de eventos dentro de genes hospedeiros (70.58%). Assim como retrocópias presentes em outros primatas, existe um enriquecimento de genes parentais relacionados com proteínas do ribossomo, 47.05% dos eventos tem como gene parental genes RPS ou RPL.

A frequência alélica dos retroCNVs varia de forma notável. Alguns eventos como SDHC e FAM133B, estão praticamente fixados em humanos. Ambos eventos

estão presentes em todos os indivíduos analisados exceto em um indivíduo mexicano e um indivíduo colombiano, respectivamente. RPL13A, por exemplo, apresenta evidência de ausência apenas em indivíduos com ancestralidade europeia (CEU e TSI). Na outra ponta do espectro de variação alélica, a retrocópia do gene RPL36A, é praticamente ausente em indivíduos com ancestralidade asiática (~10% dos alelos) e muito mais frequente em europeus (~40% dos alelos). De maneira geral estes eventos apresentam uma frequência alélica relativamente alta nos cromossomos analisados, com exceção da retrocópia do gene RPL36A, cujo alelo com retrocópia está em aproximadamente 26% dos cromossomos analisados (Figura 22).

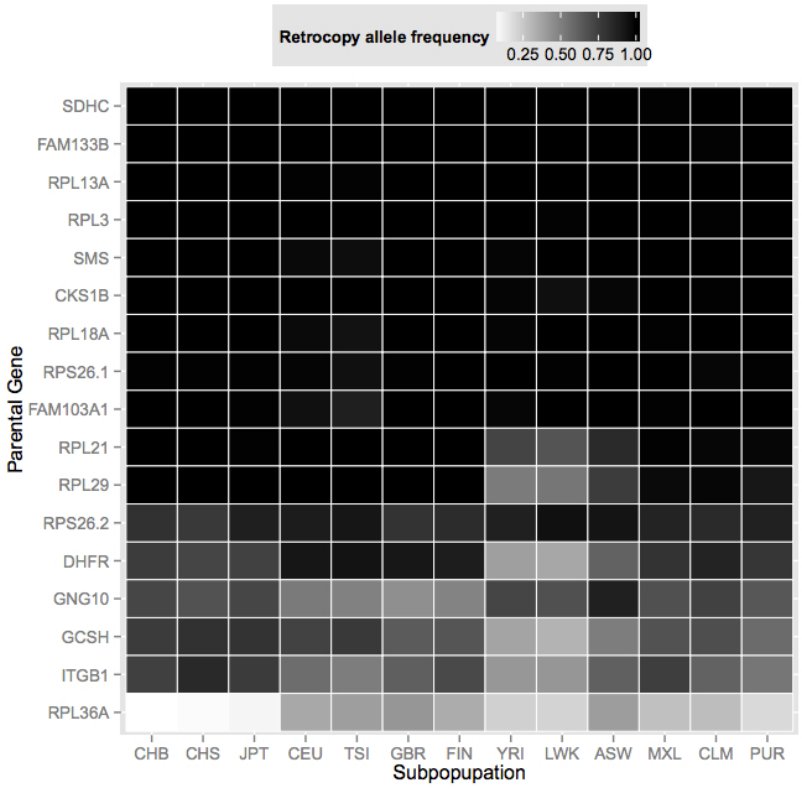


Figura 22. Frequência alélica representada em forma de *heatmap*. Cada linha refere-se a uma retrocópia e cada coluna a uma população. Cada bloco é preenchido com tons de cinza proporcionais a frequência alélica da presença da

retrocópia.

O enriquecimento de retroCNVs com frequência alélica alta nas análises anteriores tem duas possíveis explicações. A origem de novos retroCNVs pode ter sido totalmente interrompida durante a evolução humana e, portanto, somente retroCNVs mais antigos e praticamente fixados foram possíveis de serem detectados. Ou o genoma referência humano representa um grupo restrito de indivíduos, que não seria capaz de representar alelos menos frequentes. Para responder esta questão, desenvolvemos os métodos necessários para detectar retroCNVs ausentes no genoma referência humano, mas presentes em indivíduos do projeto *1000 Genomes*. Como prova de conceito, utilizamos o genoma de vinte indivíduos com maior cobertura de sequência. Para nossa surpresa, encontramos um número relativamente maior de retroCNVs ausentes no genoma referência humano, indicando um possível enriquecimento de retroCNVs de frequência alélica baixa. Em colaboração com Mathew Hahn e Daniel Schrider da Universidade de Indiana (EUA), detectamos evidência de 73 retroCNVs ausentes no genoma referência baseado em junções exon-exon. Dos 73 eventos, fomos capazes de detectar o ponto de inserção de 21 eventos com os métodos desenvolvidos em nosso laboratório (Tabela 17).

Tabela 17. Retrocópias ausentes no genoma referência humano com evidência de presença em indivíduos do projeto *1000 Genomes*.

Nome do gene parental	Cromossomo da inserção	Início da inserção	Final da inserção	Fita de inserção	Contexto	Gene hospedeiro	Fita do hospedeiro
C14orf109	chr3	169729732	169729759	-	Intergênico	-	NA
CACNA1B	chr1	147499917	147500462	?	Intergênico	-	NA
CBX3	chr15	40854166	40854191	-	Intrônico	C15orf57	-
FBXL5	chr13	40620249	40620275	+	Intergênico	-	NA
FOXK2	chr6	159771977	159772000	-	Intergênico	-	NA

Nome do gene parental	Cromossomo da inserção	Início da inserção	Final da inserção	Fita de inserção	Contexto	Gene hospedeiro	Fita do hospedeiro
GAPDH	chr5	56485970	56485994	-	Intrônico	GPBP1	+
HNRNPC	chr6	114017509	114017537	+	Intergênico	-	NA
KRR1	chr10	23199446	23199474	+	Intergênico	-	NA
LAPTM4B	chr6	167333951	167333973	+	Intergênico	-	NA
MATR3	chr12	113886996	113887027	-	Intergênico	-	NA
MFF	chr15	93839688	93839714	-	Intergênico	-	NA
MIER1	chr16	77788935	77788972	-	Intergênico	-	NA
POLR2C	chr2	11405231	11405267	+	Intrônico	ROCK2	-
SKA3	chr11	108585737	108585763	-	Intrônico	DDX10	+
SLMO2	chr3	8869039	8869065	+	Intergênico	-	NA
TMEM126B	chr10	12256152	12256177	-	Intrônico	CDC123	NA
TMEM126B	chrX	121150771	121150796	-	Intergênico	-	+
TMEM66	chr1	191798702	191798728	-	Intergênico	-	NA
TYRO3	chr13	44069808	44069836	-	Intrônico	ENOX1	-
UQCR10	chr1	109650628	109650654	-	Exônico	C1orf194	-
ZNF664	chr2	3931683	3931712	-	Intergênico	-	NA

Diferente dos retroCNVs presentes no genoma referência, retroCNVs ausentes do genoma referência são mais difíceis de serem analisados quanto à identidade e tamanho da região retrocopiada. A evidência de presença se dá apenas pelas bordas dos eventos, portanto, para retroCNVs com frequência alélica maior é possível definir cerca de 500 pares de bases nas extremidades dos eventos, no entanto, como a maioria destes eventos tem frequência alélica relativamente baixa, é praticamente impossível distinguir mutações de erros de sequenciamento ou definir a extremidade exata dos eventos. Entretanto, é possível verificar o contexto genômico em que as retrocópias foram inseridas. Diferente dos retroCNVs presentes no genoma referência, a distribuição de retroCNVs em regiões intragênicas e intergênicas é mais próxima do esperado por inserções aleatórias. Encontramos que 14 eventos (66.66%) estão em regiões intergênicas enquanto sete retrocópias (33.33%) estão inseridas dentro de genes. É notável que um dos eventos, o retroCNV do gene UQCR10, é uma inserção exônica que modifica a região codificadora do gene C1orf194.

Como esperado, a frequência alélica dos retroCNVs ausentes do genoma referência é, em média, menor que retroCNVs presentes no genoma referência. Com exceção do retroCNV do gene CBX3, que está presente em 57% do cromossomos analisados. Já os retroCNVs ausentes do genoma referência apresentam frequência média de 15% dos cromossomos analisados. Também encontramos alguns retroCNVs específicos à certas subpopulações. O retroCNV do gene C14orf109, por exemplo, está presente em aproximadamente em 10% dos genomas de indivíduos com ancestralidade asiática e totalmente ausente em outras subpopulações. Similarmente, o retroCNV TMEM126.2 está presente somente em indivíduos com ancestralidade africana. Descendentes residentes nos Estados Unidos e indivíduos LWK apresentam 10% e 17% dos cromossomos com retrocópia, enquanto indivíduos YRI apenas 1% dos cromossomos contêm a retrocópia (Figura 23).

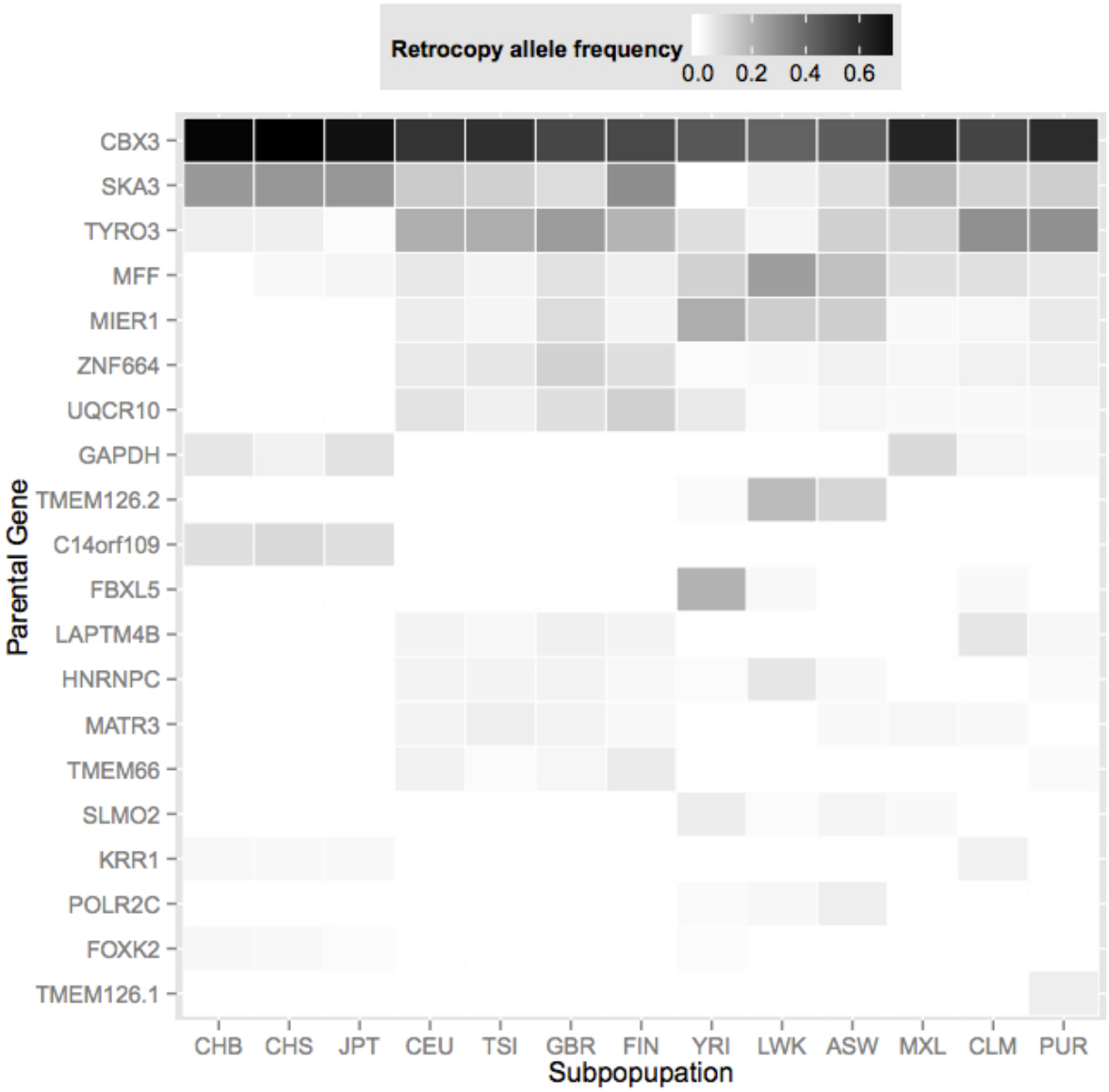


Figura 23. Frequência alélica representada em forma de heat map. Cada linha refere-se a uma retrocópia e cada coluna a uma população. Cada bloco é preenchido com tons de cinza proporcionais a frequência alélica da presença da retrocópia.

4.8. Retrocópias polimórficas somáticas.

Retrocópias são, no geral, subprodutos raros da retroposição autônoma de L1 (Kaessmann et al., 2009), portanto, devido a limitada atividade de elementos

repetitivos em tecidos somáticos (Beck et al., 2010), esperávamos encontrar um número reduzido de retroCNVs somáticos presentes em tumores de colorretais. De fato, entre todas as amostras investigadas, encontramos apenas nove candidatos, com uma média de 1,5 retroCNVs detectável por tumor. Entretanto, o número de candidatos varia de tumor para tumor, por exemplo, não detectamos nenhum candidato na amostra CMCA, enquanto a amostra do paciente AAS apresentou três possíveis eventos de retroposição de mRNA maduro (Tabela 18).

Tabela 18. Possíveis casos de retroCNVs somáticos em tumores colorretais.

ID	Amostra	Região parental	Ponto de inserção	Gene parental	Gene hospedeiro
1	AAS	chr1:155869922-155870136	chr1:222019855-222020078	RIT1	-
9	AAS	chr1:155869741-155870104	chr1:85294293-85294541	RIT1	LPAR3
6	AAS	chr1:104111427-104111981	chr1:58338430-58338901	AMY2B	DAB1
19	AAS	chr2:128394833-128395063	chr8:74814465-74814865	MYO7B	-
4	MM	chr5:110411676-110412399	chr3:7251389-7252217	TSLP	GRM7
7	LIM	chr17:19621584-19625482	chr17:19149720-19152576	SLC47A2	EPN2
16	MDS	chr19:11434176-11435124	chr19:3837334-3838403	RAB3D	ZFR2
17	SKE	chr19:46095625-46095775	chr13:36533317-36533444	GPR4	DCLK1

A fim de validar estes candidatos a retroCNVs somáticos, desenhamos um conjunto de primers flanqueando as possíveis extremidades dos candidatos selecionados, a estratégia geral utilizada está representada por um retroCNV hipotético diagramado na Figura 24A e 24B. Em termos gerais, utilizamos as evidências dos resultados das análises de sequenciamento para estimar as extremidades dos retroCNVs. Idealmente, ambas extremidades dos eventos seriam suportadas por leituras e, desta forma, seria possível amplificar ambas extremidades do retroCNV somático. Estes fragmentos deveriam ser compostos por parte do ponto de inserção e parte do gene parental, sendo que, o fragmento 3' também deveria conter um trato de múltiplas adeninas evidenciando a retroposição do

mRNA (Figura 24B). Finalmente, os fragmentos específicos do tumor, ou seja, sem um fragmento correspondente na amostra de sangue do mesmo paciente, seria sequenciado por SAGER. Todo o processo de validação destes eventos, inclusive o *design* de primers, foi conduzido pela aluna de mestrado Ana Paula de Souza Urlass.

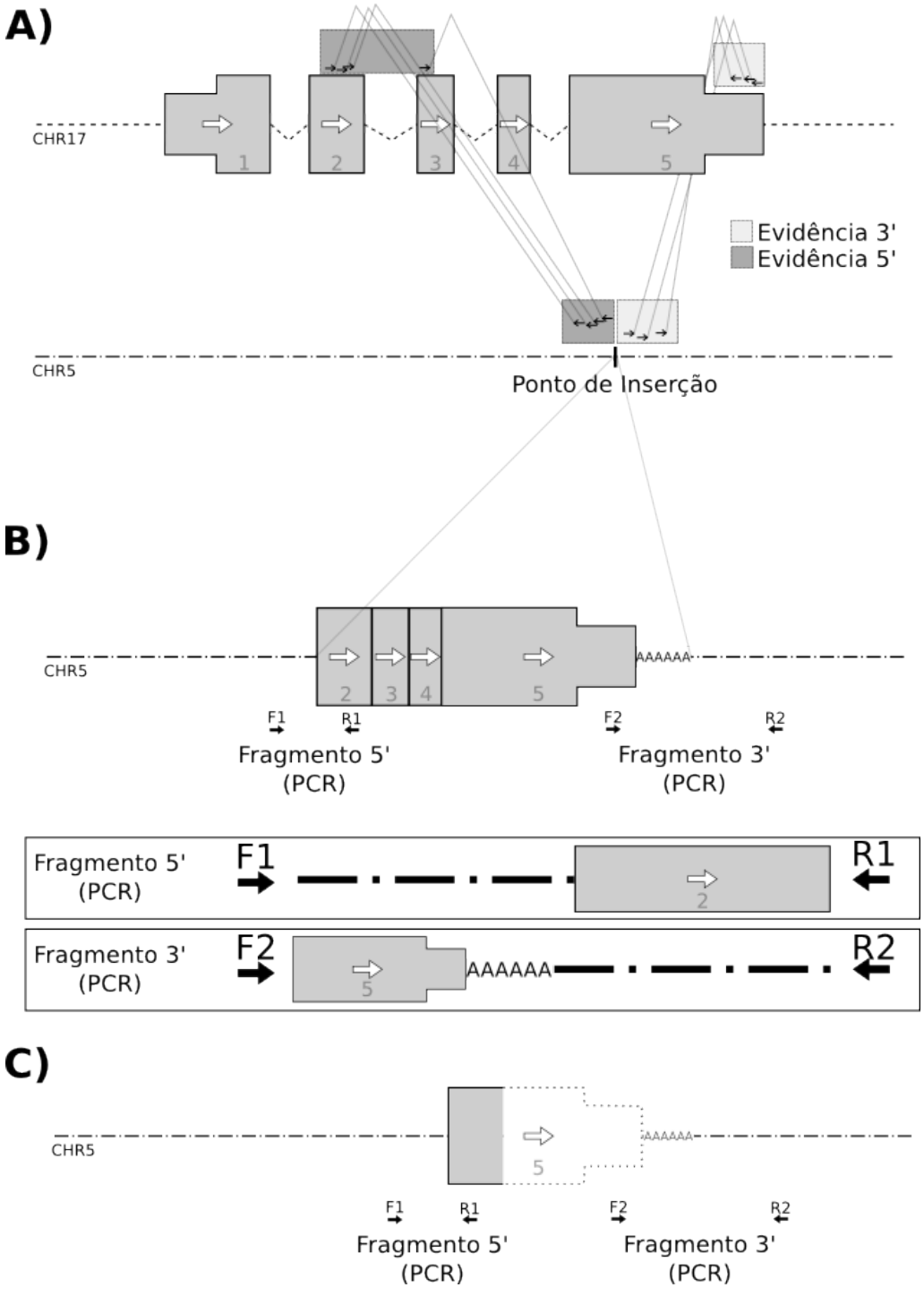


Figura 24. Esquema de detecção e validação de retroCNVs somáticos. **A)** Esquema gráfico dos agrupamentos reportando um novo retroCNV somático. O cromossomo 17 (CHR17) apresenta um gene parental esquematizado por múltiplos exons cuja orientação é dada pela seta branca. Agrupamentos podem ser divididos em "Evidência 3' " (cinza claro) e "Evidência 5' " (cinza escuro) **B)** Representação da sequência esperada de uma inserção parcial como evidenciado em A). **C)** Representação gráfica das regiões validadas dos possíveis retroCNVs somáticos. Apenas a extremidade 5' de todos os eventos foi amplificada e sequenciade, dificultando a confirmação da retroposição de mRNAs maduros.

Ao validarmos praticamente todos os eventos detectados pela versão final de nosso *pipeline* (7/8), com exceção do caso 17 do paciente SKE, que, por experiência prévia do laboratório, apresentou um elevado número de eventos falsos positivos, investigamos qualitativamente as sequências potencialmente retrocopiadas. Curiosamente, percebemos que a maioria dos eventos (1, 9, 6, 19, 4, 16) apresentam evidência de retroposição, por dados de sequenciamento, apenas na extremidade 5' dos eventos. Além disso, ao verificar as regiões com evidência de retroposição, percebemos que todos estes são inserções de um único exon (exon 3'), similar a figura (Figura 24C). Apesar de não haver evidência pelos resultados das análises de bioinformática, seguindo a lógica dos eventos de retroposição (Figura 24B), esperávamos que o restante da porção 3' do gene parental e um trato de múltiplas adeninas completassem o restante do evento no ponto de inserção. Desta forma, desenhamos os *primers* na extremidade 3', estimando o termino do gene parental e o ponto de inserção. No entanto, ao tentarmos validar a extremidade 3'

dos eventos não conseguimos amplificar nenhum fragmento 3' (Figura 24C). A combinação destes resultados com a observação de que não conseguimos validar nenhuma junção exon-exon da sequência parental dificultou a nossa conclusão de que estes eventos são, de fato, retrocópias ou intermediários de uma retroposição. Diversas hipóteses foram levantadas quanto a impossibilidade de validar o fragmento 3' destes eventos, entre elas: i) a qualidade dos primers, implicando em falsos negativos; ii) a presença de rearranjos envolvendo as extremidades 3' de retroCNVs somáticos implicando na fusão entre o ponto de inserção e a região parental, o que impossibilitaria a confirmação de que houve uma retroposição como detectado por nossos dados; iii) o uso de *primers* alternativos durante a transcriptase reversa com *primers* em alvos (Target Primed Reverse Transcription - TPRT), implicando em ausência de um trato poli(A) e uma extremidade alternativa do gene parental. Estas questões ainda estão em aberto e deverão ser reavaliadas nos próximos experimentos.

Além de avaliarmos quantitativamente os possíveis eventos de retroCNVs somáticos, avaliamos qualitativamente a movimentação de sequências potencialmente codificadoras nos genomas tumorais. Analisando com mais cuidado as descrições dos genes envolvidos nos eventos de retroCNVs somáticos, observamos que, pelo menos 4 deles podem estar envolvidos com parte do processo de tumorigênese destas amostras. Por exemplo, o primeiro evento (ID 1 e 9), tem como sequência parental o gene RIT1, codificador uma proteína que regula p38 e, portanto, envolvido com a cascata de sinais da via de MAP-K. A *up-regulação* de RIT1, seja pelo aumento de transcritos RIT1 ou pela regulação indireta de seu gene parental, levaria a *up-regulação* da via de MAP-K, frequentemente ativada em tumores de cólon. O caso 4, duplica parte do gene TSLP que está envolvido com

reposta imune, mais recentemente, descrita como um dos marcos da tumorigênese. Finalmente, o caso 16, gera uma possível retroduplicação do gene RAB3D que faz parte da família RAS de oncogenes, que também é frequentemente encontrada como ativada em tumores de cólon. É certamente curioso que, mais da metade dos eventos de possíveis retroduplicações detectados por este projeto estejam potencialmente relacionados a tumorigênese. A aluna de mestrado Ana Paula de Souza Urlass irá conduzir os experimentos para verificar o nível de expressão e eventuais quimeras dos genes hospedeiros e genes parentais para avaliar possíveis implicações funcionais nestes genes.

4.9. Expressão de retrocópias

Retrocópias podem fugir de seu destino de pseudogene não transcritos quando adquirirem uma região promotora em seu contexto de inserção. O novo contexto pode: i) conter elementos repetitivos capazes de promover a expressão de regiões adjacentes; ii) conter um gene com região promotora bem definida; ou iii) gerar uma região promotora *de novo* a partir da inserção da retrocópia. Para entender o alcance da expressão de retrocópias no genoma humano e de outros primatas, nós aplicamos os métodos previamente descritos em dados públicos de RNA-seq (Brawand et al., 2012) de cinco primatas (humano, chimpanzé, gorila, orangotango e rhesus) e seis tecidos (cérebro, cerebelo, testículo, fígado, coração e rim). Apesar do projeto ENCODE ter aumentado nosso entendimento sobre a estocacidade da expressão gênica, Pei e colaboradores também sugerem que uma fração das regiões expressas, incluindo retrocópias, apresentam função bioquímica (Pei et al., 2012).

No total, nós encontramos a expressão de 3.562 candidatos a retrocópias expressas, sendo 1.304, 1.500, 1.461, 846 e 1.324 candidatos em humano, chimpanzé, gorila, orangotango e rhesus respectivamente (Figura 25A). Com o objetivo de analisar a presença de falsos negativos em nossas análises de expressão, comparamos a expressão de genes parentais e suas respectivas retrocópias. Por exemplo, comparamos a expressão do gene NRBF2 e suas três retrocópias expressas (Figura 25B). É possível observar que o perfil de expressão em diferentes tecidos é diverso e, enquanto o gene parental tem expressão no cérebro, cerebelo, rim e testículo, suas retrocópias tem expressão testículo específico; expressão elevada em cerebelo e expressão ubíqua (Figura 25B). A fim de generalizarmos esta análise, calculamos a correlação entre o nível de expressão de genes parentais e suas retrocópias nos tecidos analisados. Não encontramos uma correlação significativa entre a expressão das retrocópias e seus genes parentais (P=0.46; Spearman=-0.0241, Figura 25C), indicando que o novo contexto de inserção permite que a retrocópia adquira um novo perfil de expressão e uma quantidade diminuta de alinhamentos falsos negativos nas retrocópias.

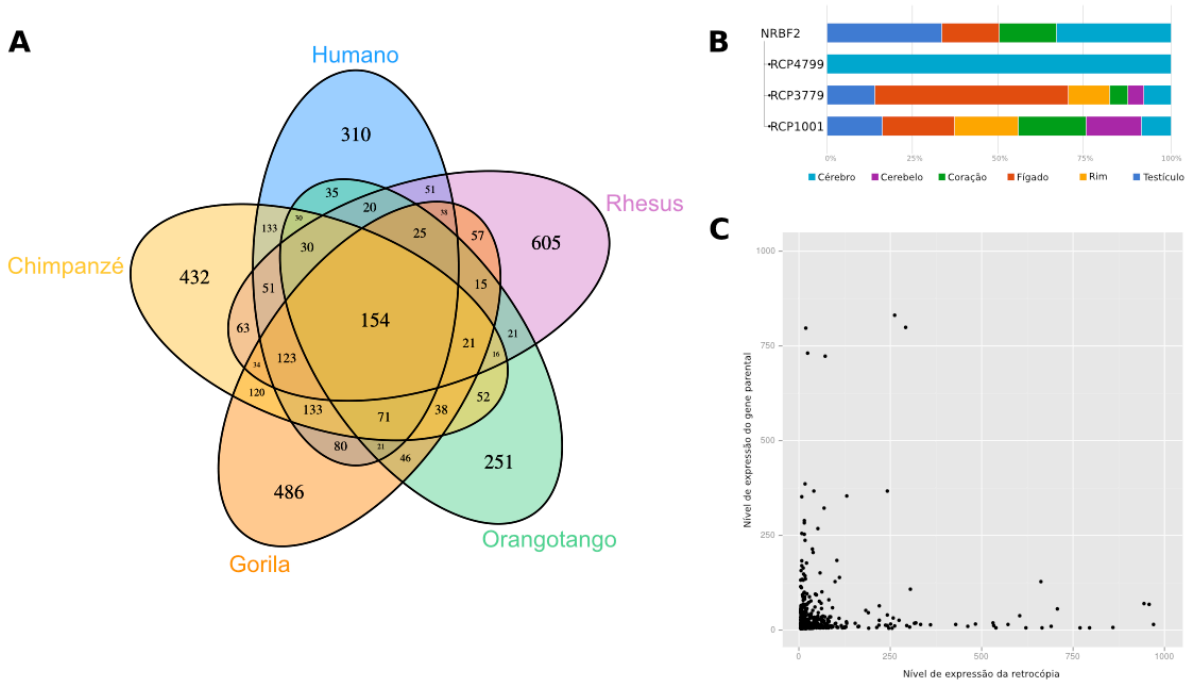


Figura 25. Retrocópias expressas no genoma de primatas. **A.** Diagrama de Veen com o número de retrocópias expressas nos cinco primatas analisados. **B.** Perfil de expressão do gene parental NRBF2 e três retrocópias expressas deste gene. **C.** Correlação entre a expressão de genes parentais e retrocópias nos diversos tecidos analisados.

A fim de entender como estas retrocópias são expressas, nós analisamos a proximidade das retrocópias expressas com regiões potencialmente reguladoras de expressão. Como esperado (Vinckenbosch et al., 2006), encontramos que um número significativo destas retrocópias estão localizadas próximo ou dentro de genes (71%; p-valor<2.2e-16; chi-quadrado=308; d.f.=2, Figura 26).

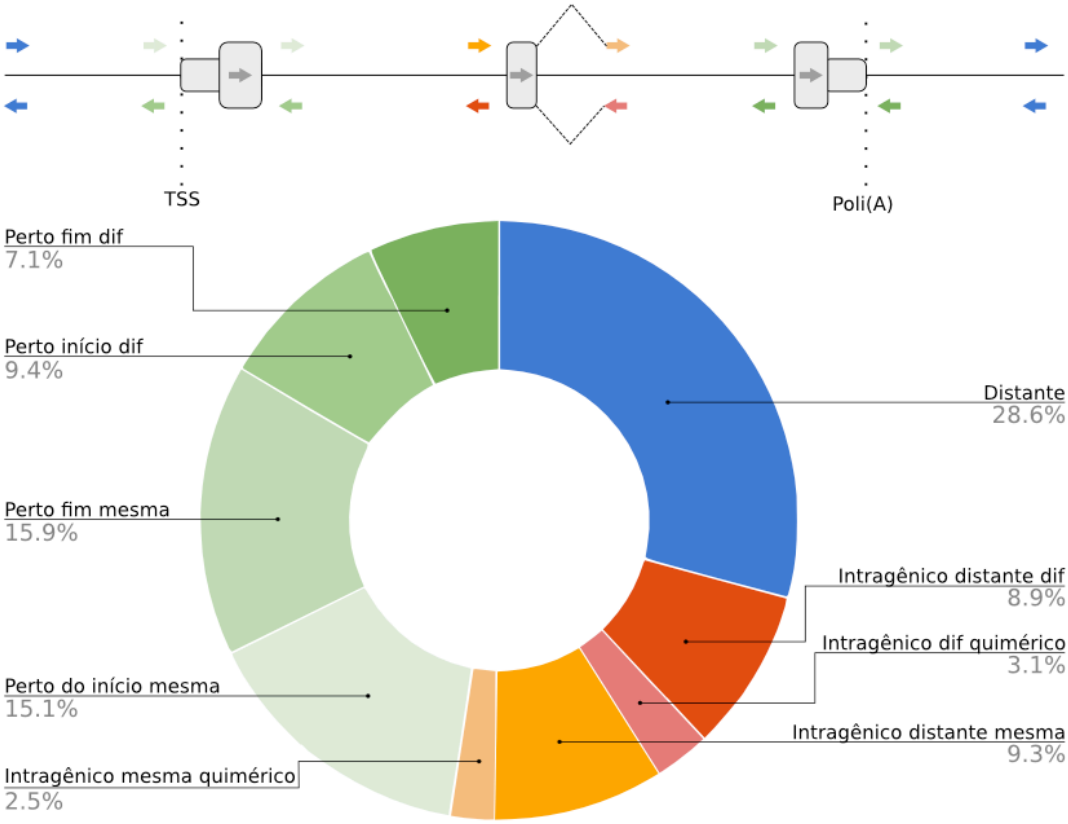


Figura 26. Contexto de retrocópias expressas no genoma humano.

Aparentemente o novo contexto regulatório das retrocópias não teve tempo suficiente para ser selecionado e, portanto, as retrocópias são mais frequentemente expressas em testículo e tecidos nervosos em comparação a tecidos mais especializados como músculo, rim e fígado, como o esperado (Jongeneel et al., 2005). Nós também observamos que retrocópias apresentam uma expressão tecido específica ou são expressas em menos tecidos que seus genes parentais (Figura 27). Por exemplo, encontramos 310, 432, 486, 251 e 605 retrocópias apresentando expressão tecido específico em humanos, chimpanzés, gorilas, orangotangos e rhesus respectivamente.

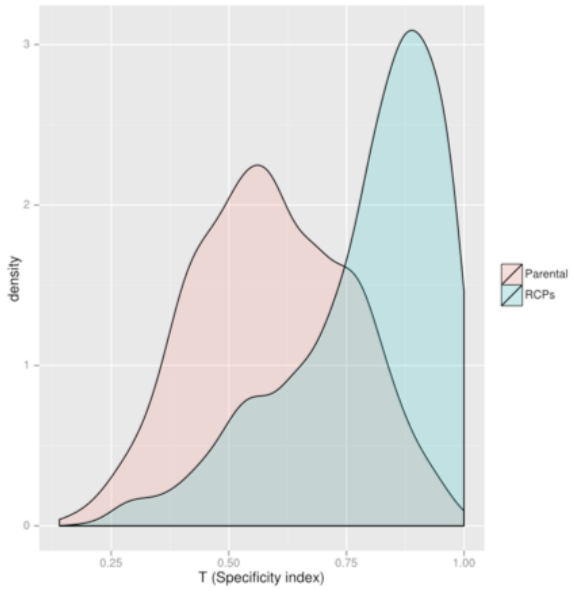


Figura 27. Distribuição do índice de especificidade da expressão de retrocópias e genes parentais.

Dado o viés de retrocópias exportadas do e para o cromossomo X (Emerson et al., 2004), nós também investigamos os vieses do número de retrocópias e retrocópias expressas em cada uma destas subclasses. Assumindo que, como mostramos anteriormente, o número de retrocópias inseridas e fixadas em um cromossomo é proporcional ao tamanho do cromossomo e que o número de retrocópias exportadas de um cromossomo é proporcional ao número de genes no cromossomo, nós calculamos o número esperado de retrocópias exportadas e importadas para o cromossomo X dos cinco primatas com dados de RNA-seq. Assim como Emerson e colaboradores, e como mostramos anteriormente, encontramos um enriquecimento de 26% e 41% no número de retrocópias exportadas e importadas para o cromossomo X no genoma humano. De forma muito similar, todos os primatas, exceto chimpanzé, apresentaram um enriquecimento de retrocópias exportadas e importadas para o cromossomo X. Este perfil é revertido quando

observamos somente as retrocópias expressas. Novamente como Emerson e colaboradores, nós observamos um enriquecimento de retrocópias expressas exportadas do cromossomo X, corroborando com a hipótese de 'desmasculinização' deste cromossomo (Emerson et al., 2004). Curiosamente, apesar de haver um enriquecimento de retrocópias importadas para o cromossomo X, a tendência é revertida quando consideramos retrocópias expressas no cromossomo X. O mesmo acontece quando avaliamos os outros primatas, com exceção a orangotango (Tabela S2).

Uma forma alternativa de exaptação de retrocópias no genoma de eucariotos é a utilização de parte da sequência das retrocópias, seja na orientação senso ou anti-senso, para formação de novos exons ou transcritos alternativos de genes hospedeiros (Baertsch et al., 2008). Para contemplar este tipo neofuncionalização utilizamos os métodos descritos anteriormente para detectar transcritos quiméricos envolvendo sequências anotadas como retrocópias. Mais uma vez, analisamos a presença de transcritos alternativos em dados de RNA-seq (Brawand et al., 2012) em seis tecidos. Analisando as retrocópias com maior suporte de transcritos quiméricos, encontramos quatro retrocópias com evidência de *splicing* alternativo sem envolver um gene hospedeiro. Duas destas retrocópias que geram transcritos alternativos são anotadas como TAF9 e MORF4L2 e são duplicações dos genes TAF9B e MORF4L1 (Tabela 19). Estas quatro retrocópias, apesar de não envolverem um gene hospedeiro, foram descritas como quiméricas, pois apresentam novos exons externos ao *locus* anotado como retrocópia. Todos os casos adicionais são casos de exonificação, ou seja, casos em que uma pequena porção da retrocópia é exaptada como exon de um gene hospedeiro. Dos seis casos restantes, quatro geram transcritos no sentido contrário do gene parental, portanto, sem

qualquer semelhança a função prévia e dois casos geram exons no mesmo sentido do gene parental, um deles é utilizado como último exon alternativo, praticamente inteiro como 3‘UTR e o segundo é um novo exon ainda não descrito na literatura do gene Transferrina que codifica um dos principais transportadores de ferro em mamíferos.

Tabela 19. Retrocópias com evidência de expressão quimérica.

Hospedeiro	Parental	Classe	Suporte	Tecido	Fita
MORF4L2	MORF4L1	Novo gene	499	Rim	
TAF9	TAF9B	Novo gene	435	Testículo e Cérebro	
FAM82B	SLC2A3	Exonificação	160	Coração, Rim, Fígado e Testículos	Oposta
CPSF4	SARNP	Exonificação	252	Todos	Oposta
CHSY1	CHSY3	Novo Gene	104	Todos	
TPT1-AS1	RCN1	Novo gene	98	Todos	Oposta
mir6080	SNRNP200	Exonificação	96	Todos	Oposta
FMO4	TOP1	Exonificação (3‘UTR)	71	Rim e Fígado	
SCP2	RASS2	Exonificação (3‘UTR)	65		Oposta
TF	ACSL3	Exonificação	52	Fígado e Cérebro	

Além de verificar quais retrocópias apresentavam evidência de expressão quimérica, com a análise de ortologia de retrocópias em primatas, pudemos verificar se alguma retrocópia presente apenas no genoma humano apresentava evidência de expressão. Estes eventos são especialmente interessantes, pois podem representar *loci* responsáveis por fenótipos específicos da espécie humana. A

detecção da expressão destes eventos é ainda mais delicada, pois, por se tratarem de eventos recentes, as sequências das retrocópias são muito similares às sequências dos transcritos parentais. Portanto, os resultados de expressão foram manualmente curados para garantir a ausência de falsos positivos. Mesmo assim, não foi possível excluir totalmente a possibilidade de existirem alguns falsos negativos. No total, foram identificados sete retrocópias (0.05%) específicas de humanos com ao menos três leituras específicas na região anotada como retrocópia. A maioria das retrocópias humano específicas são intergênicas e, portanto, devem utilizar promotores de outros genes ou promotores de elementos repetitivos. De fato, todos os eventos detectados são adjacentes a elementos repetitivos. Quatro inserções não apenas apresentam elementos repetitivos adjacentes, como ocorreram dentro de elementos repetitivos (Tabela 20).

Tabela 20. Retrocópias humano específicas com evidência de expressão *perse*.

Hospedeiro	Parental	Classe	Elemento repetitivo proximal
-	PHC1	-	L1PA14
-	NUDT4	-	L1MA4
-	RPS28	-	MER5A1
ZNF286B	FOXO3	Não codificador	Charlie
-	RSP2	-	L1M4
-	PRR13	-	LTR15
BBS5	C14orf126	3‘UTR	SVA

Apesar de nenhuma retrocópia humano específica expressa ser polimórfica, podemos verificar evidência de expressão de novas inserções ausentes no genoma

referência. Para tal, utilizamos dados publicamente disponíveis de sequenciamentos de transcriptoma de indivíduos saudáveis (www.geuvadis.org). Entretanto, esta análise apresenta diversas limitações. Como não temos a sequência completa da retrocopiada e, nas regiões detectadas, não há diferença entre a retrocópia e o gene parental, confiamos apenas na evidência de expressão quimérica destes eventos. Além disso, apesar do projeto Geuvadis (www.geuvadis.org) ter sobreposição com o indivíduos do projeto *1000 Genomes*, apenas linhagens celulares linfoblastóides tiveram seu transcriptoma sequenciado. Desta maneira, avaliamos a expressão de novas inserções apenas neste tecido. A única retrocópia com evidência de expressão detectada foi a retrocópia do gene CBX3 no cromossomo 15 dentro do gene hospedeiro C15orf57. Em camundongos este gene está anotado como Ccdc32 e foi descrito como uma das proteínas que interagem com o gene anexina2 (Li, Q. et al., 2011). O diagrama na Figura 28 representa as leituras quiméricas encontradas e sítios canônicos de *splicing*. Infelizmente não foi possível definir o final da inserção e, portanto, não foi possível identificar a sequência completa do novo transcrito nem estimar o seu nível de transcrição.

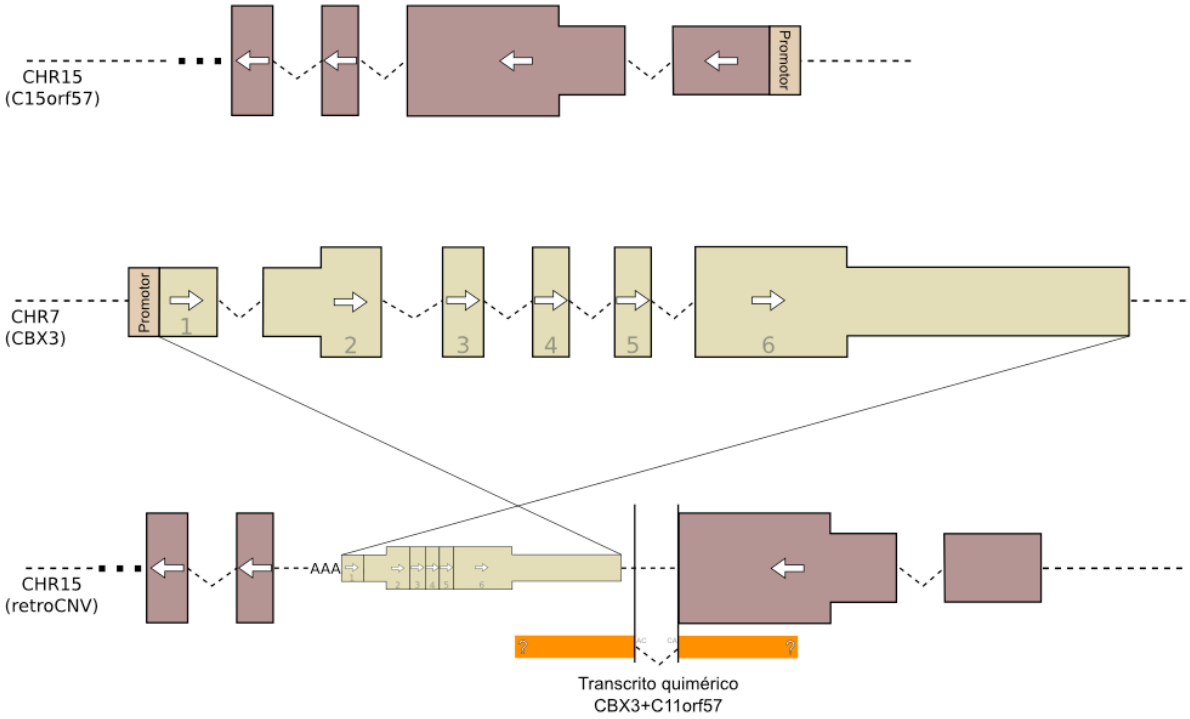


Figura 28. Diagrama representando a evidência de expressão quimérica de um gene hospedeiro (C15orf57) e um retroCNV (CBX3) ausente do genoma referência humano.

Capítulo 5.
Discussão

“Os que o imaginam sem limites esquecem que

não é ilimitado o número possível de livros”

Jorge Luis Borges - Ficções

5.1. Retrocópias no genoma humano

Retrocópias são definidas como sequências de DNA originadas pela retroposição de mRNAs de genes (codificadores de proteína ou não) e não de elementos repetitivos. Assim como o número de genes (Gerstein et al., 2007), o número exato de retrocópias no genoma de humano ainda é uma questão em aberto (Baertsch et al., 2008 e Marques et al., 2005 e Ohshima et al., 2003 e Pei et al., 2012). A detecção e enumeração de retrocópias dependem, fundamentalmente, de quatro fatores. i) A qualidade do genoma referência; ii) o conjunto de genes descritos no genoma do organismo; iii) a qualidade da sequência dos transcritos ou proteínas dos genes descritos; e iv) o método utilizado para detectar os sinais moleculares que permitem a identificação de retrocópias. Desde a publicação do genoma referência humano (Lander et al., 2001 e Venter et al., 2001), diversas atualizações foram disponibilizadas, em 2003 (hg16), 2004 (hg17), 2006 (hg18), 2009 (hg19) e, finalmente, 2014 (hg38), incrementando de maneira significativa a qualidade de montagem de todos os cromossomos humanos. O transcriptoma humano tem atualizações periódicas (Benson et al., 2013), porém ainda mais frequentes. Portanto, a caracterização do genoma e do transcriptoma humano vem sendo refinada há mais de uma década e, espera-se, que as anotações de retrocópias acompanhem esta tendência.

Devido à crença que retrocópias não teriam um papel biológico, o estudo destes eventos acompanhou a descrição de genes no genoma humano, apesar de ocupar um grau menor de importância. Inicialmente, um número reduzido de retrocópias foi independentemente descrito como casos inesperados quando se tentava isolar as sequências dos primeiros genes codificadores de proteínas (Nishioka et al., 1980 e Vanin et al., 1980). Posteriormente, baseando-se nos

cromossomos 21 e 22, surgiram as primeiras estimativas fundamentadas para o número de genes no genoma humano, e, em paralelo, baseado em poucas centenas de retrocópias detectadas nestes cromossomos, surgiram as primeiras tentativas de quantificar o número de retrocópias no genoma humano (Dunham et al., 1999 e Harrison et al., 2002 e Hattori et al., 2000). Com a publicação do genoma referência humano (Lander et al., 2001 e Venter et al., 2001), diversos trabalhos basearam-se no alinhamento de sequências proteicas e detecção de junções de exons, encontrando de dois a oito mil pseudogenes processados (potencialmente retrocópias). Com o desenvolvimento de novas tecnologias de sequenciamento e a busca exaustiva por transcritos de genes codificadores de proteínas, estabeleceu-se o conjunto mais confiável de transcritos do genoma humano, o RefSeq. Este novo recurso possibilitou o desenvolvimento de novos métodos, baseados na sequência de transcritos, para detecção de retrocópias e pseudogenes processados. Nesta nova etapa, o número de retrocópias aumentou consideravelmente, variando entre oito e treze mil eventos (Baertsch et al., 2008 e Pei et al., 2012). Nós descrevemos 7.831 retrocópias no genoma humano. Comparado aos primeiros resultados de análise completa do genoma humano (Ohshima et al., 2003 e Venter et al., 2001) e alguns poucos trabalhos recentes (Zhang, Q., 2013), encontramos um número elevado de retrocópias. Porém, quando comparamos o número de retrocópias que encontramos contra a maioria dos trabalhos recentes, encontramos um número similar (Pei et al., 2012) ou até mesmo menor (Baertsch et al., 2008). Entendemos que, neste momento, é mais importante selecionar um conjunto confiável e representativo de retrocópias, do que definir o número definitivo de retrocópias no genoma humano. Este conjunto representativo poderá então ser utilizado para

entender e avaliar quão variáveis e quais os impactos funcionais que estas duplicatas gênicas tem na biologia humana e de outros primatas.

5.2. Método de detecção de retrocópias

Este cenário é ainda mais crítico quando são consideradas outras espécies como de roedores e primatas não humanos. Baseado nesta limitação, desenvolvemos métodos e programas necessários para investigar a variação do conjunto de retrocópias em organismos com genoma referência e conjunto de transcritos publicamente disponível. Diferente da maioria dos métodos desenvolvidos até então, optamos pelo uso de sequências de mRNA maduro, ao invés de sequências proteicas, como base para o alinhamento no genoma referência e procura por eventos de retroposição. Esta escolha foi feita por diversos fatores. Primeiro, visto que a reação de transcriptase reversa se inicia pela extremidade 3' do RNA molde e tem processividade relativamente baixa, é provável que, em muitos casos, apenas regiões não traduzidas (3'UTR) sejam retrocopiadas. Portanto, boa parte das retrocópias compostas apenas por sequências não traduzidas seriam perdidas por métodos que usam sequências proteicas para detectar retrocópias.

Sabe-se que existem cerca de 600.000 cópias de elementos L1 no genoma humano. A maioria destas cópias apresenta truncamentos drásticos na porção 5' e, quando comparadas com a sequência consenso dos elementos L1, apresentam uma tendência a serem relativamente curtas (mediana 422 pares de bases). Como visto anteriormente, os métodos para detecção de retrocópias baseiam-se principalmente na procura de junções exon-exon. No entanto, o último exon de genes codificadores de proteínas é longo e tem, em média, 1.325 pares de bases (Scherer, 2008), portanto, espera-se que parte das retrocópias não tenham junções exon-exon.

Sobretudo, parte das retrocópias pequenas deve estar sub-representada em todos os métodos publicados até hoje e estas continuam elusivas e disponíveis para serem descritas, enumeradas e estudadas quanto ao seu impacto e variação no genoma humano e no genoma de outros primatas.

Pode-se criticar também o uso de RNA mensageiro para a detecção de retrocópias devido ao acumulo de mutações, que, em geral, afetam mais a similaridade de transcritos do que de proteínas que são mascaradas pelo código genético ser degenerado. Entretanto, como praticamente todas as retrocópias no genoma humano surgiram após a divergência de primatas e roedores, é possível afirmar que praticamente todas as retrocópias surgiram, no máximo, nos últimos 120 a 90 milhões de anos. Assumindo uma taxa de mutação constante de $1x10^{-9}$ mutações por ano em primatas (Scally; Durbin, 2012) e $2.6x10^{-9}$ (Kumar; Subramanian, 2002) em roedores, poderíamos calcular que sequências neutras que houvessem surgido próximo da divergência entre primatas e roedores teriam, no pior dos casos, aproximadamente 68.8% ($1-0.0000000026x.10^{-9}*120x10^{6}$) de identidade e entre primatas a identidade seria de aproximadamente 94% ($1-1x10^{-9}*60x10^{6}$). Portanto, mesmo considerando os piores casos, o alinhamento de mRNA no genoma referência dos organismos estudados deve ser suficiente para detectar retrocópias originadas há aproximadamente 100 milhões de anos e, portanto, compartilhadas entre mamíferos.

Uma terceira crítica ao desenvolvimento de um *pipeline* específico para a detecção de retrocópias em genomas referência seria a existência de métodos publicados (Baertsch et al., 2008 e Karro et al., 2007) e bancos públicos de retrocópias (Karro et al., 2007 e Khelifi et al., 2005 e Pei et al., 2012). Para justificar nossa decisão de reimplementar este *pipeline*, avaliamos manualmente dois bancos

de dados (pseudogene.org e GENCODE v16) que são utilizados como catálogos referência de pseudogenes processados. Encontramos uma baixa sobreposição (67%) entre ambos os bancos de dados indicando um possível enriquecimento de falso positivos e falsos negativos em ambos. Entretanto, quando comparamos as retrocópias do GENCODE com nossos resultados, encontramos que 87% das retrocópias detectadas pelo nosso *pipeline* também estavam no GENCODE. A fim de analisarmos as retrocópias e pseudogenes processados específicos de cada *pipeline*, comparamos um pequeno conjunto aleatório de eventos e encontramos algumas inconsistências no banco do GENCODE. Por exemplo, encontramos sequências exportadas do DNA mitocondrial, pseudogenes processados de 47 mil pares de bases, isto é, contendo introns e eventos resultantes de duplicação genômica de retrocópias. Em contraste, não encontramos nenhum falso positivo em nosso banco. Em conjunto, estes resultados deixam claro a complexidade de identificar e anotar retrocópias em genomas complexos. Quando consideramos a dimensão do genoma humano e a enganosa simplicidade de seu código, é razoável supor que pequenas sutilezas nos parâmetros dos *pipelines* de detecção de retrocópias possam ser responsáveis por um drástico aumento de falsos positivos e falsos negativos. Mesmo que os parâmetros e métodos ideais sejam encontrados, é quase que impossível chegar a uma lista de retrocópias contendo todos os verdadeiros positivos e nenhum falso negativo. Este argumento tem como base o conjunto de genes de genoma humano. Uma breve busca pela literatura para identificar trabalhos que questionam a definição de gene (Gerstein et al., 2007 e Harrow et al., 2012), ou definição de DNA funcional (Kellis et al., 2014) que são definições fundamentais para o entendimento da biologia do ser humano. Desta forma, entendemos que discussões sobre os melhores métodos e estratégias para

enumeração retrocópias no genoma humano estarão sob a luz do balanço entre sensibilidade e especificidade e antes de avaliar os valores de falsos negativos e falsos positivos encontrados, faz-se necessário considerar o objetivo por trás do método desenvolvido.

5.3. Retrocópias no genoma de outros primatas.

A literatura de retrocópias e pseudogenes processados é principalmente focada no genoma referência de humanos, camundongos e linhagens de Drosophila. Apesar de existirem algumas publicações estudando o número de retrocópias em outros primatas, as atualizações do genoma referência destes organismos e a ausência total de estudos de retrocópias em outros primatas como gorila, orangotango e primatas do novo mundo nos estimularam a gerar um catálogo de retrocópias no genoma destes organismos e compara-las com organismos mais bem estudados. Encontramos um número de retrocópias muito similar em primatas do velho mundo (Catarrhini). Os genomas de humanos, chimpanzés, gorilas, orangotangos e rhesus apresentam aproximadamente 7.500 retrocópias de aproximadamente 2.500 genes parentais (Tabela 3 e Tabela 8). As variações encontradas para cada organismo podem, em geral, ser atribuídas à qualidade do genoma referência e transcriptoma da espécie, ou ainda, e talvez mais interessante, à retrocópias específicas de cada linhagem ou espécie. Para nossa surpresa encontramos cerca de 10.000 retrocópias nos genomas referência de primatas do novo mundo (Platyrrhini), um aumento de aproximadamente 50% quando comparado ao número de retrocópias descritas nos genomas de Catarrhinis.

A fim de entendermos melhor a super-representação de retrocópias em primatas do novo mundo, comparamos os genomas referência destes organismos

quanto ao tamanho, número de genes, número de transcritos, composição do genoma referência anotada como LINE ou SINE. Curiosamente, não encontramos nenhuma diferença significativa (Tabela 7) entre estas características. Assumindo que a resposta para o maior número de retrocópias deve ser consequência direta da maior atividade de elementos L1, aprofundamos as análises de elementos repetitivos e comparamos, entre as subfamílias mais frequentes de L1, quais apresentavam uma diferença significativa, quando comparadas às subfamílias de L1 no genoma de Catarrhinis. Encontramos duas subfamílias super-representadas no genoma de Platyrrhini. Enquanto L1PA7 e L1P3 correspondem a 5 e 1% respectivamente dos elementos L1 mais frequentes no genoma de Catarrhini, em Platyrrhini estes elementos correspondem a 25 e 5% respectivamente (Figura 18A). Para investigarmos a possível expansão de L1PA7 e L1P3, realizamos um alinhamento múltiplo de todas as ORF2p de todos L1PA7 no genoma dos primatas analisados (Figura 18B). Encontramos que, enquanto parte dos L1PA7 em Platyrrhini são similares a L1PA7 em Catarrhini e, portanto ancestral à divergência de ambos os grupos, cerca de 25% dos elementos L1 anotados como L1PA7 em Platyrrhini agrupam apenas entre si, indicando uma possível expansão específica no ramo de primatas do novo mundo. De acordo com nossa hipótese de que o maior número de retrocópias em Platyrrhini pode ser, ao menos parcialmente, explicada pela expansão L1PA7, Ohshima e colaboradores encontraram que parte das retrocópias no genoma humano tem substituições não sinônimas equiparáveis a elementos L1PA7 e, portanto, esta subfamília de elementos L1 seria uma das responsáveis pela explosão de retrocópias no genoma de primatas (Ohshima et al., 2003).

5.4. Retrocópias ortólogas entre primatas e roedores

Aproveitando o benefício de ter acesso ao genoma referência de camundongos (Mouse Genome Sequencing Consortium et al., 2002) e ratos (Gibbs et al., 2004) nós também identificamos as retrocópias compartilhadas entre primatas e roedores. Encontramos que mais de 90% das retrocópias de humanos não estão presentes no genoma de roedores e, portanto, surgiram após a divergência entre estas linhagens. Portanto o nosso achado fortalece diversos trabalhos (Marques et al., 2005 e Ohshima et al., 2003 e Zhang, Z. et al., 2004) que sugerem uma explosão na formação de retrocópias (e Alus) há aproximadamente 40-50 milhões de anos atrás, ou seja, em um ancestral comum a todos os primatas que analisamos. É intrigante imaginar que praticamente todas as retrocópias (99%) no genoma humano (e outros primatas), não tem um análogo em roedores, de sorte que, se algumas destas retrocópias forem funcionais, estas funções, ou não existem em camundongos (e vice versa) ou foram selecionadas de forma independentemente.

Adicionalmente, nós identificamos apenas 63 retrocópias compartilhadas entre primatas e roedores. A maioria destes eventos parecem ser funcionais: elas são transcritas, apresentam evidência de codificação de proteínas e parecem estar sobre seleção purificadora. Destes eventos, 42% retrocópias estão relacionados com o cromossomo X. Quatorze genes foram exportados do cromossomo X para outros cromossomos autossomos, gerando novas cópias não relacionadas aos cromossomos sexuais. Segundo Emerson e colaboradores (Emerson et al., 2004), a super-representação de genes exportados para autossomos pode ser explicada de duas formas: i) Mecanisticamente, quando há um viés para gerar retrocópias de genes expressos no cromossomo X, ou ii) Por seleção natural. Uma forma de seleção, o antagonismo sexual, prediz que variantes que beneficiem machos

e fêmeas devem acumular, respectivamente, em autossomos e no cromossomo X (Ellegren; Parsch, 2007). Os resultados encontrados por nós e Emerson e colaboradores indicam que há um viés para geração de retrocópias em autossomos com genes parentais em cromossomos sexuais, portanto, é possível que estas variantes beneficiem machos em detrimento de fêmeas. Adicionalmente, este viés pode ser explicado pelo fato de vários genes do cromossomo X serem silenciados durante a meiose em machos (Turner, 2007) e, portanto, cópias autossômicas destes genes seriam mais eficientes em machos. O mesmo vale para genes importados ao cromossomo X. Treze retrocópias das 63 compartilhadas entre humanos e camundongos foram importadas para o cromossomo X. Emerson e colaboradores também investigam este viés e apresentam duas explicações: i) um viés mecanístico, onde mais retrocópias seriam inseridas no cromossomo X que em autossomos é investigado e encontra-se que nem toda super-representação pode ser explicada pelo viés mecanicista, apesar de já terem sido descritos vieses semelhantes para elementos repetitivos como LINEs, ou ii) seria resultado de um viés de seleção natural. Novamente, o modelo de seleção por antagonismo sexual pode influenciar o desvio de retrocópias importadas ao cromossomo X. Este modelo também prediz que variantes no cromossomo X que beneficiem fêmeas em detrimento de machos estão presentes em dois terços dos cromossomos na espécie e, portanto, pode ser mais selecionado positivamente (em fêmeas) que negativamente (em machos).

O fato de encontrarmos que 78% das retrocópias compartilhadas entre primatas e roedores são anotadas como genes codificadores de proteína sugere que estes *loci* surgiram ou não como pseudogenes, adquiriram a capacidade de serem transcritos, adquiriram nova ou mantiveram a função codificante do seu parental e

sofreram seleção natural. É possível imaginar que mesmo os 12 eventos não anotados como codificantes possam i) ser codificante, porém ainda não tiveram suas proteínas descritas ou ii) apresentem função não codificadora. Sobretudo, assumindo que o número de retrocópias antes da divergência de humanos e camundongos era pequeno (devido a baixa atividade de LINEs), podemos especular que várias retrocópias, hoje tidas como pseudogenes processados e fósseis de transcritos, serão futuramente utilizadas como substrato para seleção e compõem um reservatório de possibilidades para futuras especiações.

5.5. Retrocópias compartilhadas entre primatas

É comum utilizar o número de mutações sinônimas, ou mutações neutras, em sequências codificadoras de proteínas para estimar qual o tempo decorrido de uma duplicação genômica ou de um gene codificador de proteína (Marques et al., 2005 e Ohshima et al., 2003 e Zhang, Q., 2013). Entretanto, as estimativas de idade são evidências indiretas e análises de ortologia podem ser utilizadas para coletar evidências diretas do compartilhamento de sequências entre diferentes espécies, possibilitando, portanto, uma precisão maior em relação a estas informações. Ao aplicarmos o nosso método, encontramos cerca de 5.700 retrocópias compartilhadas entre todos os genomas de Catarrhini e aproximadamente 4.100 entre Platyrrhinis e Catarrhinis (Figura 20). Portanto, seguindo a linhagem que deu origem a nossa espécie, nos doze milhões de anos que separam a última espécie comum entre primatas do novo e do velho mundo e o primeiro macaco do velho mundo presente em nossos dados (rhesus), cerca de 1.700 retrocópias foram criadas e hoje estão presentes em todos indivíduos Catarrhinis, incluindo os humanos. Estas retrocópias podem ser consideradas oportunidades para criação de versões alternativas de

genes funcionais, novas cópias neofuncionais ou ainda, duplicatas gênicas com perfil distinto de expressão.

Analisando o conjunto de retrocópias surgidos a cada período da evolução de primatas, realizamos uma estimativa da taxa média de criação e fixação de retrocópias em cada um destes períodos (Tabela 13). No geral, encontramos que a taxa de criação de retrocópias por milhão de anos iniciou-se alta nos primatas ancestrais (142 retrocópias por milhão de ano) e decaiu bruscamente até o ancestral comum entre humanos, chimpanzés e gorilas (45 retrocópias por milhão de ano) e decai novamente na linhagem de humanos para 21 retrocópias por milhão de ano. Ainda não há informação suficiente na literatura para entendermos completamente quais as razões deste decaimento. Entretanto, podemos fazer algumas especulações. Por exemplo, sabemos que a atividade de elementos LINE1 diminui nos últimos milhões de anos na linhagem dos primatas (Konkel et al., 2011). Provavelmente, devido aos diversos mecanismos de restrição a atividade de elementos repetitivos, tal como a amplificação da família APOBEC3 (Muckenfuss et al., 2006), ou atividade de PIWIs (Kuramochi-Miyagawa et al., 2008 e Marchetto et al., 2013) ou siRNAs(Watanabe et al., 2008). Sobretudo, é interessante imaginar que independente da taxa de retrocópias criadas e fixadas a cada milhão de ano, durante toda a evolução da linhagem de primatas, retrocópias foram, um dos fatores para geração de variabilidade genética.

5.6. Retrocópias espécie específicas

A análise de ortologia de retrocópias no genoma de primatas permitiu não somente a descrição de retrocópias mais antigas, e, portanto compartilhadas entre todos os primatas, mas também a identificação de retrocópias específicas às

espécies analisadas. Diferente de humanos, chimpanzés e gorilas que, por ter um último ancestral comum relativamente recente (aproximadamente oito milhões de anos), orangotango, rhesus e sagui não tem uma segunda espécie mais próxima com genoma referência publicado e, portanto, podem apresentar um conjunto maior de candidatos falso positivos para as retrocópias específicas de cada espécie. Uma vez que o genoma referência de espécies mais próximas como, por exemplo, *Pongo pigmaeus,* aproximadamente 2.3 milhões de anos do último ancestral comum com orangotango (Zhang, Y. et al., 2001) e *Chlorocebus aethiops,* aproximadamente 9.9 milhões de anos do último ancestral comum rhesus (Steiper; Young, 2006), sejam publicados, seria interessante verificar quantas e quais retrocópias são, de fato, específicas. Entretanto, para as 127, 228 e 212 retrocópias específicas de humanos, chimpanzés e gorilas, devem haver poucos falso positivos.

A fim de entender melhor uma possível contribuição destas retrocópias espécie específicas como recursos para adaptação e especiação buscamos na literatura, exemplos de retrocópias humano específicas com evidência de funcionalização. As retrocópias NANOGP8 e CSNK2A3 estão associadas ao desenvolvimento de tumores e são exemplos de mudança do contexto de expressão (Fairbanks et al., 2012) (Hung et al., 2010).

Exemplos de possíveis alterações em relação aos genes próximos aos pontos de inserção também foram observados entre estas retrocópias espécie-específicas. Por exemplo, a retrocópia do gene DTD2 foi inserida na porção 3‘UTR do gene BBS5, e é um exemplo da criação de um contexto novo de regulação pós-transcricional pela inserção de sequências adicionais na região 3‘UTR ou criando pares de genes parental e retrocópia co-regulados (Poliseno et al., 2010). Outro exemplo é a retrocópia do gene AK4 próximo ao gene DENND5B que gera um

transcrito antisenso ao gene hospedeiro e pode ser considerada um bom exemplo de inserções intragênicas ou próximas de genes hospedeiros que criam a possibilidade de alterar o padrão de *splicing* de todo o *locus* transcrito, criando ou fusionando novos domínios proteicos. Portanto, cada nova retrocópia, sejam elas específicas ou compartilhadas por várias espécies é uma variações que cria oportunidades para modificação de contexto e de genes existentes e permitem que organismos se adaptem a novos ambientes ou condições.

5.7. Retrocópias polimórficas germinativas.

Estima-se que entre 60 e 100 elementos L1 ainda estão ativos no genoma humano (Brouha et al., 2003). A atividade destes elementos repetitivos continua gerando inserções de LINEs (L1HS) e SINEs (Alus e SVAs) no genoma humano e, consequentemente, é responsável por gerar variação entre indivíduos e populações (Beck et al., 2010). Retrocópias são subprodutos da retroposição, em *trans*, mediada pela atividade da transcriptase reversa de elementos L1 e, portanto, é razoável supor que retrocópias também foram criadas em ancestrais de humanos e humanos contemporâneos. Baseado nesta hipótese, e no fato da literatura descrever uma retrocópia como polimórfica (Anagnou et al., 1988), buscamos, nos sequenciamentos realizados pelo projeto *1000 Genomes,* evidências de polimorfismos de presença e ausência de retrocópias humano específicas e chamamos estes eventos de retroCNVs. Assim como Anagnou e colaboradores, encontramos que a retrocópia do gene DHFR (DHFRP1) não está totalmente fixada na espécie humana e, adicionalmente, descrevemos 16 retrocópias presentes no genoma referência polimórficas quanto a presença e ausência em indivíduos do projeto *1000 Genomes*. Por estarem presentes no genoma referência (hg19/

GRCh37), estes retroCNVs devem ter uma representatividade relativamente alta nos indivíduos utilizados para a construção do genoma referência e, portanto, devem ter uma frequência alélica relativamente alta. De fato, os retroCNVs presentes no genoma referência estão, em média, em 75% dos cromossomos analisados, com casos próximos da fixação como, por exemplo, a retrocópia DHFRP1 que está presente em 90% dos indivíduos com ancestralidade europeia (Tabela 15) (Schrider et al., 2013). Também investigamos se seria possível a existência de retrocópias com representatividade baixa, a ponto de não estarem no genoma referência (1000 Genomes Project Consortium et al., 2012). Para nossa surpresa, baseado na análise de 22 genomas, encontramos 20 novas retrocópias ausentes no genoma humano de referência. Ao genotipar estes retroCNVs notamos que a frequência alélica destas inserções são significativamente menores que as frequência alélica das retrocópias presentes no genoma referência (média 15% em todas as populações). A única exceção é o gene CBX3, que está presente em 57% dos cromossomos analisados.

Ao analisar as sequências dos retroCNVs, presentes ou ausentes no genoma referência, encontramos que praticamente todos são muito similares ao seu respectivo gene parental e muitas vezes não apresentam alteração na sequência retrocopiada equivalente a sequência proteica codificada pelo gene parental. Nos questionamos se estes eventos deveriam ser classificados como retrogenes ou pseudogenes processados. A favor da classificação destes retroCNVs como retrogenes, temos a definição de genes pela semelhança com outros genes codificadores de proteína, sem necessariamente apresentar evidência de expressão proteica (Gerstein et al., 2007). A favor da classificação de pseudogenes, temos a ausência de evidência de expressão e o fato de que a maioria das retrocópias no genoma humano acumularam mutações que impedem a sua tradução. Entretanto,

lembramos que assim como retrocópias, *locus* hoje anotados como genes, podem acumular mutações e tornarem-se pseudogenes no futuro (Pei et al., 2012). Portanto, a classificação do *locus* deve ser feita não baseada em condições futuras, mas na condição atual do *locus*. Desta forma, devido a alta identidade e a ausência de mutações que destruam a sequência proteica em vários retroCNVs, podemos assumir que muitos destes eventos não deveriam ser anotados como pseudogenes processados, mas como novos retrogenes gerados por retroposição de mRNAs.

Outra questão importante é o impacto funcional destas retrocópias no genoma humano. Infelizmente, a similaridade entre retroCNVs e genes parentais dificultam a detecção de casos de expressão *per se*, já que é impossível discernir entre a expressão do gene parental e da retrocópia e pequenos erros de sequenciamento podem gerar falsos positivos. Entretanto, podemos investigar a expressão quimérica de retrocópias e genes hospedeiros. Encontramos evidência de expressão quimérica da retrocópia do gene CBX3, um gene com função de silenciamento transcricional por formação de heterocromatina (Smallwood, A. et al., 2012), inserida entre o segundo e terceiro exon do gene C15orf57. Apesar de não estar anotado, este gene hospedeiro apresenta semelhança com o genes CDC, que tem atividade de transporte de metabólitos para o interior da célula e é expresso no sangue, bem como em vários tecidos como testículo. Finalmente, também detectamos inserções exônicas, como, por exemplo, a inserção do gene UQCR10 na região codificadora do gene C1orf154. A inserção não é no mesmo quadro de leitura do gene hospedeiro, portanto, há uma destruição da proteína previamente sintetizada pelo gene C1orf154. Neste caso, é possível afirmar que indivíduos heterozigotos com alelos contendo a retrocópia inserida dentro na região exônica do gene C1orf154

expressam, além da versão referência do gene C11org154, um novo gene, contendo a fusão entre o gene C1orf154 e UQCR10.

Logo após a publicação dos nossos resultados, dois trabalhos muito similares também investigaram a existência de polimorfismos de presença e ausência de retrocópias no genoma humano. Ewing e colaboradores utilizaram sequenciamento de segunda geração e estratégias muito similares para descrever retroCNVs não só em humanos, mas também em camundongos, chimpanzés e genomas tumorais (Ewing et al., 2013). Segundo o número de retroCNVs identificados em humanos, é estimado que haja uma inserção a cada 6.000 nascimentos. Além disso, este trabalho descreveu, pela primeira vez, a presença de retroCNVs em tumores. Posteriormente, Abyzov e colaboradores utilizaram os dados do projeto *1000 Genomes* para identificar retrocópias presentes no genoma referência com evidência de ausência e novas inserções ausentes do genoma referência (Abyzov et al., 2013). Curiosamente, os eventos reportados como polimórficos e presentes no genoma referência, frequentemente estão conservados em outros primatas e envolvem regiões maiores ou menores que a retrocópia em si. Portanto, ao menos para estes eventos, diferente do polimorfismo de ausência da inserção ou presença da inserção, a variação se dá pela deleção de retrocópias previamente fixadas no genoma de primatas. Abyzov e colaboradores também reportam 147 inserções ausentes no genoma referência, com apenas 16 pontos de inserção, baseado na análise de 974 indivíduos do projeto *1000 Genomes*. Sobretudo, encontramos que é provável que as variações de presença e ausência detectadas entre espécies de primatas também atuem entre indivíduos da mesma espécie na forma de retroCNVs, porém, trabalhos adicionais serão necessários para que se entenda qual a extensão desta variação e quais os possíveis impactos na biologia humana.

5.8. Retrocópias polimórficas somáticas em tumores.

Tumores frequentemente apresentam um estado de hipometilação genômica ou mutações em vias responsáveis pelo silenciamento de retroposição de elementos repetitivos (Ross et al., 2014). A desregulação dos mecanismos de silenciamento tem como conseqüência o incremento do número de cópias somáticas destes elementos e, portanto, elevação do potencial mutagênico causado pela retroposição de elementos repetitivos, sejam eles autônomos (LINEs e HERVs) ou não autônomos (por exemplo, SINEs e SVAs) (Beck et al., 2011). Visto que a retroposição de mRNAs maduros é mediada pela maquinaria de transcriptase reversa de LINEs, é de se esperar que, com a diminuição das restrições da retroposição, também haja um maior número de retroCNVs somáticos em tumores que em tecidos normais (Cooke et al., 2014). Apesar de retroCNVs serem, a priori, um dos mecanismos para criação de novos genes ou novas variantes, hipoteticamente, é possível que a inserção de uma sequência potencialmente codificadora possa colaborar para a tumorigênese de diversas formas: i) geração de cópias funcionais de oncogenes com perfil de expressão distinto do gene parental; ii) mutação de genes supressores de tumores pela inserção de retroCNVs, causando, por exemplo, a modificação do seu perfil de *splicing*. iii) alteração da regulação do gene parental e, finalmente, iv) modificação da expressão do gene hospedeiro (Lee et al., 2012). A fim de investigar o alto potencial mutagênico dos retroCNVs somáticos, desenvolvemos e aplicamos os métodos descritos acima em seis amostras de tumor de cólon, que sabidamente é descrito como um dos tumores com maior nível de retroposição somática (Solyom et al., 2012).

Nós encontramos sete possíveis inserções de genes codificadores de proteína em cinco genomas tumorais de câncer colorretal (Tabela 18). Curiosamente, todas as inserções são de tamanho muito reduzido e, portanto, não sobrepõem junções exon-exon. Adicionalmente, não encontramos evidência da presença de tratos de múltiplas adeninas na porção 3' das inserções ou repetições diretas, apesar de termos validado a presença de todas as inserções através de amplificação com primers específicos seguido de sequenciamento. Trabalhos anteriores mostram que, de fato, quando elementos L1 são retropostos em tumores há uma depreciação na qualidade da transcriptase reversa e a maioria das retroposições detectadas são severamente truncadas em sua porção 5' (Helman et al., 2014 e Lee et al., 2012 e Solyom et al., 2012). Entretanto, apesar da ausência destes sinais moleculares, é difícil especular sobre quais outros mecanismos poderiam ser responsáveis pela presença de uma inserção de regiões 3'UTR em uma região aleatória do genoma. Entre as possibilidades estão deleções e recombinações de regiões relativamente distantes. Adicionalmente, Cooke e colaboradores, descrevem que apenas 10% das retroposições somáticas de mRNA estão no mesmo cromossomo do gene parental, em contraste, nós encontramos que aproximadamente 60% (quatro retroCNVs) das retroposições encontradas e validadas em nosso trabalho estão no mesmo cromossomo que os genes parentais.

Sobretudo, RetroCNV somáticos foram inicialmente descritos em 2013 e a primeira análise em larga escala foi publicada em 2014. Nesta análise em larga escala, Cooke e colaboradores (Cooke et al., 2014) encontraram apenas 42 retroCNVs somáticos em 17 tumores dos 660 tumores analisados. É possível que, devido aos parâmetros estringentes do método utilizado, o qual exige, por exemplo, duas junções exon-exon de regiões codificadoras, haja um enriquecimento de falsos

negativos nos resultados reportados. Entretanto, como pioneiro, este trabalho indica que alguns tumores são mais suscetíveis à retroposição de mRNAs, em especial, tumores escamosos pulmonares e tumores de cólon. Assim como os resultados aqui apresentados, os autores concluem que a retroposição de mRNAs e criação de retroCNVs somáticos são uma nova classe de mutação, ou variação genética, ocorrida durante o desenvolvimento de tumores.

5.9. Expressão de retrocópias

Vários estudos tem reportado um número crescente de retrocópias expressas e potencialmente funcionais (Harrison et al., 2005 e Kalyana-Sundaram et al., 2012 e Poliseno et al., 2010 e Yano et al., 2004). O fato de algumas retrocópias expressas não apresentarem função estritamente codificadora de proteína e terem um papel importante na regulação de genes parentais (Poliseno et al., 2010) é ainda mais inesperado. O mecanismo de expressão destas retrocópias, apesar de elusivo, parte do pressuposto de que retrocópias sequestram regiões regulatórias de seu novo contexto genômico, sendo transcritas *per se* ou transcritas de carona quando inseridas dentro de regiões transcritas de genes (Vinckenbosch et al., 2006). Nós utilizamos dados de RNA-seq e um *pipeline* de alta especificidade para definir a expressão de *loci* anotados como retrocópias e identificamos um conjunto de aproximadamente 3.600 retrocópias transcritas em cinco primatas e seis tecidos. Apesar de ser um conjunto restrito, nós também identificamos retrocópias intragênicas expressas como parte de transcritos quiméricos de seus genes hospedeiros. Entre os transcritos quiméricos, encontramos duas retrocópias ausentes do genoma referência, que podem estar gerando variações entre indivíduos.

Sob uma perspectiva mais geral, nós também identificamos conjuntos de retrocópias apresentando expressão espécie específica e/ou expressão tecido-específica. E, similar a Marques e colaboradores, nós identificamos um enriquecimento de retrocópias expressas em tecido nervoso central (cérebro e cerebelo) e testículo, ambos descritos como tecidos com permissividade maior para expressão de *loci* funcionais e também não funcionais. Adicionalmente, também comparamos o número de tecidos em que retrocópias e seus respectivos genes parentais são expressos e, enquanto retrocópias apresentam uma especificidade maior de tecido, genes parentais tem uma expressão mais ubíqua.

Ainda sobre a questão de quão similar é a expressão da retrocópias e seus respectivos genes parentais, avaliamos qual a correlação da expressão de ambos os grupos em seis tecidos. Encontramos que as retrocópias e seus genes parentais não tem uma correlação direta de expressão e, portanto, além de eliminarmos a possibilidade de enriquecimento de alinhamentos falsos positivos em retrocópias, encontramos evidência de que o contexto diferente da inserção da retrocópia e seu gene parental geram uma oportunidade para que a retrocópia seja expressa em diferentes níveis e diferentes tecidos, podendo atuar como um mecanismo para mudar o perfil de expressão de um gene duplicado, como esperado do ponto de vista teórico.

Capítulo 6. Conclusões

“Se um viajante eterno a atravessasse em qualquer direção comprovaria ao cabo de séculos que os mesmos volumes se repetem na mesma desordem”

Jorge Luis Borges - Ficções

Desenvolvemos e aplicamos o *pipeline* de detecção de retrocópias em sete genomas de primatas e dois roedores, nominalmente, humanos, chimpanzés, gorilas, orangotangos, rhesus, saguis, macacos esquilo, camundongos e ratos. Também disponibilizamos os resultados encontrados em forma de uma ferramenta *web*, a RCPedia.

Encontramos que retrocópias são fatores de variação genética inter-espécies. Apesar de roedores e primatas do velho mundo apresentarem cerca de 7.500 retrocópias em seus genomas referência, confirmamos que o conjunto de retrocópias destas linhagens originou-se independentemente e a maioria dos 63 eventos compartilhados entre ambas linhagens são atualmente anotados como funcionais.

Entre primatas, Platyrrhinis (primatas do novo mundo) apresentam um enriquecimento de aproximadamente 50% mais retrocópias (~10.000 eventos) quando comparados aos genomas de Catarrhinis (primatas do velho mundo) e este enriquecimento pode ser decorrente da maior atividade de dois elementos transponíveis, L1PA7 e L1P3, nestes genomas.

A maioria das retrocópias em humanos (~53%) são compartilhadas por todas as espécies de primatas estudadas. Se inicialmente, a taxa de criação e fixação de retrocópias foi alta (~152 retrocópias por milhão de anos), ela decresceu a medida que as especiações foram ocorrendo, e atualmente, é menor em humanos (21 retrocópias por milhão de anos).

Retrocópias também são um fator de variação genética intra-espécie (chamado por nós de retroCNV germinativo) e cobrem todo o espectro de frequência

alélica, com enriquecimento de alelos de frequência baixa e não representados no genoma referência humano (16 retroCNVs presentes no genoma referência e 20 ausentes do genoma referência). RetroCNVs também aumentam a variabilidade do transcriptoma humano criando, por exemplo, transcritos quiméricos com seus respectivos genes hospedeiros. Portanto, é possível que o impacto das retrocópias polimórficas seja ainda maior, mas ainda pouco explorado devido as limitações amostrais e técnicas para explorar a expressão dos retroCNVs.

RetroCNVs somáticos são potencialmente criados em um contexto tumoral e compreendem em uma nova classe de mutação. Análises futuras sobre a retroposição somática de mRNA em tumores primários, linhagens tumorais e indivíduos com mutações relacionadas a ativação de elementos repetitivos devem elucidar a frequência e influência de retroCNVs somáticos em patologias humanas.

REFERÊNCIAS

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. **Nature**, v. 467, n. 7319, p. 1061–1073, 2010.
1000 Genomes Project Consortium; Abecasis, G. R.; Auton, A.; et al. An integrated map of genetic variation from 1,092 human genomes. **Nature**, v. 491, n. 7422, p. 56–65, 2012.
Abyzov, A.; Iskow, R.; Gokcumen, O.; et al. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. **Genome research**, 2013.
Adra, C. N.; Ellis, N. A.; McBurney, M. W. The family of mouse phosphoglycerate kinase genes and pseudogenes. **Somatic cell and molecular genetics**, v. 14, n. 1, p. 69–81, 1988.
Anagnou, N. P.; Antonarakis, S. E.; O'Brien, S. J.; Modi, W. S.; Nienhuis, A. W. Chromosomal localization and racial distribution of the polymorphic human dihydrofolate reductase pseudogene (DHFRP1). **American journal of human genetics**, v. 42, n. 2, p. 345–352, 1988.
Aravin, A. A.; Sachidanandam, R.; Girard, A.; Fejes-Toth, K.; Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. **Science (New York, N.Y.)**, v. 316, n. 5825, p. 744–747, 2007.
Badge, R. M.; Alisch, R. S.; Moran, J. V. ATLAS: a system to selectively identify human-specific L1 insertions. **American journal of human genetics**, v. 72, n. 4, p. 823–838, 2003.
Baertsch, R.; Diekhans, M.; Kent, W. J.; Haussler, D.; Brosius, J. Retrocopy contributions to the evolution of the human genome. **BMC genomics**, v. 9, p. 466, 2008.
Baillie, J. K.; Barnett, M. W.; Upton, K. R.; et al. Somatic retrotransposition alters the genetic landscape of the human brain. **Nature**, v. 479, n. 7374, p. 534–537, 2011.
Balasubramanian, S.; Zheng, D.; Liu, Y.-J.; et al. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. **Genome biology**, v. 10, n. 1, p. R2, 2009.
Batzer, M. A.; Deininger, P. L. Alu repeats and human genomic diversity. **Nature reviews. Genetics**, v. 3, n. 5, p. 370–379, 2002.
Beck, C. R.; Collier, P.; Macfarlane, C.; et al. LINE-1 retrotransposition activity in human genomes. **Cell**, v. 141, n. 7, p. 1159–1170, 2010.
Beck, C. R.; Garcia-Perez, J. L.; Badge, R. M.; Moran, J. V. LINE-1 elements in structural variation and disease. **Annual review of genomics and human genetics**, v. 12, p. 187–215, 2011.
Becker, K. G.; Swergold, G. D.; Ozato, K.; Thayer, R. E. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. **Human Molecular Genetics**, v. 2, n. 10, p. 1697–1702, 1993.
Benson, D. A.; Cavanaugh, M.; Clark, K.; et al. GenBank. **Nucleic acids research**, v. 41, n. Database issue, p. D36–42, 2013.
Berget, S. M.; Moore, C.; Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 8, p. 3171–3175, 1977.
Bogerd, H. P.; Wiegand, H. L.; Hulme, A. E.; et al. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. **Proceedings of the National**

Academy of Sciences of the United States of America, v. 103, n. 23, p. 8780–8785, 2006.
Boissinot, S.; Entezam, A.; Young, L.; Munson, P. J.; Furano, A. V. The insertional history of an active family of L1 retrotransposons in humans. **Genome research**, v. 14, n. 7, p. 1221–1231, 2004a.
Boissinot, S.; Roos, C.; Furano, A. V. Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. **Journal of molecular evolution**, v. 58, n. 1, p. 122–130, 2004b.
Brawand, D.; Soumillon, M.; Necsulea, A.; et al. The evolution of gene expression levels in mammalian organs. **Nature**, v. 478, n. 7369, p. 343–348, 2012. Nature Publishing Group.
Brosius, J. Retroposons--seeds of evolution. **Science (New York, N.Y.)**, v. 251, n. 4995, p. 753, 1991.
Brouha, B.; Schustak, J.; Badge, R. M.; et al. Hot L1s account for the bulk of retrotransposition in the human population. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 9, p. 5280–5285, 2003.
Chiu, Y.-L.; Greene, W. C. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. **Annual review of immunology**, v. 26, p. 317–353, 2008.
Chow, L. T.; Roberts, J. M.; Lewis, J. B.; Broker, T. R. A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. **Cell**, v. 11, n. 4, p. 819–836, 1977.
Conrad, D. F.; Pinto, D.; Redon, R.; et al. Origins and functional impact of copy number variation in the human genome. **Nature**, v. 464, n. 7289, p. 704–712, 2010.
Cooke, S. L.; Shlien, A.; Marshall, J.; et al. Processed pseudogenes acquired somatically during cancer development. **Nature communications**, v. 5, p. 3644, 2014.
Cost, G. J.; Feng, Q.; Jacquier, A.; Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. **The EMBO journal**, v. 21, n. 21, p. 5899–5910, 2002.
Coufal, N. G.; Garcia-Perez, J. L.; Peng, G. E.; et al. L1 retrotransposition in human neural progenitor cells. **Nature**, v. 460, n. 7259, p. 1127–1131, 2009.
Craig, N. L. **MOBILE DNA II**. NY, 1980.
Dewannieux, M; Heidmann, T. LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. **Cytogenetic and genome research**, v. 110, n. 1-4, p. 35–48, 2005.
Dewannieux, Marie; Esnault, C.; Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. **Nature genetics**, v. 35, n. 1, p. 41–48, 2003.
Dunham, I.; Shimizu, N.; Roe, B. A.; et al. The DNA sequence of human chromosome 22. **Nature**, v. 402, n. 6761, p. 489–495, 1999.
Ehsani, S.; Tao, R.; Pocanschi, C. L.; et al. Evidence for retrogene origins of the prion gene family. **PloS one**, v. 6, n. 10, p. e26800, 2011.
Ellegren, H.; Parsch, J. The evolution of sex-biased genes and sex-biased gene expression. **Nature reviews. Genetics**, v. 8, n. 9, p. 689–698, 2007.
Emerson, J. J.; Kaessmann, H.; Betrán, E.; Long, M. Extensive gene traffic on the mammalian X chromosome. **Science (New York, N.Y.)**, v. 303, n. 5657, p. 537–540, 2004.
Esnault, C.; Maestre, J.; Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. **Nature genetics**, v. 24, n. 4, p. 363–367, 2000.
Esteller, M. Non-coding RNAs in human disease. **Nature reviews. Genetics**, v. 12, n. 12, p. 861–874, 2011.

Evrony, G. D.; Cai, X.; Lee, E.; et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. **Cell**, v. 151, n. 3, p. 483–496, 2012.
Ewing, A. D.; Ballinger, T. J.; Earl, D.; et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. **Genome biology**, v. 14, n. 3, p. R22, 2013.
Ewing, A. D.; Kazazian, Haig H. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. **Genome research**, v. 20, n. 9, p. 1262–1270, 2010.
Fairbanks, D. J.; Fairbanks, A. D.; Ogden, T. H.; Parker, G. J.; Maughan, P. J. NANOGP8: evolution of a human-specific retro-oncogene. **G3 (Bethesda, Md.)**, v. 2, n. 11, p. 1447–1457, 2012.
Farley, A. H.; Luning Prak, E. T.; Kazazian, H. H. More active human L1 retrotransposons produce longer insertions. **Nucleic acids research**, v. 32, n. 2, p. 502–510, 2004.
Feng, Q.; Moran, J. V.; Kazazian, H. H.; Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. **Cell**, v. 87, n. 5, p. 905–916, 1996.
Frazer, K. A.; Murray, S. S.; Schork, N. J.; Topol, E. J. Human genetic variation and its contribution to complex traits. **Nature reviews. Genetics**, v. 10, n. 4, p. 241–251, 2009.
Gasior, S. L.; Roy-Engel, A. M.; Deininger, P. L. ERCC1/XPF limits L1 retrotransposition. **DNA repair**, v. 7, n. 6, p. 983–989, 2008.
Gerstein, M. B.; Bruce, C.; Rozowsky, J. S.; et al. What is a gene, post-ENCODE? History and updated definition. **Genome research**, v. 17, n. 6, p. 669–681, 2007.
Gibbs, R. A.; Weinstock, G. M.; Metzker, M. L.; et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. **Nature**, v. 428, n. 6982, p. 493–521, 2004.
Goodier, J. L.; Kazazian, Haig H. Retrotransposons revisited: the restraint and rehabilitation of parasites. **Cell**, v. 135, n. 1, p. 23–35, 2008.
Harrison, P. M.; Hegyi, H.; Balasubramanian, S.; et al. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. **Genome research**, v. 12, n. 2, p. 272–280, 2002.
Harrison, P. M.; Zheng, D.; Zhang, Z.; Carriero, N.; Gerstein, M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. **Nucleic acids research**, v. 33, n. 8, p. 2374–2383, 2005.
Harrow, J.; Frankish, A.; Gonzalez, J. M.; et al. GENCODE: the reference human genome annotation for The ENCODE Project. **Genome research**, v. 22, n. 9, p. 1760–1774, 2012.
Hattori, M.; Fujiyama, A.; Taylor, T. D.; et al. The DNA sequence of human chromosome 21. **Nature**, v. 405, n. 6784, p. 311–319, 2000.
Havecker, E. R.; Gao, X.; Voytas, D. F. The diversity of LTR retrotransposons. **Genome biology**, v. 5, n. 6, p. 225, 2004.
Hazkani-Covo, E.; Sorek, R.; Graur, D. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. **Journal of molecular evolution**, v. 56, n. 2, p. 169–174, 2003.
Helman, E.; Lawrence, M. L.; Stewart, C.; et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. **Genome research**, 2014.

Horn, A. V.; Klawitter, S.; Held, U.; et al. Human LINE-1 restriction by APOBEC3C is deaminase independent and mediated by an ORF1p interaction that affects LINE reverse transcriptase activity. **Nucleic acids research**, 2013.
Houtsmuller, A. B.; Rademakers, S.; Nigg, A. L.; et al. Action of DNA repair endonuclease ERCC1/XPF in living cells. **Science (New York, N.Y.)**, v. 284, n. 5416, p. 958–961, 1999.
Hung, M.-S.; Lin, Y.-C.; Mao, J.-H.; et al. Functional polymorphism of the CK2alpha intronless gene plays oncogenic roles in lung cancer. **PloS one**, v. 5, n. 7, p. e11418, 2010.
International HapMap Consortium. The International HapMap Project. **Nature**, v. 426, n. 6968, p. 789–796, 2003.
Iskow, R. C.; McCabe, M. T.; Mills, R. E.; et al. Natural mutagenesis of human genomes by endogenous retrotransposons. **Cell**, v. 141, n. 7, p. 1253–1261, 2010.
Jongeneel, C. V.; Delorenzi, M.; Iseli, C.; et al. An atlas of human gene expression from massively parallel signature sequencing (MPSS). **Genome research**, v. 15, n. 7, p. 1007–1014, 2005.
Kaessmann, H.; Vinckenbosch, N.; Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. **Nature reviews. Genetics**, v. 10, n. 1, p. 19–31, 2009.
Kalyana-Sundaram, S.; Kumar-Sinha, C.; Shankar, S.; et al. Expressed pseudogenes in the transcriptional landscape of human cancers. **Cell**, v. 149, n. 7, p. 1622–1634, 2012.
Karro, J. E.; Yan, Y.; Zheng, D.; et al. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. **Nucleic acids research**, v. 35, n. Database issue, p. D55–60, 2007.
Kashiwabara, S.-I.; Noguchi, J.; Zhuang, T.; et al. Regulation of spermatogenesis by testis-specific, cytoplasmic poly(A) polymerase TPAP. **Science (New York, N.Y.)**, v. 298, n. 5600, p. 1999–2002, 2002.
Kazazian, H H; Moran, J V. The impact of L1 retrotransposons on the human genome. **Nature genetics**, v. 19, n. 1, p. 19–24, 1998.
Kazazian, Haig H. Mobile elements: drivers of genome evolution. **Science (New York, N.Y.)**, v. 303, n. 5664, p. 1626–1632, 2004.
Kellis, M.; Wold, B.; Snyder, M. P.; et al. Defining functional DNA elements in the human genome. **Proceedings of the National Academy of Sciences**, v. 111, n. 17, p. 6131–6138, 2014.
Kent, W J. BLAT---The BLAST-Like Alignment Tool. **Genome research**, v. 12, n. 4, p. 656–664, 2002.
Kent, W James; Sugnet, C. W.; Furey, T. S.; et al. The human genome browser at UCSC. **Genome research**, v. 12, n. 6, p. 996–1006, 2002.
Khachane, A. N.; Harrison, P. M. Assessing the genomic evidence for conserved transcribed pseudogenes under selection. **BMC genomics**, v. 10, p. 435, 2009.
Khelifi, A.; Adel, K.; Duret, L.; et al. HOPPSIGEN: a database of human and mouse processed pseudogenes. **Nucleic acids research**, v. 33, n. Database issue, p. D59–66, 2005.
Kojima, K. K.; Okada, N. mRNA retrotransposition coupled with 5' inversion as a possible source of new genes. **Molecular biology and evolution**, v. 26, n. 6, p. 1405–1420, 2009.
Konkel, M. K.; Walker, J. A.; Batzer, M. A. LINEs and SINEs of primate evolution. **Evolutionary Anthropology: Issues, News, and Reviews**, v. 19, n. 6, p. 236–249, 2011.

Krebs, J. E.; Goldstein, E. S.; Kilpatrick, S. T. **Lewin's GENES X**. 10th ed. Jones & Bartlett Learning, 2009.
Krzywinski, M.; Schein, J.; Birol, I.; et al. Circos: an information aesthetic for comparative genomics. **Genome research**, v. 19, n. 9, p. 1639–1645, 2009.
Kubo, S.; Seleme, M. D. C.; Soifer, H. S.; et al. L1 retrotransposition in nondividing and primary human somatic cells. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 21, p. 8036–8041, 2006.
Kulpa, D. A.; Moran, John V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. **Nat. Struct. Mol. Biol.**, v. 13, n. 7, p. 655–660, 2006.
Kumar, S.; Subramanian, S. Mutation rates in mammalian genomes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 2, p. 803–808, 2002.
Kuramochi-Miyagawa, S.; Watanabe, T.; Gotoh, K.; et al. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. **Genes & development**, v. 22, n. 7, p. 908–917, 2008.
Lander, E. S.; Linton, L. M.; Birren, B.; et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 2001. Nature Publishing Group.
Langmead, B.; Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Publishing Group**, v. 9, n. 4, p. 357–359, 2012.
Larkin, M. A.; Blackshields, G.; Brown, N. P.; et al. Clustal W and Clustal X version 2.0. **Bioinformatics (Oxford, England)**, v. 23, n. 21, p. 2947–2948, 2007.
Lau, N. C.; Seto, A. G.; Kim, J.; et al. Characterization of the piRNA complex from rat testes. **Science (New York, N.Y.)**, v. 313, n. 5785, p. 363–367, 2006.
Lau, N. C.; Robine, N.; Martin, R.; et al. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. **Genome research**, v. 19, n. 10, p. 1776–1785, 2009.
Lee, E.; Iskow, R.; Yang, L.; et al. Landscape of somatic retrotransposition in human cancers. **Science (New York, N.Y.)**, v. 337, n. 6097, p. 967–971, 2012.
Levy, S.; Sutton, G.; Ng, P. C.; et al. The diploid genome sequence of an individual human. **PLoS biology**, v. 5, n. 10, p. e254, 2007.
Li, H.; Handsaker, B.; Wysoker, A.; et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics (Oxford, England)**, v. 25, n. 16, p. 2078–2079, 2009.
Li, Q.; Laumonnier, Y.; Syrovets, T.; Simmet, T. Yeast two-hybrid screening of proteins interacting with plasmin receptor subunit: C-terminal fragment of annexin A2. **Acta pharmacologica Sinica**, v. 32, n. 11, p. 1411–1418, 2011.
Li, W. H.; Gojobori, T.; Nei, M. Pseudogenes as a paradigm of neutral evolution. **Nature**, v. 292, n. 5820, p. 237–239, 1981.
Liu, J.; Nau, M. M.; Zucman-Rossi, J.; et al. LINE-I element insertion at the t(11;22) translocation breakpoint of a desmoplastic small round cell tumor. **Genes, chromosomes & cancer**, v. 18, n. 3, p. 232–239, 1997.
Liu, Y.-J.; Zheng, D.; Balasubramanian, S.; 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**, v. 10, p. 480, 2009.
Lomedico, P.; Rosenthal, N.; Efstratidadis, A.; et al. The structure and evolution of the two nonallelic rat preproinsulin genes. **Cell**, v. 18, n. 2, p. 545–558, 1979.
Luan, D. D.; Korman, M. H.; Jakubczak, J. L.; Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. **Cell**, v. 72, n. 4, p. 595–605, 1993.

Lynch, M. **The Origins of Genome Architecture**. 1st ed. Sinauer Associates Inc, 2007.
Maestre, J.; Tchénio, T.; Dhellin, O.; Heidmann, T. mRNA retroposition in human cells: processed pseudogene formation. **The EMBO journal**, v. 14, n. 24, p. 6333–6338, 1995.
Magiorkinis, G.; Gifford, R. J.; Katzourakis, A.; De Ranter, J.; Belshaw, R. Env-less endogenous retroviruses are genomic superspreaders. **Proceedings of the National Academy of Sciences**, v. 109, n. 19, p. 7385–7390, 2012.
Mandal, P. K.; Ewing, A. D.; Hancks, D. C.; Kazazian, H. H. Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. **Human Molecular Genetics**, v. 22, n. 18, p. 3730–3748, 2013.
Marchetto, M. C. N.; Narvaiza, I.; Denli, A. M.; et al. Differential L1 regulation in pluripotent stem cells of humans and apes. **Nature**, 2013.
Mardis, E. R. A decade's perspective on DNA sequencing technology. **Nature**, v. 470, n. 7333, p. 198–203, 2011.
Marques, A. C.; Dupanloup, I.; Vinckenbosch, N.; Reymond, A.; Kaessmann, H. Emergence of young human genes after a burst of retroposition in primates. **PLoS biology**, v. 3, n. 11, p. e357, 2005.
Martin, S. L. The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. **Journal of Biomedicine and Biotechnology**, v. 2006, n. 1, p. 45621, 2006.
Martin, S. L. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. **RNA Biology**, v. 7, n. 6, p. 706–711, 2010.
Martin, S. L.; Cruceanu, M.; Branciforte, D.; et al. LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. **Journal of molecular biology**, v. 348, n. 3, p. 549–561, 2005.
McEntee, G.; Minguzzi, S.; O'Brien, K.; et al. The former annotated human pseudogene dihydrofolate reductase-like 1 (DHFRL1) is expressed and functional. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 37, p. 15157–15162, 2011.
Mercer, T. R.; Dinger, M. E.; Mattick, J. S. Long non-coding RNAs: insights into functions. **Nature reviews. Genetics**, v. 10, n. 3, p. 155–159, 2009.
Miki, Y.; Nishisho, I.; Horii, A.; et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. **Cancer Research**, v. 52, n. 3, p. 643–645, 1992.
Morse, B.; Rotherg, P. G.; South, V. J.; Spandorfer, J. M.; Astrin, S. M. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. **Nature**, v. 333, n. 6168, p. 87–90, 1988.
Mouse Genome Sequencing Consortium; Waterston, R. H.; Lindblad-Toh, K.; et al. Initial sequencing and comparative analysis of the mouse genome. **Nature**, v. 420, n. 6915, p. 520–562, 2002.
Muckenfuss, H.; Hamdorf, M.; Held, U.; et al. APOBEC3 proteins inhibit human LINE-1 retrotransposition. **The Journal of biological chemistry**, v. 281, n. 31, p. 22161–22172, 2006.
Muotri, A. R.; Chu, V. T.; Marchetto, M. C. N.; et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. **Nature**, v. 435, n. 7044, p. 903–910, 2005.
Myers, J. S.; Vincent, B. J.; Udall, H.; et al. A comprehensive analysis of recently integrated human Ta L1 elements. **American journal of human genetics**, v. 71, n. 2, p. 312–326, 2002.

Navarro, F. C. P.; Galante, P. A. F. RCPedia: a database of retrocopied genes. **Bioinformatics (Oxford, England)**, v. 29, n. 9, p. 1235–1237, 2013.
Nishioka, Y.; Leder, A.; Leder, P. Unusual alpha-globin-like gene that has cleanly lost both globin intervening sequences. **Proceedings of the National Academy of Sciences of the United States of America**, v. 77, n. 5, p. 2806–2809, 1980.
Ohshima, K.; Hattori, M.; Yada, T.; et al. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. **Genome biology**, v. 4, n. 11, p. R74, 2003.
Ostertag, E. M.; Kazazian, H H. Biology of mammalian L1 retrotransposons. **Annual review of genetics**, v. 35, p. 501–538, 2001a.
Ostertag, E. M.; Kazazian, H H. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. **Genome research**, v. 11, n. 12, p. 2059–2065, 2001b.
Pei, B.; Sisu, C.; Frankish, A.; et al. The GENCODE pseudogene resource. **Genome biology**, v. 13, n. 9, p. R51, 2012. BioMed Central Ltd.
Perez, S. I.; Tejedor, M. F.; Novo, N. M.; Aristide, L. Divergence Times and the Evolutionary Radiation of New World Monkeys (Platyrrhini, Primates): An Analysis of Fossil and Molecular Data. **PloS one**, v. 8, n. 6, p. e68029, 2013.
Piskareva, O.; Schmatchenko, V. DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. **FEBS letters**, v. 580, n. 2, p. 661–668, 2006.
Poliseno, L.; Salmena, L.; Zhang, J.; et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. **Nature**, v. 465, n. 7301, p. 1033–1038, 2010.
Pruitt, K. D.; Brown, G. R.; Hiatt, S. M.; et al. RefSeq: an update on mammalian reference sequences. **Nucleic acids research**, 2013.
Ray, D. A.; Walker, J. A.; Batzer, M. A. Mobile element-based forensic genomics. **Mutation research**, v. 616, n. 1-2, p. 24–33, 2007.
Reuter, M.; Berninger, P.; Chuma, S.; et al. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. **Nature**, v. 480, n. 7376, p. 264–267, 2011.
Ross, R. J.; Weiner, M. M.; Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. **Nature**, v. 505, n. 7483, p. 353–359, 2014.
Sakai, H.; Koyanagi, K. O.; Imanishi, T.; Itoh, T.; Gojobori, T. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. **Gene**, v. 389, n. 2, p. 196–203, 2007.
Scally, A.; Dutheil, J. Y.; Hillier, L. W.; et al. Insights into hominid evolution from the gorilla genome sequence. **Nature**, v. 483, n. 7388, p. 169–175, 2012.
Scally, A.; Durbin, R. Revising the human mutation rate: implications for understanding human evolution. **Nature reviews. Genetics**, v. 13, n. 10, p. 745–753, 2012.
Scherer, S. **A Short Guide to the Human Genome**. Cold Spring Harbor Laboratory Press, 2008.
Schrider, D. R.; Navarro, F. C. P.; Galante, P. A. F.; et al. Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans. (J. M. Akey, Ed.)**PLoS genetics**, v. 9, n. 1, p. e1003242, 2013.
Schwartz, S.; Kent, W. J.; Smit, A.; et al. Human-mouse alignments with BLASTZ. **Genome research**, v. 13, n. 1, p. 103–107, 2003.
Seleme, M. D. C.; Vetter, M. R.; Cordaux, R.; et al. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. **Proceedings**

of the National Academy of Sciences of the United States of America, v. 103, n. 17, p. 6611–6616, 2006.
She, X.; Rohl, C. A.; Castle, J. C.; et al. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. **BMC genomics**, v. 10, p. 269, 2009.
Sheen, F. M.; Sherry, S. T.; Risch, G. M.; et al. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. **Genome research**, v. 10, n. 10, p. 1496–1508, 2000.
Shemesh, R.; Novik, A.; Edelheit, S.; Sorek, R. Genomic fossils as a snapshot of the human transcriptome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 5, p. 1364–1369, 2006.
Smallwood, A.; Hon, G. C.; Jin, F.; et al. CBX3 regulates efficient RNA processing genome-wide. **Genome research**, v. 22, n. 8, p. 1426–1436, 2012.
Smallwood, S. A.; Kelsey, G. De novo DNA methylation: a germ cell perspective. **Trends in genetics : TIG**, v. 28, n. 1, p. 33–42, 2012.
Solyom, S.; Ewing, A. D.; Rahrmann, E. P.; et al. Extensive somatic L1 retrotransposition in colorectal tumors. **Genome research**, v. 22, n. 12, p. 2328–2338, 2012.
Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. **Molecular and cellular biology**, v. 21, n. 6, p. 1973–1985, 2001.
Srikantha, T.; Landsman, D.; Bustin, M. Retropseudogenes for human chromosomal protein HMG-17. **Journal of molecular biology**, v. 197, n. 3, p. 405–413, 1987.
Stajich, J. E.; Block, D.; Boulez, K.; et al. The Bioperl toolkit: Perl modules for the life sciences. **Genome research**, v. 12, n. 10, p. 1611–1618, 2002.
Steiper, M. E.; Young, N. M. Primate molecular divergence dates. **Molecular phylogenetics and evolution**, v. 41, n. 2, p. 384–394, 2006.
Stewart, C.; Kural, D.; Strömberg, M. P.; et al. A comprehensive map of mobile element insertion polymorphisms in humans. **PLoS genetics**, v. 7, n. 8, p. e1002236, 2011.
Tay, Y.; Kats, L.; Salmena, L.; et al. Coding-Independent Regulation of the Tumor Suppressor PTEN by Competing Endogenous mRNAs. **Cell**, v. 147, n. 2, p. 344–357, 2011. Elsevier Inc.
Torrents, D.; Suyama, M.; Zdobnov, E.; Bork, P. A genome-wide survey of human pseudogenes. **Genome research**, v. 13, n. 12, p. 2559–2567, 2003.
Turner, J. M. A. Meiotic sex chromosome inactivation. **Development (Cambridge, England)**, v. 134, n. 10, p. 1823–1831, 2007.
Ueda, S.; Nakai, S.; Nishida, Y.; Hisajima, H.; Honjo, T. Long terminal repeat-like elements flank a human immunoglobulin epsilon pseudogene that lacks introns. **The EMBO journal**, v. 1, n. 12, p. 1539–1544, 1982.
Ullu, E.; Tschudi, C. Alu sequences are processed 7SL RNA genes. **Nature**, v. 312, n. 5990, p. 171–172, 1984.
Vanin, E. F. Processed pseudogenes: characteristics and evolution. **Annual review of genetics**, v. 19, p. 253–272, 1985.
Vanin, E. F.; Goldberg, G. I.; Tucker, P. W.; Smithies, O. A mouse alpha-globin-related pseudogene lacking intervening sequences. **Nature**, v. 286, n. 5770, p. 222–226, 1980.
Venter, J.; Adams, M.; Myers, E.; Li, P.; Mural, R. The Sequence of the Human Genome. **Science (New York, N.Y.)**, 2001.

Vinckenbosch, N.; Dupanloup, I.; Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 9, p. 3220–3225, 2006.
Walsh, C. P.; Chaillet, J. R.; Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. **Nature genetics**, v. 20, n. 2, p. 116–117, 1998.
Wang, J.; Song, L.; Grover, D.; et al. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. **Human mutation**, v. 27, n. 4, p. 323–329, 2006.
Watanabe, T.; Totoki, Y.; Toyoda, A.; et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. **Nature**, v. 453, n. 7194, p. 539–543, 2008.
Wei, W.; Gilbert, N.; Ooi, S. L.; et al. Human L1 retrotransposition: cis preference versus trans complementation. **Molecular and cellular biology**, v. 21, n. 4, p. 1429–1439, 2001.
Wheeler, D. A.; Srinivasan, M.; Egholm, M.; et al. The complete genome of an individual by massively parallel DNA sequencing. **Nature**, v. 452, n. 7189, p. 872–876, 2008.
Whitcomb, J. M.; Hughes, S. H. Retroviral reverse transcription and integration: progress and problems. **Annual review of cell biology**, v. 8, p. 275–306, 1992.
Wilde, C. D.; Crowther, C. E.; Cowan, N. J. Diverse mechanisms in the generation of human beta-tubulin pseudogenes. **Science (New York, N.Y.)**, v. 217, n. 4559, p. 549, 1982.
Witherspoon, D. J.; Marchani, E. E.; Watkins, W. S.; et al. Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions. **Human heredity**, v. 62, n. 1, p. 30–46, 2006.
Wu, T. D.; Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. **Bioinformatics (Oxford, England)**, v. 26, n. 7, p. 873–881, 2010.
Xiong, Y.; Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. **The EMBO journal**, v. 9, n. 10, p. 3353–3362, 1990.
Yano, Y.; Saito, R.; Yoshida, N.; et al. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. **Journal of molecular medicine (Berlin, Germany)**, v. 82, n. 7, p. 414–422, 2004.
Yu, Z.; Morais, D.; Ivanga, M.; Harrison, P. M. Analysis of the role of retrotransposition in gene evolution in vertebrates. **BMC bioinformatics**, v. 8, p. 308, 2007.
Zhang, J.; Espinoza, L. A.; Kinders, R. J.; et al. NANOG modulates stemness in human colorectal cancer. p. 1–9, 2012. Nature Publishing Group.
Zhang, Q. The role of mRNA-based duplication in the evolution of the primate genome. **FEBS letters**, v. 587, n. 21, p. 3500–3507, 2013.
Zhang, Y.; Ryder, O. A.; Zhang, Y. Genetic divergence of orangutan subspecies (Pongo pygmaeus). **Journal of molecular evolution**, v. 52, n. 6, p. 516–526, 2001.
Zhang, Z. D.; Cayting, P.; Weinstock, G.; Gerstein, M. Analysis of nuclear receptor pseudogenes in vertebrates: how the silent tell their stories. **Molecular biology and evolution**, v. 25, n. 1, p. 131–143, 2008.
Zhang, Z.; Harrison, P.; Gerstein, M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. **Genome research**, v. 12, n. 10, p. 1466–1482, 2002.

Zhang, Z.; Harrison, P. M.; Liu, Y.; Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. **Genome research**, v. 13, n. 12, p. 2541–2558, 2003.
Zhang, Z.; Carriero, N.; Gerstein, M. Comparative analysis of processed pseudogenes in the mouse and human genomes. **Trends in genetics : TIG**, v. 20, n. 2, p. 62–67, 2004.

Apêndice(s)

Tabela S1. Retrocópias humano específicas.

Parental	Cromossomo	Início	Fim	Fita
HS6ST1	chr1	21754761	21758727	+
AIDA	chr1	78275455	78277620	-
VOPP1	chr1	148193179	148196113	+
RPL21	chr1	212224809	212225386	-
RPL23A	chr1	228262	228787	-
PNRC2	chr1	118319133	118321531	-
RAP1GDS1	chr1	144043970	144044174	-
RAP1GDS1	chr1	206506677	206506881	-
SMS	chr1	160864679	160866356	+
RPL7	chr1	97144339	97145196	-
NUDT4	chr1	145136108	145139946	-
GCSH	chr1	168024597	168025731	-
MORF4L1	chr1	220426792	220428562	+
PHKA1	chr1	91358549	91359470	-
ARID3B	chr1	81967270	81971426	+
RAP1GDS1	chr10	44983781	44983971	-
DYNC1I2	chr10	52024724	52027095	+
RPL13A	chr10	98510023	98510680	+
SRP9	chr10	93565800	93567290	-
CKS1B	chr10	29986864	29987649	+
FAM36A	chr10	70392094	70392593	+
GNG10	chr11	10292761	10293835	-
CSNK2A1	chr11	11373322	11374847	-
RPL36A	chr11	16996193	16996591	-
RPL26L1	chr11	2356365	2357013	-
ZNF283	chr11	128441186	128441335	-
PYROXD1	chr11	106694627	106698744	-
BMPR1A	chr11	121231079	121234010	-
RPS28	chr11	82400570	82400967	+
FABP5	chr11	59548556	59549224	-
DNAJB6	chr11	127810157	127811705	-
RPL18A	chr12	104659052	104659669	+
PGAM1	chr12	104424519	104426269	+
RPL41	chr12	93477059	93477492	-
AK4	chr12	31766180	31769517	-
PGAM1	chr12	94033662	94035392	-
RPL14	chr12	63359082	63359804	+
UHRF1	chr12	20704503	20707317	+
PHC1	chr12	55803470	55808727	-
RPS26	chr13	101192070	101192526	-
DGKZ	chr13	44542459	44545843	+
RBM8A	chr14	60864425	60867232	+
RPL3	chr14	99439638	99439817	+
RPL3	chr14	99439148	99439638	-
BNIP3	chr14	28733635	28735166	+
NANOG	chr15	35375427	35377509	-
RNF145	chr15	23499340	23501553	-
RNF145	chr15	20831936	20834149	-
RPL21	chr16	9250199	9250778	-
HNRNPA1	chr16	51679667	51681027	+
RAB43	chr16	46656773	46660897	-
NPIPL3	chr16	22545319	22547539	+
PAIP1	chr17	18553680	18556015	+
SDHC	chr17	1760573	1761755	-
FOXO3	chr17	18569236	18576494	-

Parental	Cromossomo	Início	Fim	Fita
TWF1	chr17	27528564	27531563	-
AK4	chr17	29672334	29675762	+
RPS2	chr17	19349226	19350188	-
RPS26	chr17	43685906	43686369	+
RPS7	chr17	26794796	26795581	+
DHFR	chr18	23747811	23751321	-
RPL6	chr18	6462091	6463028	-
ITGB1	chr19	14732345	14733056	-
PLEKHA3	chr19	42026596	42028737	-
TPM3	chr19	42011984	42014101	+
RPSA	chr19	24009927	24010921	+
PRR13	chr19	40448540	40449651	-
EIF3F	chr2	58478564	58479841	+
RPS28	chr2	232120779	232121182	+
C20orf30	chr2	51056580	51056913	-
FAM49B	chr2	170514642	170517867	-
MXRA7	chr2	162196011	162197865	-
RPL22	chr2	108531345	108533690	+
C14orf126	chr2	170361263	170361514	+
H3F3A	chr2	175584528	175585583	+
HNRNPC	chr2	190787895	190789612	+
EIF3E	chr2	165430251	165430530	+
VDAC2	chr2	65432212	65433412	+
BAK1	chr20	31276721	31278856	-
MPPE1	chr22	22239623	22240328	-
RPL41	chr22	36234312	36234744	-
HMGB2	chr3	22423307	22424093	+
TCEA1	chr3	37317028	37319650	+
PBX2	chr3	142894904	142898107	+
TMEM183B	chr3	149699448	149701153	-
PSMC1	chr3	68684836	68686393	+
ARMC10	chr3	94224397	94226494	-
METTL15	chr3	156429192	156432814	-
C1orf106	chr3	111902197	111904847	-
RPL22	chr3	169201007	169201808	-
HNRNPA3	chr3	75263613	75264906	+
RPS26	chr4	114135112	114135576	-
MTRF1L	chr4	189659525	189663178	+
TECR	chr4	87870690	87871257	-
CDC42	chr4	22728045	22729646	-
RAC1	chr4	46725687	46726624	-
RPLP0	chr5	165809310	165809691	+
CKS1B	chr5	61807580	61808309	-
PSMC1	chr5	106530856	106531031	+
FAM133B	chr5	60670885	60672859	+
RAP1B	chr5	75465910	75470179	-
RPL31	chr5	59725670	59726129	-
RPL10	chr5	168043316	168044051	+
RPL41	chr5	55240443	55240878	+
HMGN2	chr5	75537026	75538227	-
FAM103A1	chr6	166998987	167000150	-
RPL23A	chr6	171054595	171055067	+
RPL29	chr6	118320091	118320745	+
EIF4H	chr7	27495991	27498476	-
RPS26	chr7	122321347	122321816	+
RPL21	chr7	20042348	20042915	+
EEF1A1	chr7	22549936	22551681	-
RWDD4	chr7	39892296	39894859	-

Parental	Cromossomo	Início	Fim	Fita
CDC26	chr7	129049694	129050391	-
RPS26	chr8	101907975	101908432	+
LSM12	chr8	35381097	35383292	-
ZNF322	chr9	99957633	99962427	-
RPS26	chr9	9090877	9091323	+
RALGAPA1	chr9	108282023	108290006	+
SLC4A1AP	chr9	30558588	30559479	-
RPL9	chrX	23854740	23855463	-
FAM45A	chrX	129629107	129631549	+
FAM3C	chrX	23093703	23096507	+
ANKRD11	chrX	145700249	145702282	-
GAPDH	chrY	21489384	21490475	+
SFPQ	chrY	15206830	15209635	+
CTBP2	chrY	59001390	59002804	+

Tabela S02. X Movimentos intercromossomais de retrocópias e retrocópias expressas.

Humano - RETROCÓPIAS			
Direção	Esperado	Observado	*p-valor*
Do X	283	357	7,28E-06
Dos autossomos	7126	7052	
Para X	354	500	1,83E-15
Para autossomos	7055	6909	
Humano - RETROCÓPIAS EXPRESSAS			
Direção	Esperado	Observado	*p-valor*
Do X	48	85	5,15E-08
Dos autossomos	1202	1165	
Para X	60	60	1
Para autossomos	1190	1190	
Chimpanzé - RETROCÓPIAS			
Direção	Esperado	Observado	*p-valor*
Do X	292	256	4,21E-02

Dos autossomos	6774	6810	
Para X	336	446	7,80E-10
Para autossomos	6732	6622	
Chimpanzé - RETROCÓPIAS EXPRESSAS			
Direção	Esperado	Observado	*p-valor*
Do X	57	75	1,49E-02
Dos autossomos	1328	1310	
Para X	66	64	0,8008
Para autossomos	1319	1321	
Gorila - RETROCÓPIAS			
Direção	Esperado	Observado	*p-valor*
Do X	291	362	2,14E-05
Dos autossomos	6793	6722	
Para X	338	458	2,25E-11
Para autossomos	6746	6626	
Gorila - RETROCÓPIAS EXPRESSAS			
Direção	Esperado	Observado	*p-valor*
Do X	54	80	3,02E-04
Dos autossomos	1259	1233	
Para X	63	69	0,4385
Para autossomos	1250	1244	
Orangotango - RETROCÓPIAS			
Direção	Esperado	Observado	*p-valor*

Do X	283	379	5,41E-09
Dos autossomos	6261	6165	
Para X	337	468	2,35E-13
Para autossomos	6207	6076	
Orangotango - RETROCÓPIAS EXPRESSAS			
Direção	Esperado	Observado	*p-valor*
Do X	33	72	3,96E-12
Dos autossomos	740	701	
Para X	40	56	0,009379
Para autossomos	733	717	
Rhesus - RETROCÓPIAS			
Direção	Esperado	Observado	*p-valor*
Do X	317	337	2,50E-01
Dos autossomos	6716	6696	
Para X	378	526	5,06E-15
Para autossomos	6655	6507	
Rhesus - RETROCÓPIAS EXPRESSAS			
Direção	Esperado	Observado	*p-valor*
Do X	55	79	9,28E-04
Dos autossomos	1168	1144	
Para X	66	66	1
Para autossomos	1157	1157	

Lista de Anexos

SÚMULA CURRICULAR

Fábio Cassarotti Parronchi Navarro
Limeira - 18/05/1984

EDUCAÇÃO

1999/2001
Organização Einstein de Ensino - Limeira - SP
Ensino médio normal e técnico em processamento de dados

2004/2009
Universidade Federal de São Carlos - São Carlos - SP
Engenharia de Computação
Graduação

2010/Atual
Universidade São Paulo - São Paulo - SP
Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)
Doutorado Direto

OCUPAÇÃO

2010-2014
Bolsista de Doutorado, CAPES - ProEx

PUBLICAÇÕES

Schrider, D. R.* ; **Navarro, F. C. P.*** ; Galante, P. A. F. ; Parmigiani, R. B. ; Camargo, A. A. ; Hahn, M. W. ; De Souza, S. J. . **Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans**. PLOS Genetics (Online), v. 9, p. e1003242, 2013.

Navarro, F. C. P. ; Galante, P. A. F. . **RCPedia: a database of retrocopied genes.** Bioinformatics, v. 29, p. 1235-1237, 2013.

Kroll, J. E. ; Galante, P. A. F. ; Ohara, D. T. ; **Navarro, F. C. P.** ; Ohno-Machado, L. ; De Souza, S. J. . **SPLOOCE: A new portal for the analysis of human splicing variants**. RNA Biology, v. 9, p. 1339, 2012.

Galante, P. A. F.; Parmigiani, R. B.; Zhao, Q.; Caballero, O. L.; de Souza, J. E.; **Navarro, F. C. P.;** Gerber, A. L.; Nicolas, M. F.; Salim, A. C. M.; Silva, A. P. M.; Edsall, L.; Devalle, S.; Almeida, L. G.; Ye, Z.; Kuan, S.; Pinheiro, D. G.; Tojal, I.; Pedigoni, R.

G.; de Sousa, R. G. M. A.; Oliveira, T. Y. K.; de Paula, M. G.; Ohno-Machado, L.; Kirkness, E. F.; Levy, S.; da Silva, W. A.; ,Vasconcelos, A. T. R.; Ren, B.; Zago, M. A.; Strausberg, R. L.; Simpson, A. J. G.; De Souza, S. J.; Camargo, A. A.; **Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual**. Nucleic Acids Research, v. 39, p. 6056-6068, 2011.

Navarro F. C. P., Galante P. A. F.
A genome-wide landscape of retrocopies in primate genomes (submetido).

Donnard E. R.; Carpinetti P. A.; **Navarro F. C. P.**; Perez R. O.; Habr-Gama A.; Parmigiani R. B.; Camargo A. A.; Galante P. A. F. **ICRmax: an optimized approach to detect tumor-specific InterChromosomal Rearrangements for Clinical Application.** (submetido).

Donnard E. R. ; Asprino P. F.; Correa B.; Bettoni F.; Koyama F. ; **Navarro F. C. P.**; Perez R. O.; Mariadason J.; Siebe O.; Straussberg R.; Simpson A. J.G.; de Souza S. J.; Reis L. F. L.; Jardim D. L.F.; Parmigiani R. B.; Galante P. A.F.; Camargo A. A. **Mutational analysis of genes coding for cell surface proteins in colorectal cancer reveal novel altered pathways, druggable mutations and mutated epitopes for targeted therapy.** (submetido).

PARTICIPAÇÃO EM EVENTOS

Human Evolution - EMBO Conferences. Polymorphic retrotransposition of mRNAs in human and primates. 2014. (Congresso).

X-meeting. Evolution impact of a primate specific ACSL3 retrocopy inserted on Transferrin. 2013. (Congresso).

Cold Spring Harbor - The Biology of Genomes. A genome wide landscape of retrocopied protein-coding genes in primate genomes. 2013. (Congresso).

X-meeting. RetroCNVs a germinative source of genomic variation on human populations. 2012. (Congresso).

X-meeting. RetrogenesDB: A database of retrogenes in eukaryotes. 2011. (Congresso).

2nd São Paulo School of Translational Science. Detection of Genome-wide Structural Variation using Next Generation Sequencing Data. 2011. (Congresso).

Gordon Research Conference on Human Genetics & Genomics. RetrogeneDB: A database of retrogenes in eukaryotes. 2011. (Congresso).

X-meeting. Detection of Genome-wide Structural Variation using Next Generation Sequencing Data. 2010. (Congresso).

Fórum Internacional do Software Livre. Câncer: Como o processamento e armazenamento de dados distribuidos podem ajudar. 2010. (Palestra).

X-meeting. An *in silico* approach to select cancer/testis antigens genes in *mus Musculus*. 2009. (Congresso).

Fórum Internacional de Software Livre. Gambiarra: criando um jogo educativo em Python. 2008. (Palestra).

Oficina Desenvolvendo para o OLPC. 2007. (Oficina).

Fórum Internacional de Software Livre. 2007. (Outra).

Cell Broadband Engine Architecture Programming Workshop - Technical Briefings. 2006. (Oficina).

Fórum Internacional do Software Livre. 2005. (Outra).

PRÊMIOS RECEBIDOS

Travel Grant - Human Evolution Leicester. 2014.

Best Poster Award - X-Meeting 2013: "Evolution impact of a primate specific ACSL3 retrocopy inserted on Transferrin.", AB3C - Associação Brasileira de Bioinformática e Biologia Computacional. 2013.

Prêmio Viagem - IQ/USP. 2013.

Best Poster Award - X-Meeting 2012: "RetroCNVs a germinative source of genomic variation on human populations.", AB3C - Associação Brasileira de Bioinformática e Biologia Computacional. 2012.

Melhores estagiários, Portugal Telecom Inovação. 2008.

EXPERIÊNCIA ACADÊMICA

Biologia Molecular (Medicina). Monitor. 2012.

Biologia Molecular Computacional (Química). Monitor. 2011.

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 9 2013, pages 1235–1237
doi:10.1093/bioinformatics/btt104

Databases and ontologies

Advance Access publication March 1, 2013

RCPedia: a database of retrocopied genes

Fábio C. P. Navarro[1,2] and Pedro A. F. Galante[1,*]

[1]Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo 01308-060, Brazil and [2]Departmento de Bioquímica, Universidade de São Paulo, São Paulo 05508-000, Brazil

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Retrocopies are copies of mature RNAs that are usually devoid of regulatory sequences and introns. They have routinely been classified as processed pseudo-genes with little or no biological relevance. However, recent findings have revealed functional roles for retrocopies, as well as their high frequency in some organisms, such as primates. Despite their increasing importance, there is no user-friendly and publicly available resource for the study of retrocopies.

Results: Here, we present RCPedia, an integrative and user-friendly database designed for the study of retrocopied genes. RCPedia contains a complete catalogue of the retrocopies that are known to be present in human and five other primate genomes, their genomic context, inter-species conservation and gene expression data. RCPedia also offers a streamlined data representation and an efficient query system.

Availability and implementation: RCPedia is available at http://www.bioinfo.mochsl.org.br/rcpedia.

Contact: pgalante@mochsl.org.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 21, 2012; revised on February 20, 2013; accepted on February 22, 2013

1 INTRODUCTION

Retrocopies are gene copies that are generated by reverse transcription and genomic integration of transcribed mRNAs. Although retrocopies have been described since the early 1980s (Vanin, 1985), their functional roles have only recently been revealed (Ciomborowska *et al.*, 2013; McEntee *et al.*, 2011; Poliseno *et al.*, 2010). Retrocopies occur frequently in many genomes, including those of primates (Marques *et al.*, 2005), and some retrocopies are transcribed and have putative functions [see (Kaessmann *et al.*, 2009) for a review].

Interestingly, retrocopies have idiosyncrasies that simplify their identification. The four main characteristics are as follows: (i) an original multi-exonic parental gene copy in the genome; (ii) a mono-exonic region, without intronic regions; (iii) a poly-A stretch located in the 3′-most region; or (iv) direct repeats of 8–12 nucleotides (nt) flanking them [see (Kaessmann *et al.*, 2009) for a review]. These characteristics make retrocopy identification through computational pipelines reasonably straightforward, especially for species for which well-assembled genomes and transcriptomes are available.

Despite this, there is still a lack a publicly available and easy-to-use resources dedicated to the study of retrocopies (Kaessmann *et al.*, 2009), making it necessary either to use manual and multi-step approaches to explore retrocopies or to use non-specialized databases, such as the pseudogene databases (e.g. http://www.pseudogene.org/), that contain only basic and/or restricted information. Here, we describe RCPedia, a publicly available database that was developed for the study of retrocopies. RCPedia contains a myriad of information on retrocopied genes from six primate genomes (human, chimp, gorilla, orangutan, rhesus and marmoset), as well as a streamlined graphical data representation and an efficient information query system.

2 DATA RETRIEVAL AND CURATION

2.1 Data sources

The detection of retrocopies in eukaryotic genomes relies on two fundamental datasets: (i) a reference genome sequence and (ii) a set of known transcripts from each organism. The current version of RCPedia is based on genomic data from the UCSC Genome Browser (http://genome.ucsc.edu): human (hg19), chimpanzee (panTro3), gorilla (gorGor3), orangutan (ponAbe2), rhesus (rheMac2) and marmoset (calJac3). We used RefSeq sequences (http://www.ncbi.nih.gov/RefSeq) as the source of known transcripts, except for gorilla for which there are no RefSeq data. For gorilla, we used Ensembl transcripts (http://www.ensembl.org/). To evaluate retrocopy expression, we re-analysed the publicly available RNA-seq data from six tissues (brain, cerebellum, heart, liver, kidney and testis) of five primates (human, chimp, gorilla, orangutan and rhesus) (Brawand *et al.*, 2011).

2.2 Identifying orthologous retrocopies

The next step was to determine retrocopy conservation among the six primates. To avoid misidentification, we defined orthologous retroposition events based on conservation of the retrocopy and the flanking genomic regions. All retrocopies and their flanking regions (3 kb up- and downstream, without repetitive sequences) were aligned against the other primate genomes using BLAT [(Kent, 2002) with the following parameters: -mask = lower; tileSize = 12; -minScore = 50; -minIdentity = 0]. Only loci that matched the retrocopy and its flanking regions were considered as orthologous and, therefore, conserved.

2.3 Expression data

To detect retrocopies that were expressed, we developed a stringent multi-step pipeline. First, we searched for chimeric transcripts by analysing all intragenic retrocopies. We used GSNAP (parameters: -t 30; -B 4; –nofails; -A sam; -m 2; -n 1)

*To whom correspondence should be addressed.

to align all RNA-seq reads against genomic loci containing intragenic retrocopies (Wu and Nacu, 2010). Then, we selected only the alignments (alignment score >20) that showed two separated blocks (distance between blocks: >42 nt), where one read overlapped the retrocopy and the other aligned with the host gene. Alignments that were not defined by a canonical splicing site (GT-AG) were also filtered out. Intragenic retrocopies that contained at least five reads and showed this alignment pattern were considered to be expressed. Second, we searched for retrocopy expression *per se* by aligning all the reads against their respective genomes and transcriptomes. The alignment against the transcriptome data was important for removing false positive alignments derived from exon–exon junctions. Only unique genome matches (alignment score: >40) that were filtered by aligning them with the transcriptome data were used for gene expression analysis. At least five supporting reads were required for a retrocopy to be considered as expressed.

3 DATABASE IMPLEMENTATION

RCPedia is a database and a front-end interface. The database was build over MySQL (http://www.mysql.com). The website was developed mainly using PHP (http://www.php.net) based on CakePHP (http://cakephp.org) as the framework for the development of an efficient Model-View-Controller front-end. All genomic annotation and gene expression data were processed using Perl (http://www.perl.org) scripts developed in-house. Briefly, all coding transcripts from RefSeq (and Ensembl for gorilla) were downloaded and aligned against their respective reference genomes using BLAT [(Kent, 2002) with the following parameters: -mask = lower; -tileSize = 12; -minIdentity = 75; -minScore = 100]. All alignments were processed and sequences with >75% identity, and either a sequence alignment length >50% or, at least, 120 matched nucleotides, were selected. Based on the expected genomic characteristics for retrocopies, we designed a four-step strategy to identify them. First, any alignment containing gaps >15 kb in length was eliminated. This step eliminated transcripts with large (large) introns but kept retroelements, such as Long Interspersed Elements (LINEs) (~6 kb) and Short Interspersed Elements (SINEs) (<1 kb), that are frequently inserted inside retrocopied loci. Second, we retrieved the exon–exon boundary positions from the parental genes. Next, we mapped these boundary positions onto the retrocopies and searched for gaps between them. Putative retrocopy alignments that contained one or more gaps were excluded because they are unlikely to have been derived from retroduplications. Third, only gene copies that contained >50 nt from two or more exons of the parental genes were selected. Finally, we defined the retrocopy set by selecting all remaining alignments and, if necessary, grouping any alignments that were mapped onto the same genomic locus (Supplementary Fig. S1).

4 DATABASE QUERY INTERFACE AND OUTPUT VISUALIZATION

4.1 The query system

The RCPedia query system is easy-to-use, complete and fast. It includes gene (e.g. GAPDH), chromosome (e.g. chr17), genomic position orientation (e.g. chr17:28 102 500–29 112 200), gene alias (e.g. RAS) and gene annotation keyword (e.g. kinase or oncogene) searches, making it easy for the user to explore the genes and genomic locations that match their retrocopy events.

4.2 Results

Because there are many unnamed retrocopies, the search output results in RCPedia are based on parental gene names. The results of a query can be presented from two data visualization perspectives: (i) the parental gene perspective, which helps the user to visualize all retrocopied events of a given parental gene, as well as their genomic loci, and their identity to retrocopies, for example (for the full dataset, see the website) and (ii) the retrocopy perspective, which displays information, such as their genomic context, identity to the parental gene, conservation in other species, and retrocopy expression (see Supplementary Fig. S2 for a schematic view).

5 USING RCPedia

To show how RCPedia can be used, we selected the human gene DHFR as a sample query. RCPedia reported five retrocopies for DHFR in the human genome (Supplementary Fig. S2). Interestingly, one of the retrocopies was present only in the human genome. Another retrocopy was expressed in four human tissues (Supplementary Fig. S2), and it was reported previously that this locus is expressed and has a putative function (McEntee *et al.*, 2011).

6 CONCLUSION

RCPedia is a well-organized, user-friendly and streamlined graphical representation resource dedicated to the study of retrocopies in primate genomes. To the best of our knowledge, RCPedia is the most comprehensive and publicly available database in this field, although some resources providing similar information (Karro *et al.*, 2007; Khelifi *et al.*, 2005; Ortutay and Vihinen, 2008). We strongly believe that RCPedia will significantly improve the annotation and functional characterization of retrocopies present in primate genomes.

ACKNOWLEDGEMENTS

The authors thank A. A. Camargo, LFL Reis and all members of the Bioinformatics Group for suggestions. They are grateful to D. T. Ohara for helpful technical support.

Funding: PAFG was supported by FAPESP (2012/24731-1) and D43TW007015 from the Fogarty International Center, National Institutes of Health. FCPN was supported by CNPq fellowship. Funding to pay the Open Access publication charges was provided by Hospital Sírio-Liobanês.

Conflict of Interest: none declared.

REFERENCES

Brawand,D. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Ciomborowska,J. *et al.* (2013) "Orphan" retrogenes in the human genome. *Mol. Biol. Evol.*, **30**, 384–396.

Kaessmann,H. *et al.* (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.*, **10**, 19–31.

Karro,J.E. *et al.* (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.*, **35**, D55–D60.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Khelifi,A. *et al.* (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic Acids Res.*, **33**, D59–D66.

Marques,A.C. *et al.* (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.*, **3**, e357.

McEntee,G. *et al.* (2011) The former annotated human pseudogene dihydrofolate reductase-like 1 (DHFRL1) is expressed and functional. *Proc. Natl Acad. Sci. USA*, **108**, 15157–15162.

Ortutay,C. and Vihinen,M. (2008) Pseudogenequest - service for identification of different pseudogene types in the human genome. *BMC Bioinformatics*, **9**, 299.

Poliseno,L. *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.

Vanin,E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.*, **19**, 253–272.

Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

OPEN ACCESS Freely available online

PLOS | GENETICS

Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans

Daniel R. Schrider[1☯]*, Fabio C. P. Navarro[2,3,4☯]*, Pedro A. F. Galante[2,4], Raphael B. Parmigiani[2,4], Anamaria A. Camargo[2,4], Matthew W. Hahn[1], Sandro J. de Souza[2,5]

1 Department of Biology and School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States of America, **2** São Paulo Branch, Ludwig Institute for Cancer Research, São Paulo, Brazil, **3** Departamento de Bioquímica, Universidade de São Paulo, São Paulo, Brazil, **4** Centro de Oncologia Molecular–Hospital Sírio-Libanês, São Paulo, Brazil, **5** Brain Institute, Federal University of Rio Grande do Norte, Natal, Brazil

Abstract

The era of whole-genome sequencing has revealed that gene copy-number changes caused by duplication and deletion events have important evolutionary, functional, and phenotypic consequences. Recent studies have therefore focused on revealing the extent of variation in copy-number within natural populations of humans and other species. These studies have found a large number of copy-number variants (CNVs) in humans, many of which have been shown to have clinical or evolutionary importance. For the most part, these studies have failed to detect an important class of gene copy-number polymorphism: gene duplications caused by retrotransposition, which result in a new intron-less copy of the parental gene being inserted into a random location in the genome. Here we describe a computational approach leveraging next-generation sequence data to detect gene copy-number variants caused by retrotransposition (retroCNVs), and we report the first genome-wide analysis of these variants in humans. We find that retroCNVs account for a substantial fraction of gene copy-number differences between any two individuals. Moreover, we show that these variants may often result in expressed chimeric transcripts, underscoring their potential for the evolution of novel gene functions. By locating the insertion sites of these duplicates, we are able to show that retroCNVs have had an important role in recent human adaptation, and we also uncover evidence that positive selection may currently be driving multiple retroCNVs toward fixation. Together these findings imply that retroCNVs are an especially important class of polymorphism, and that future studies of copy-number variation should search for these variants in order to illuminate their potential evolutionary and functional relevance.

Citation: Schrider DR, Navarro FCP, Galante PAF, Parmigiani RB, Camargo AA, et al. (2013) Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans. PLoS Genet 9(1): e1003242. doi:10.1371/journal.pgen.1003242

Editor: Joshua M. Akey, University of Washington, United States of America

Received May 30, 2012; **Accepted** November 28, 2012; **Published** January 24, 2013

Funding: DRS is supported by National Institutes of Health Genetics, Cellular and Molecular Sciences Training Grant GM007757. FCPN is supported by a CAPES fellowship. MWH is supported by NSF grant DBI-0855494 and a fellowship from the Alfred P. Sloan Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dschride@indiana.edu (DRS); fnavarro@usp.br (FCPN)

☯ These authors contributed equally to this work.

Introduction

In recent years it has become apparent that changes in gene copy-number introduced by genomic duplication and deletion events are an important force driving adaptive evolution [1]. Examples of adaptive gene gains and losses have been found in a variety of organisms, including humans [2–4] and *Drosophila melanogaster* [5,6]. Much attention has focused on gene duplications in particular, as they may facilitate the evolution of new gene functions [7,8]. Given that all new gene duplicates must arise as polymorphisms, and the fact that genomic duplications and deletions can have negative phenotypic consequences [9–11], massive efforts have been made to identify regions of the genome differing in copy-number, referred to as copy-number variants (CNVs), among humans [2,12–15] and other species (e.g., refs. [16–18]). These studies have revealed extensive copy-number variation especially within humans, with any two African individuals differing in copy-number at over 100 genes [2,19].

It has been suggested that in humans the vast majority of gene duplications contributing to this variation result in a new copy located adjacent to the original gene [14]. However, a substantial number of new duplicates are inserted far from the original locus in humans and other mammals [20,21], including genes duplicated by retrotransposition [22,23]. These retrocopies, which are created when a messenger RNA transcript is reverse-transcribed and reinserted into a different location in the genome, are an especially interesting class of gene duplicate for several reasons. First, a new retrocopy will contain an entire coding sequence except when derived from an incomplete transcript. In addition, retrocopies occasionally carry promoter elements located downstream of the retrotranscribed transcript's transcription start site but located upstream of an alternative transcription start site [24]. Evidence that a substantial proportion of gene retrotransposition events result in functional gene copies, called retrogenes, come from both mammals [25,26] and *Drosophila* [27]. In addition, patterns of gene movement onto and off of the X chromosome in mammals and off of the X in *D. melanogaster* suggest that many retrogenes are subject to positive selection (e.g., refs. [28–30]). Finally, processed pseudogenes, inactivated gene copies created by retrotransposition, have also been shown to influence expression

Author Summary

Recent studies of human genetic variation have revealed that, in addition to differing at single nucleotide polymorphisms, individuals differ in copy-number at many regions of the genome. These copy-number variants (CNVs) are caused by duplication or deletion events and often affect functional sequences such as genes. Efforts to reveal the functional impact of CNVs have identified many variants increasing the risk of various disorders, and some that are adaptive. However, these studies mostly fail to detect gene duplications caused by retrotransposition, in which an mRNA transcript is reverse-transcribed and reinserted into the genome, yielding a new intron-less gene copy. Here we describe a method leveraging next-generation sequence data to accurately detect gene copy-number variants caused by retrotransposition, or retroCNVs, and apply this method to hundreds of whole-genome sequences from three different human subpopulations. We find that these variants account for a substantial number of gene copy-number differences between individuals, and that gene retrotransposition may often result in both deleterious and beneficial mutations. Indeed, we present evidence that two of these new gene duplications may be adaptive. These results imply that retroCNVs are an especially important class of CNV and should be included in future studies of human copy-number variation.

levels of the parental gene copy, potentially disrupting its function [31,32].

Despite the potentially important evolutionary and phenotypic consequences of retrogenes, current CNV-detection approaches are largely unable to find them. In fact, only one study of copy-number variation in humans was able to detect any polymorphic retrogenes [2]. Previously, we developed a method capable of leveraging next-generation sequence data to detect gene copy-number variants caused by retrotransposition, or retroCNVs, and used it to reveal that 13% of gene copy-number polymorphisms in *D. melanogaster* are caused by retrotransposition [30]. Although a similar method has been applied to detect retroCNVs in humans [33], there has been no detailed analysis of retroCNVs in humans to date. Here we apply an improved method to a number of sequenced human genomes, including data from the 1000 Genomes Project [34]. We find a surprising amount of variation due to retroCNVs within the human population—accounting for ~12 genes differing in copy-number between any two individuals. By comparing retroCNV patterns to retrogene divergence, we reveal that retrotransposition is an important source of both adaptive and deleterious mutations in humans. We also find evidence that some of these retroCNVs may currently be under positive selection in humans. These findings underscore the functional and evolutionary importance of gene duplication via retrotransposition, and suggest that further study of retrogenes will illuminate the extent to which these retroCNVs affect human phenotypes and drive adaptive evolution.

Results/Discussion

RetroCNVS are common in human populations

In order to detect polymorphic retrocopies of protein coding genes segregating in human populations, we searched for evidence of retrocopy insertion sites using sequence reads from two human genomes that we sequenced ourselves with the SOLiD technology (denoted AAC and SJS), and additional genomes from the 1000 Genomes Project [34]. Briefly, this approach works by searching for paired-end reads spanning insertion sites of retrocopies present in the reference genome but absent from a resequenced genome (Figure 1a), or vice-versa (Figure 1b). We also searched low-coverage genomes resequenced for the 1000 Genomes Project [34] for exon-exon junction-spanning reads indicative of retroCNVs (Figure 1c), similar to our previous approach [30]. Because the whole genome must be searched in order to discover retroCNV insertions absent from the reference genome, such retroCNVs were initially discovered using a smaller set of 17 individuals (Table S1; Materials and Methods). These retroCNVs were then genotyped using paired-end sequence data from three subpopulations from the 1000 Genomes Project: 52 Yoruban individuals in Nigeria (referred to as the YRI subpopulation), 41 individuals of European ancestry in Utah (referred to as CEU), and 56 Han Chinese individuals and Japanese individuals from Tokyo (referred to as ASI). Because of this ascertainment scheme, these retroCNVs are expected to be biased towards higher frequencies than if they were discovered using the entire set of sequenced genomes. RetroCNVs present in the reference genome were identified using paired-end reads from all individuals sequenced for the 1000 Genomes Project, and are therefore unaffected by any ascertainment bias. We correct for this difference in ascertainment schemes where necessary in the analyses presented here. We find that our computational approach for retroCNV identification has high specificity and sensitivity, allowing us to estimate the contribution of retrotransposition to gene copy-number polymorphism in humans.

We identified 91 retroCNVs in total, finding that these polymorphisms account for 11.9 genes differing in copy-number between any two African individuals on average. Given that a recent comparison of pairs of individual human genomes has revealed gene copy-number differences at 105 genes on average (based on data from ref. [2]), our results suggest that retroCNVs could account for a sizable minority of human gene copy-number polymorphisms (although retroCNVs may often be non-functional). We were able to determine the insertion sites of 39 retroCNVs (18 present in the reference genome; 21 absent from the reference), and verify that retrocopy presence was the derived state for each of these (Materials and Methods); the remaining 52 retroCNVs were identified from reads spanning exon-exon junctions only and therefore have unknown insertion loci. While many of these retrocopies may contain only fragments of coding sequence, perhaps due to the low processivity of reverse-transcriptase or partial degradation of the mRNA used as template, we found that at least 41.8% (accounting for ~6 complete gene copy-number differences between any two African genomes) of the retrocopies across all genomes are complete or near-complete retrogenes which may have the potential to be functional (see Materials and Methods).

To estimate the fraction of false positive retrogenes in our analysis, we attempted to validate all retroCNVs with known insertion sites by PCR amplification followed by sequencing. We confirmed 10 of 11 retroCNVs present in the reference genome (90.9%) that we were able to assay, and 17 of 21 (80.5%) retroCNVs absent from the reference genome. In the case of retroCNVs absent from the reference genome our experimental design does not allow us to differentiate between false positives and retroCNVs we could not amplify due to experimental difficulties such as low primer specificity (Materials and Methods), and most retroCNVs we could not amplify (whether present or absent in the reference) were flanked by repetitive elements. It therefore seems plausible that some or all of the four retroCNVs absent from the reference genome that we could not confirm are actually true positives. However, even if we conservatively assume that these

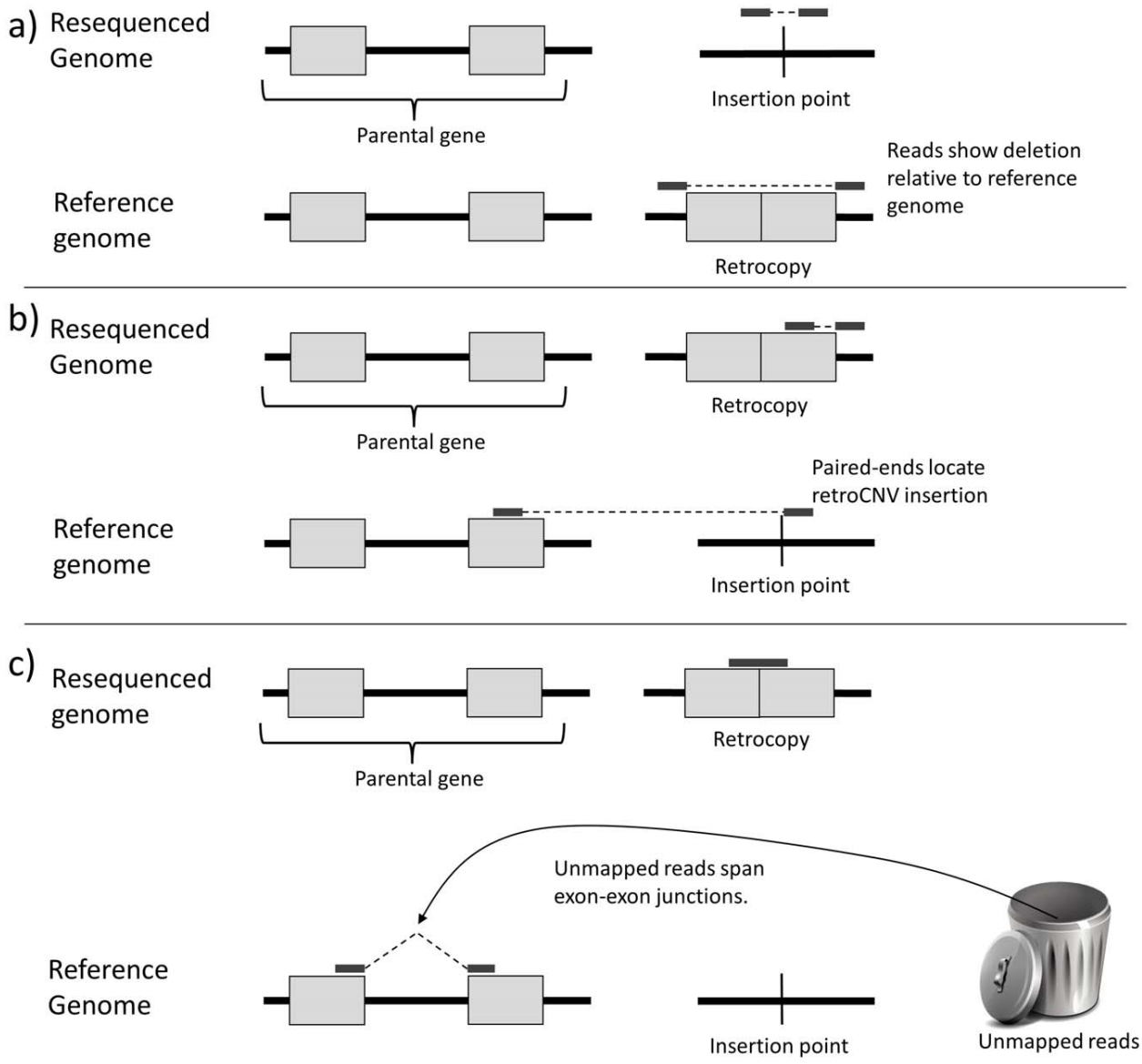


Figure 1. Detecting retroCNVs using sequence reads. a) RetroCNVs present in the reference genome are detected by searching for retrocopies in the reference that are absent from a sequenced individual, as revealed by paired-end reads spanning the location of the retroCNV and mapping too far apart from one another. b) RetroCNVs absent from the reference genome are detected by using paired-end reads to detect retroCNV insertion sites, and c) using reads that span exon-exon junctions but do not map to the reference genome.
doi:10.1371/journal.pgen.1003242.g001

four cases are false positives, our false positive rate across the set of 39 retroCNVs with known insertion loci is acceptably low (15.6%; validation results are listed in Table S2 and genomes used for validation are listed in Table S3). The remaining 52 retroCNVs may contain a higher fraction of false positives, and their relatively high fraction of singletons (67.3%) is consistent with this. However, we have previously shown that the exon-exon junction approach used to detect these retroCNVs is quite accurate [30]; thus, many of these 52 retroCNVs are likely true events, and the large number of singletons could in part be explained by somatic mutations in the cell lines used to obtain DNA for the individuals in the 1000 Genomes Project, in addition to false positives. In any case, the omission of these retroCNVs does not qualitatively affect any of the analyses described below. We estimate that the approach using paired-end reads to discover retroCNVs (whether present in or absent from the reference genome) was able to detect at least 77.4% of singleton retroCNVs inserted in non-repetitive sequence in the 17 discovery genomes. The false negative rate decreases dramatically for retroCNVs present more than once in the discovery set—we estimate that retroCNVs present in just two samples would be discovered ~95% of the time (Materials and Methods). In addition, the exon-exon junction approach has previously been shown to be highly sensitive [30]; this implies that our dataset contains the vast majority of retroCNVs present in the genomes we examined during the discovery phase of our study. All retroCNVs included in our dataset, and their insertion coordinates

when known, are listed in Table S2. The sets of genome sequences and retroCNVs included in each of our analyses are summarized in Table S4.

Insertion patterns of retroCNVs

In contrast to tandem duplications caused by replication slippage, or sometimes by non-allelic homologous recombination (NAHR), retrotransposition results in a new gene duplicate located far from the parental copy. Unlike our previous examination of gene retrotransposition in *D. melanogaster* [30], in this study we were able to locate the insertion site of new retrocopies and therefore to examine precise patterns of gene movement caused by this type of duplication. Although there is an excess of fixed retrogene movements onto and off of the human and mouse X chromosomes relative to expectations [29], we do not see such a pattern in our set of retroCNVs (Table 1), suggesting differences in the contribution of adaptive evolution to polymorphic and fixed retrogenes. As we have previously done in *D. melanogaster*, here we conducted a statistical test for differences in patterns of movement between retroCNVs and fixed functional retrogenes. If gene movements onto and off of the X are neutral, then we expect the same proportion of such events among polymorphic retrocopies and fixed functional retrogenes; however, if movements involving the X chromosome are often adaptive, then we will observe a higher fraction of this class of movements among fixed retrogenes. We do in fact find a significantly higher fraction of fixed functional retrogenes than retroCNVs moving to and from the X chromosome ($P = 0.0067$; Fisher's exact test using fixed retrogene data from ref. [29]), lending further support to the hypothesis that natural selection is driving gene movement to and from mammalian X chromosomes [29]. This result remains significant when we only examine retroCNVs discovered in females ($P = 0.0079$), and is therefore not an artifact of reduced power to detect X-linked retroCNVs in males. Because retroCNVs absent from the reference genome were discovered using a different ascertainment scheme than retroCNVs present in the reference genome, combining them in this analysis could impact our results. However, this would only result in a deficit of retroCNVs moving to or from the X chromosome if such retroCNVs were more likely to be confined to lower allele frequencies by purifying selection than other retroCNVs, and there is no reason to expect such a difference in selective pressures. Moreover, after imposing the same ascertainment scheme on both retroCNVs present in and absent from the reference genome (Materials and Methods), we observe a similar but non-significant deficit of retroCNVs moving to or from the X (none of the 9 retroCNVs in this set involve movements to or from the X; $P = 0.11$). When we test separately for an excess of fixed functional retrogenes moving off of the X or moving onto the X, we do not see significance in either case ($P = 0.150$ for movements off of the X; Table S5; $P = 0.0650$ for movements onto the X; Table S6). However, although we have lower statistical power in these comparisons, we do observe trends suggestive of natural selection. Moreover, the excess of fixed functional retrogenes moving off of the X is significant when we compare retroCNVs to data from ref. [35] ($P = 0.0077$; Table S5); when we examine all retroCNVs, including those with an unknown insertion site, we also see a significant excess of fixed retrogenes originating on the X chromosome when comparing our data to both ref. [29] and ref. [35] ($P = 0.032$ and $P = 3.6 \times 10^{-4}$ respectively; Table S7). Combined with the observation that processed pseudogenes do not exhibit a bias of movement from the X [29], our data strongly suggest that natural selection is responsible for the excess of functional retrogenes moving off of the X chromosome in mammals, and perhaps onto the X chromosome as well. These observations could be the result of positive selection driving the fixation of new functional retrogenes moving to or from the X, selection to maintain such genes once they are established, or both of these mechanisms.

Table 1. RetroCNVs versus fixed retrogenes moving from an autosome to an autosome (A→A) from the X chromosome to the X (X→X), from the X to the autosomes (X→A), or vice-versa (A→X).

	RetroCNVs	Fixed retrogenes*
A→A or X→X	36	70
A→X or X→A	3	29

*Data from Emerson et al. [29].

doi:10.1371/journal.pgen.1003242.t001

While it is widely believed that gene duplicates created by retrotransposition are almost always dead-on-arrival pseudogenes because they do not carry all regulatory elements from the parental copy with them, it has been shown that a retrocopy inserted into another gene will often exploit that gene's regulatory machinery in order to be expressed [26]. We therefore examined the insertion point of our retroCNVs to determine how many were inserted into existing genes. We found that over one-half (20 of 39) of retroCNVs were inserted into genes, with all but one of these retroCNVs being inserted into an intron (Table S2). This does not represent a significant deviation from what one would expect if retrocopy insertions were distributed uniformly across the genome, as introns make up roughly 40% of the human genome ($P = 0.60$; χ^2 test). Although there does not appear to be a strong bias in polymorphism data, we compared retroCNVs to the 7,831 retrocopies (functional or otherwise) identified in the reference genome (Materials and Methods), nearly all of which are fixed, and found a deficit of fixed human retrocopies in introns compared to retroCNVs: 50.0% of retroCNVs versus 31.8% of fixed retrocopies are found in introns (Table 2; $P = 0.022$; Fisher's exact test; $P = 0.012$ using fixed retrocopies from ref. [26] with $d_S < 0.1$ when compared to their parent gene). Again, similar to the reasoning laid out above, this implies that retrocopies inserted into introns are often deleterious, as was suggested by Vinckenbosch et al. [26]. Indeed, the results in Table 2 suggest that roughly one-half of intronic retrocopy insertions are eliminated by purifying selection. A similar deficit of fixed intronic retrocopies is observed when we impose the same ascertainment scheme on all retroCNVs, as described in Materials and Methods (62.5% of retroCNVs found in introns versus 31.8% of fixed retrocopies), although this comparison is no longer significant ($P = 0.12$), perhaps in part due to diminished statistical power. Because this is a comparison of patterns of retroCNVs that may not be functional to fixed retrocopies that are mostly pseudogenes, the simplest interpretation of this result is that the insertion of

Table 2. RetroCNVs versus fixed retrocopies inserted in intronic versus intergenic sequence.

	RetroCNVs	Fixed retrocopies
Intronic insertions	19	2,492
Intergenic insertions	19	5,339

doi:10.1371/journal.pgen.1003242.t002

retrocopies into genes may often be deleterious even when the inserted retrocopy is non-functional. Thus, intronic insertions may often be deleterious regardless of the content of the inserted sequence. This interpretation is supported by the observation that tandem duplications occurring within introns are often subject to purifying selection in *Drosophila* [17].

If the above interpretation is correct, then it could imply that roughly half of the genic retroCNVs we detect here are deleterious and would not be allowed by selection to reach fixation. This interpretation is substantiated by the lower allele frequencies of intronic versus intergenic retroCNVs when examining only retroCNVs present in the reference genome (avg. frequency in YRI is 0.46 for intronic and 0.72 for intergenic retroCNVs; $P = 0.75$) or absent from the reference genome (0.11 for intronic versus 0.16 for intergenic; $P = 0.95$). We performed this comparison separately for retroCNVs present and absent from the reference genome in order to control for ascertainment bias, as these retroCNVs had different ascertainment schemes. While these differences are not significant, they are consistent with selection acting against intronic insertions, especially given evidence that non-retroCNV insertions within introns are often deleterious as discussed above. Consistent with this interpretation, it has been noted that fixed retrocopy insertions are less likely to be intronic than expected if retrocopies are inserted with uniform probability across the genome [26], although there is evidence of an insertion bias associated with chromatin accessibility in *Drosophila* [36]. Overall, there is substantial evidence that insertions of retrocopies or other sequence into introns are often deleterious.

Since one would presume that retrocopies inserted into introns are also more likely to be expressed, our results suggest that retrotransposition could be an important source of new functional gene copies as well as potentially deleterious mutations. An additional possible functional consequence of the insertion of retroCNVs into introns is the formation of sense-antisense pairs, as we previously suggested [37]. Consistent with this possibility, we find that 10 of 20 retrocopies inserted into another gene are on that gene's minus strand (Table S2). We also find that one retroCNV, a copy of RPL3, switches strands mid-sequence, most likely due to 5′ inversion during retrotransposition [38].

Segregating chimeric genes created by retrotransposition

Another interesting consequence of the insertion of a retrocopy into an intron of a host gene is the possibility of chimeric transcription of the host and the retrocopy. Chimeric genes are likely an important source of new gene functions [39], and the large fraction of retroCNVs inserted into introns suggests that retrotransposition could be an important source of these genes. Indeed, there are several known cases of retrotransposition resulting in functional chimeric genes in humans [40,41] and *Drosophila* [6,42,43], with some of these genes showing evidence for adaptive evolution [6,44].

In order to search for evidence of chimeric transcripts among the 20 retroCNVs inserted within existing genes, we examined RNA-seq data from lymphoblast tissues from 60 HapMap individuals of European descent [45]. We found that 20% (4 of 20) of these retroCNVs show evidence of chimeric expression. The chimeric transcript *CBX3-C15orf57*, where the *CBX3* retroCNV is inserted in-between the second and third exons of *C15orf57*, shows evidence of expression as a chimera in 20 individuals. The chimeric combination *SDHC-RPA1* forms a sense-antisense pair, with *SDHC* inserted in-between the fifth and sixth exon of *RPA1*; the chimeric transcript is expressed in 6 individuals. *UQCR10-C1orf194*, in which *UQCR10* is inserted into the second exon of *C1orf194* is expressed in a single individual. An examination of the sequencing read confirming the validity of this retroCNV reveals that the *UQCR10* portion of this transcript is not in proper reading frame. The *RPL18A-TXNRD1* combination, in which *RPL18A* is inserted in-between the third and fourth exons of *TXNRD1*, was also found to be expressed in one individual. We also found evidence of chimeric transcripts derived from SKA3-DDX10 in a breast cancer cell line and in a lymphoid cell line (HCC1954 and HCC1954-BL from ref. [46]), both derived from an individual genotyped for *SKA3*. The *SKA3* retroCNV is inserted in-between the tenth and eleventh exons of *DDX10*, forming a sense-antisense pair.

Because three of these chimeric transcripts involve either a sense-antisense pair or the retroCNV apparently being inserted out of reading frame, they may be nonfunctional and perhaps deleterious. Alternatively, it has been suggested that chimeric transcripts could result in novel protein coding regions even if they are not in sense-sense orientation or proper reading frame [25]. In addition, we have only examined expression data for chimeric transcripts from lymphoblast cell lines for the majority of our retroCNVs, and two additional cell lines for a single retroCNV (*SKA3*; Materials and Methods), and may therefore be underestimating the number of segregating chimeric genes caused by the incorporation of retroCNVs into existing genes. While further work is required to determine the number of these new genes and their functional consequences, our results suggest that retrotransposition could be a source of evolutionary novelty creating not only new gene duplicates but new genes with potentially novel functions.

Evidence that positive selection may be acting on retroCNVs

In order to examine the population dynamics of retroCNVs, we used both insertion presence/absence information at retroCNV insertions and evidence of retrotransposition from exon-exon junction-spanning reads to genotype 39 retroCNVs whose insertions we were able to locate. After estimating allele frequencies for these retroCNVs in three human populations (Materials and Methods), we noticed that several had very high derived-allele frequencies (Figure 2; frequencies listed in Table S2). While this observation is consistent with positive selection driving retroCNVs to fixation, the fact that many of our retroCNVs were ascertained in a sample of 17 genomes (AAC, SJS, and 15 individuals from the 1000 Genomes Project) biases our frequency spectra towards higher frequency variants. We therefore searched for more direct evidence of adaptive natural selection acting on individual retroCNVs. Although previous genome-wide studies of copy-number variation have searched for evidence of natural selection sweeping duplications towards fixation [2,14], these searches were conducted at regions containing the parental copy and not necessarily the daughter copy. This was because location of the daughter locus was not known, and was simply assumed to be proximate to the parental locus. These approaches would therefore fail to detect evidence of positive selection on dispersed duplications, a limitation that does not affect our analysis because we have identified the exact location of the new duplicates. Conversely, if the insertion sites of duplicates are not known, many previous studies of ongoing selective sweeps in humans [47,48] may have detected the signature of positive selection on an inserted sequence that was not known to lie in the selected region.

In addition to examining the correct locus, testing for adaptive evolution requires accurate genotyping. We therefore genotyped all 39 retroCNVs with known insertion sites as homozygous for retroCNV presence, heterozygous, or homozygous absent using

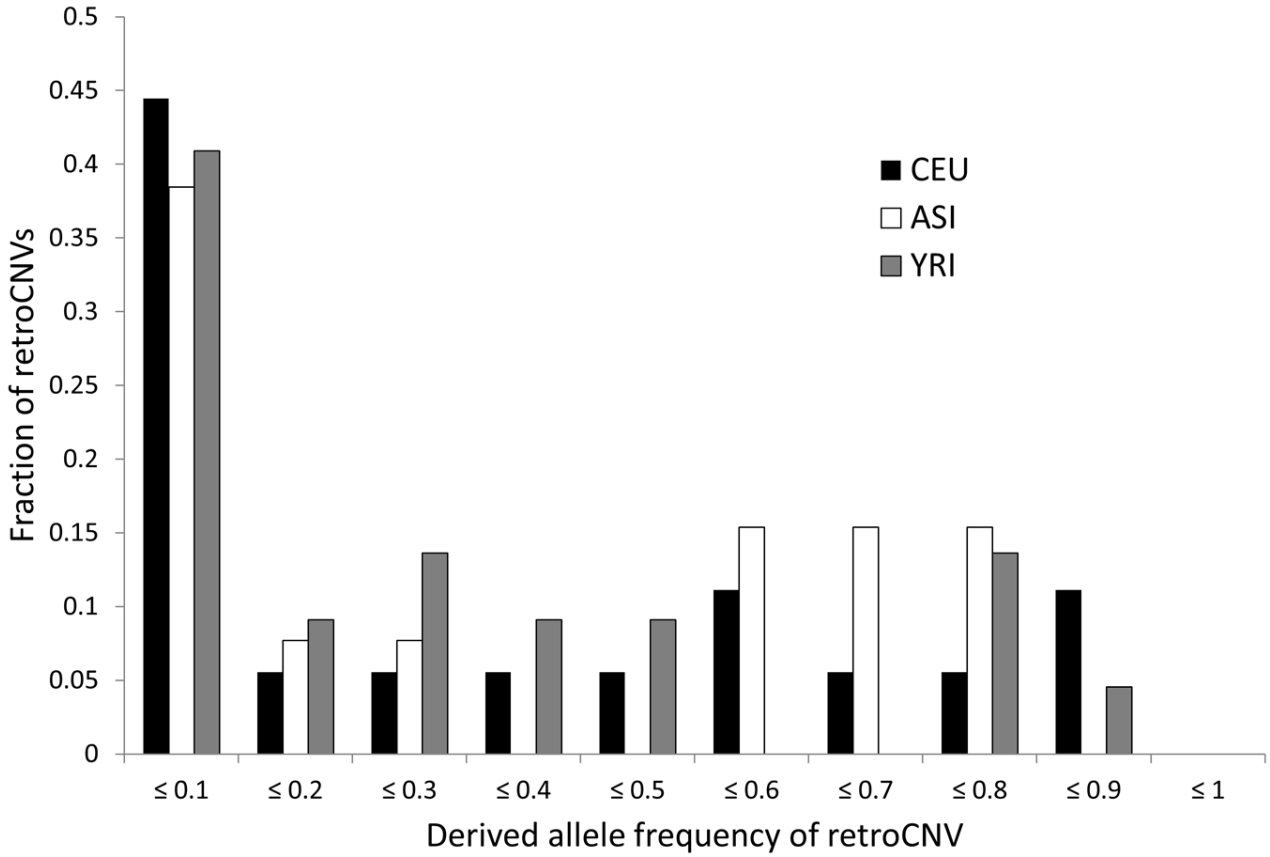


Figure 2. Estimated derived allele frequencies of retroCNVs segregating in three human subpopulations. Allele frequencies were calculated as described in the Materials and Methods. RetroCNVs fixed in or absent from a given subpopulation are not shown.
doi:10.1371/journal.pgen.1003242.g002

our short-read sequences. In order to assess our genotyping accuracy, we initially compared our genotyping results for the retroCNV of *DHFR* to those of Conrad et al. [2], who were able to genotype this retroCNV as well. We found that our genotypes agreed for 100% of individuals genotyped as homozygous for retroCNV presence by Conrad et al., for 85% of individuals genotyped as heterozygous, and for 98% of individuals genotyped as homozygous absent. Because Conrad et al. [2] may have committed genotyping errors as well, these percentages can be thought of as a lower bound on our genotyping accuracy, suggesting that our genotyping is highly accurate. In order to gain additional confidence in our genotyping accuracy, we analyzed the genotypes of two available trios from the 1000 Genomes Project, finding that no analyzed retroCNVs violated Mendelian inheritance (Table S8), although these genomes had higher coverage than the rest of our data set. In addition, we experimentally validated the genotypes of *DHFR* and *GNG10* (discussed below) in 36 individuals (Table S3) and found that our genotyping is also accurate in genomes with lower coverage, with 94.4% and 91.7% of genotyping calls confirmed for these two retroCNVs, respectively. At these two retroCNVs we correctly genotyped 85.3% of heterozygous individuals and 100% of homozygotes, similar to our results in comparison to those of Conrad et al. [2].

The action of positive selection on an allele results in a rapid increase in the frequency of the haplotype containing the selected allele in the population. The swift nature of this rise in frequency results in a decrease in genetic diversity among chromosomes containing the selected allele compared to neutral expectations. We therefore examined nucleotide diversity (π) in regions flanking retroCNV insertions, finding several retroCNVs with a marked reduction in diversity among haplotypes containing the retroCNV relative to the other haplotypes in the population (Materials and Methods). However, a deficit of diversity is expected among haplotypes sharing a derived allele regardless of its selective importance [49]. With this in mind, we used coalescent simulations [50] to ask whether the ratio of π among haplotypes containing a retroCNV to π among haplotypes lacking it, which we refer to as π_{der}/π_{anc}, was lower than expected under neutrality (Materials and Methods). This is similar to the haplotype-based test first suggested by Hudson et al. [51], the sole difference being that we contrast π between the derived and ancestral allelic classes, rather than the number of segregating sites. For a polymorphism segregating in the absence of selection, we expect the observed ratio of π_{der}/π_{anc} to be typical when compared to those generated from the neutral coalescent for derived alleles of the same sample frequency. For a polymorphism sweeping to fixation, on the other hand, relatively little diversity is expected among chromosomes containing the selected allele that is rapidly rising in frequency, and this allelic class would therefore exhibit a lower π_{der}/π_{anc} ratio than polymorphisms of the same frequency simulated under neutrality.

We were able to perform this test on 17 retroCNVs in the CEU subpopulation, 16 in YRI, and 13 in ASI (Materials and Methods). Two retrocopies are candidates for positive selection according to this test: the retrocopy of *DHFR* appears to be experiencing positive selection in individuals of European descent ($P = 0.0083$;

Figure 3), as does a retrocopy of *GNG10* in both Europeans ($P = 0.0094$; Figure S1) and Africans ($P < 1.1 \times 10^{-4}$; Figure S2). If we correct for multiple testing by conservatively assuming that all 46 tests for selection that we conducted were independent—even though many tests were of the same retroCNVs but in different subpopulations—the false discovery rate (FDR) for the *DHFR* and *GNG10* retroCNVs in Europeans is 0.14, while the FDR for the *GNG10* retroCNV is 0.0051 in Africans. As stated above, a deficit of diversity is expected within haplotypes containing a new mutation under the neutral coalescent. However, this deficit is less pronounced for polymorphisms with relatively high derived-allele frequencies such as the *DHFR* and *GNG10* retroCNVs because the amount of diversity associated with any allele is proportional to its frequency. The reductions in heterozygosity shown in Figure 3, Figure S1, and Figure S2 may therefore be suggestive of positive selection; this interpretation is supported by the results of our coalescent-based test that takes allele frequency into account.

The *DHFR* retroCNV, previously discovered by Anagnou et al. [52], is inserted into the fourth intron of *PSM8*, forming a sense-antisense pair. The ORF of this retrocopy perfectly matches that of the parental copy of *DHFR* in the reference genome [53]. *DHFR* codes for dihydrofolate reductase, deficiency of which causes megaloblastic anemia and neurological disease [54], and is required for nucleotide synthesis [55]. *DHFR* has an important role in cell growth, and its inhibition has been used in antibacterial [56] and antitumor drugs [57]. This retrocopy also exhibited a similar reduction in nucleotide diversity in the Asian subpopulation, although this pattern was not significant by our test ($P = 0.099$; Figure S3). *GNG10*, which has been associated with melanoma [58], has a retrocopy that forms a sense-sense pair with *SBF2*, which has been implicated in Charcot-Marie-Tooth disease [59]. To gain further confidence in these results, we compared the π_{der}/π_{anc} ratios observed for these candidates to those calculated from random regions flanking SNPs with similar derived allele frequencies, finding that relatively few SNPs in the human genome exhibited lower π_{der}/π_{anc} ratios than these retroCNVs, even though some of these loci are likely themselves under positive selection. For example, just 2.5% and 5.5% of loci in the genome exhibited lower ratios of π_{der}/π_{anc} than the *DHFR* retroCNV in Europeans and the *GNG10* retroCNV in Africans, respectively (Materials and Methods).

Although we experimentally determined that our genotype calls at these two retroCNVs were quite accurate, genotyping error could still affect the analyses described above. We therefore conducted a further test based on integrated haplotype scores (iHS), a statistic designed to detect extended haplotypes characteristic of ongoing sweeps, around these two retroCNV insertions [48]. Importantly, this test is not dependent on our genotype assignments. We find that only 1.2% of random genomic regions exhibit stronger biases toward extreme iHS values than the region containing the *GNG10* retroCNV in Africans, the strongest candidate identified by our coalescent-based test (Materials and Methods). Additionally, only 5.7% of random genomic regions exhibit more extreme iHS values than the *DHFR* retrocopy in Asians, where we observed a suggestive but non-significant signal of selection in our coalescent-based test. We cannot know with certainty that natural selection is responsible for the patterns of diversity around these two retroCNVs, or that the retroCNVs themselves rather than polymorphisms in linkage disequilibrium with them are the targets of any such selection. Nonetheless, our findings that the haplotypes containing these retroCNVs exhibit reduced diversity and reside within regions identified by an

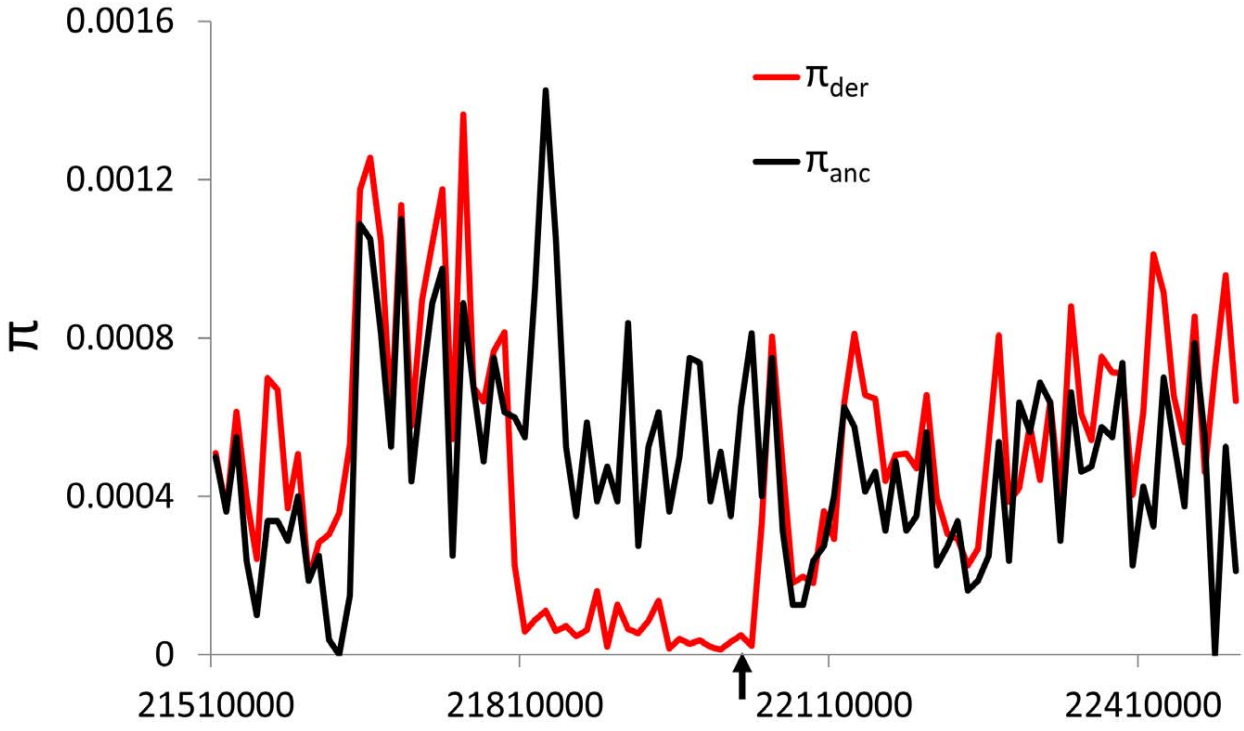


Figure 3. Reduced nucleotide diversity on chromosome 18 among chromosomes containing the *DHFR* retroCNV in CEU. π is shown in 10 kilobase windows for chromosomes containing the *DHFR* retroCNV (red) and those lacking this retroCNV (black). The location of the retroCNV insertion is marked by an arrow. While there is little difference in nucleotide diversity distal to the retroCNV, there is a recombination hotspot in that region (data from ref. [65]).
doi:10.1371/journal.pgen.1003242.g003

extended haplotype test suggest that these retroCNVs should be considered candidates for adaptive natural selection. This evidence that multiple retroCNVs currently segregating in human subpopulations could potentially confer an increase in fitness suggests that retrotransposition could be an important source of adaptive alleles in humans.

Conclusions

Given the evolutionary significance of gene retrotransposition in humans and other species, we sought to examine the extent of gene copy-number variation caused by retroCNVs in human subpopulations. This effort resulted in the first set of gene duplication polymorphisms caused by retrotransposition in humans obtained from next-generation sequence data. Experimental validation shows that our methodology has high sensitivity and precision. These data reveal that retroCNVs are quite common, accounting for roughly a dozen gene copy-number differences between any two African genomes on average. Our data also provide direct evidence that gene retrotransposition events are often adaptive. First, a comparison of retroCNV insertion patterns with fixed retrogenes supports the hypothesis that the excess of retrogenes moving onto and off of the X chromosome during mammalian evolution is driven by natural selection [29]. Moreover, our high genotyping accuracy combined with our ability to locate the insertion sites of many common retroCNVs allowed us to detect signatures of natural selection acting on these variants. We find evidence that at least two retroCNVs detected in this study may be affected by adaptive natural selection. Indeed, because we may not have perfect power to detect all polymorphisms under positive selection, we may be underestimating the fraction of retroCNVs undergoing selective sweeps. This result implies that retrotransposition could be an important force driving ongoing human adaptation.

We also find that many retroCNVs are inserted into the introns of existing genes. While we find that these retroCNVs are less likely to reach fixation than intergenically inserted retrocopies and may therefore often be deleterious, these retroCNVs are more likely to be expressed [26]. Moreover, five particularly interesting cases of this type of retroCNV result in a chimeric transcript consisting of sequence from the retroCNVs and the gene in which it was inserted. Given that chimeric genes can have important functional consequences [44], and that we are very likely underestimating the fraction of chimeras among retroCNVs, retrotransposition could be an important source of chimeric proteins with the potential to perform novel functions. Taken together, these results imply that gene retrotransposition has been and may continue to be an important source of adaptive alleles in humans, and could be an underappreciated source of mutations with negative phenotypic consequences as well.

Materials and Methods

Data sources

The human genome reference sequence (hg19/GRCh37) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/). Gene models and transcript sequences of protein-coding genes were downloaded from version 57 of Ensembl [60]. Human reference mRNA sequences were downloaded from NCBI Reference Sequence project (http://www.ncbi.nlm.nih.gov/RefSeq/). Alignments, raw sequences, and unmapped reads from resequenced whole genomes were obtained from the 1000 Genomes Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/). We also sequenced two individual human genomes using the SOLiD3 platform; DNA samples from these individuals were donated to the Tumor Bank from the Hospital Alemão Oswaldo Cruz in São Paulo, Brazil after informed consent was obtained. These sequences were aligned to the reference genome using the mapping/pairing pipeline from BioScope (v3.1; http://www.solidbioscope.com/) with default parameters. The sets of individual genomes and retroCNVs examined in each phase of our analysis are listed in Table S4. Additionally, RNA-seq (paired-end) data from 60 HapMap individuals [45] were searched for evidence of chimeric transcripts.

Sequencing two individual human genomes

The two individuals sequenced here (AAC and SJS) filled out consent forms and donated DNA to the Tumor Bank from the Hospital Alemão Oswaldo Cruz; this databank was approved by the Hospital's Institutional Review Board. Twenty micrograms of genomic DNA were sheared using HydroShear to generate fragments with an average size of 2.0 kb. DNA fragments were then repaired to generate blunt ends and ligated to adaptors. DNA fragments of 2–3 kb were size-selected in agarose gels and subsequently circularized by ligation of a biotinylated internal adaptor. After removing non-circularized fragments, circularized DNA was treated with DNA polymerase I for nick-translation, followed by digestion with T7 exonuclease and S1 nuclease, which generated tags longer than 50 bp from the adaptor edges. Digested products were ligated with P1 and P2 adaptors, purified and amplified with 12 PCR cycles. A total of 96 picograms of the resulting library were then used for emulsion PCR. Approximately 300 million beads from each library were deposited on one slide, followed by 50 bp mate-pair sequencing on a SOLiD3 instrument, according to the manufacturer's protocol.

Identification of retroCNVs present in the reference genome

In order to detect retroCNVs present in the human reference genome, we first identified retrocopies present in the reference using a pipeline consisting of four steps: i) We aligned all human RefSeq transcripts to the human genome reference sequence; ii) All alignments overlapping multi-exon genes or the gene of the transcript's origin were removed. iii) Intronless alignments containing at least two exons from the parental gene, and exons mapped adjacently (without gaps) were selected; iv) Finally, we grouped sequences mapping to the same genomic region and removed putative retrocopies appearing to arise from genomic duplication. Using this approach, we found 7,831 retrocopies, which is similar to the number found in other databases, such as pseudogene.org (www.pseudogene.org) and Hoppsigen (http://pbil.univ-lyon1.fr/databases/hoppsigen).

In order to determine whether any of these 7,831 were retroCNVs segregating in humans, we downloaded alignments for all individuals from the 1000 Genomes Project that had whole-genome paired-end data and examined paired-end reads lying within 5 kb of a retrocopy. Paired-end reads that mapped further apart from one another than expected (indicative of a deletion) and that spanned a retrocopy without overlapping it were kept as evidence of a retroCNV (Figure 1a). Putative retroCNVs spanned by more than five paired-end reads were examined, and those not appearing to be artifacts of misalignment were subjected to experimental validation.

Identification of retroCNVs absent from the reference genome using reads at insertion sites

In order to detect retroCNVs not present in the human reference genome we examined paired-end read alignments from

15 individuals from the 1000 Genomes Project (Table S1), including two high-coverage parent-offspring trios. Examining these genomes and the genomes of AAC and SJS, we searched for paired-ends with one read mapped entirely within exonic sequence of a known gene (the putative parental gene) and the other read mapped to a distinct genomic region: i.e. on a different chromosome or on the same chromosome with a mapping distance higher than the average insert size of the paired-end library (a putative retroCNV insertion site; Figure 1b). We then removed insertion sites located within 2 kb of known retrocopies as they may represent alignment artifacts, insertion points overlapping retrotransposons (defined by RepeatMasker), and insertion sites supported by five or fewer non-redundant paired-end reads mapping to exonic regions of a single parental gene. All 39 candidates containing an insertion site were manually curated to remove those resulting from alignment artifacts, and subjected to experimental validation (for details, see "Experimental validation of retroCNVs" below).

Identification of retroCNVs absent from the reference genome using reads from exon–exon junctions

In order to search for additional retrotransposition events in low-coverage human genomes, we aligned unmapped reads from low-coverage genomes from the 1000 Genomes Project (the same genomes from ref. [34] used in the genotyping step described below) to human transcript sequences using BWA with default parameters (similar to the approach described in ref. [30]). Only Illumina and 454 reads were included in this analysis, as we noticed that the shorter SOLiD reads used in the 1000 Genomes Project introduced an extremely high number of false positives. Reads mapping across exon-exon junctions within these transcripts were taken as initial evidence of retrotransposition (Figure 1c). In particular, a gene was considered retrotransposed if there was i) at least one read in at least one individual spanning an exon-exon junction with at least 10 bp of the read crossing the junction, or ii) at least two distinct reads with different sequences (whether in the same individual or not) with at least 5 bp crossing an exon-exon junction. We only considered alignments having no more than 4% mismatches, and no more than 0.2*min(r,l) mismatches, where r and l are the number of bases in the read mapping to the left and right sides of the exon-exon junction, respectively. We used BLAT [61] to search for exon junction sequences (20 bp on either side of the junction) and to determine which of these junctions had partial or complete matches in the reference genome with the potential to introduce false positives. We removed from the analysis junctions with a BLAT hit in the reference genome with at least 90% identity and 10 bp on either side of the junction mapping to the reference genome. BLAT hits spanning the junction by less than 10 bp were kept in the analysis, but the number of base pairs spanning the junction was added to the mapping cutoffs required for calling retrogenes as described above. For example, if an exon-exon junction mapped to the reference genome with 7 bp of the match spanning the junction, two reads would need at least 12 bp spanning the junction, or one read would need at least 17 bp spanning the junction in order to call a retroCNV. All aligments reporting a putative retrotransposed gene were examined manually and reproduced using BLAT, and alignments that could be explained by reasons other than a retrotransposition event (e.g. reads mapping to the reference genome with a few mismatches) were removed.

In order to find the insertion site of retroCNVs identified using the exon-exon junction approach, all alignments for each of the individuals with whole-genome sequences from the 1000 Genomes Project were downloaded and paired-end reads with one read mapped to the 5′ or 3′ exon of a putative parental gene were extracted. Since genome coverage for most of these individuals is low, we merged all reads from these individuals and then selected insertion sites supported by more than five paired-end reads summing across individuals. For this analysis we have also excluded: i) insertion sites related to two or more parental genes; ii) insertion sites located within 2 kb of known retrocopies; iii) and insertion points overlapping retrotransposons. Insertion sites were manually curated in order to remove those resulting from misalignment.

Controlling for different ascertainment schemes

RetroCNVs present in and absent from the reference genome have different ascertainment schemes, with retroCNVs present in the reference genome discovered by examining all sequenced individuals in our data set and retroCNVs absent from the reference discovered in a smaller discovery set, or from exon-exon junction spanning reads (Table S4). Ascertainment bias could therefore affect observed patterns of retroCNV insertions when these two sets of retroCNVs are combined. We therefore repeated our comparisons of fixed and polymorphic retrocopies with respect to X versus autosomes and introns versus intergenic regions after imposing the same ascertainment scheme on both retroCNVs present in and absent from the reference. This ascertainment scheme required a retroCNV to have support for the non-reference allele from more than five read-pairs in at least one of the 17 discovery genomes (Table S1), and ignored evidence from exon-exon junction spanning reads. Note that this ascertainment scheme is more stringent for both retroCNVs present in and absent from the reference genome, and therefore the number of retroCNVs discovered is reduced substantially. When comparing allele frequencies of intronic and intergenic retroCNVs, we simply performed the analysis separately for retroCNVs present in the reference genome and retroCNVs absent from the reference genome, thereby preventing differences in ascertainment from affecting the results. The results of our coalescent-based tests for selection are not affected by ascertainment bias as each test is conditioned on the observed allele frequency of the retroCNV being tested.

Genotyping retroCNVs in human populations

We performed *in silico* genotyping for our complete set of retroCNVs identified using all three methods: from the reference genome absent in sequenced individuals, from paired-ends supporting insertion sites absent from the reference genome, and from exon-exon junction-spanning reads. These retroCNVs were genotyped in CEU (n = 41 unrelated individuals with paired-end data), YRI (n = 52), and ASI (CHB+JPT; n = 56) individuals with Illumina paired-end sequence data generated for the 1000 Genomes Project [34]. Genotyping proceeded as follows: for the set of retroCNVs present in the reference genome, we searched for paired-end reads for which one read mapped to the retroCNV itself and the other read mapped to the genomic region flanking the retroCNV (evidence of retroCNV presence). We also searched for paired-end reads spanning (without overlapping) the retroCNV regions (evidence of retroCNV absence). For the set of retroCNVs not present in the reference genome, we searched for paired-end reads for which one read mapped to the exonic region of a parental gene and the other read mapped to the insertion point of the retroCNV (evidence of retroCNV presence). We also searched for paired-end reads mapping to both sides of the insertion point and presenting the expected distance and orientation (evidence of retroCNVs absence). Heterozygous individuals were identified as those exhibiting evidence for both retroCNV presence and

absence. Reads spanning exon-exon junctions by 5 bp (plus any additional bases required due to partial matches of the exon junction in the reference genome as described above) were also used for determining whether a retroCNV was present in a given individual. For each of these strategies only one supporting read or read-pair was required for genotyping. For one gene, *CACNA1B*, heterozygotes could not reliably be distinguished from homozygotes. Allele frequencies were calculated for this retroCNV from the fraction of individuals with the presence allele (whether heterozygous or homozygous), in the same manner as the other 38 retroCNVs for which the insertion was located (see below). This retroCNV was omitted from tests for positive selection.

Assessing the completeness of retroCNV sequences

RetroCNVs were considered complete or nearly complete if the retrocopy contained at least part of the 5′-most and 3′-most exons in the retroposed transcript. For retroCNVs present in the reference genome, we simply examined the sequence of the retrocopy. For retroCNVs absent from the reference genome, all isoforms of the parental gene that could potentially have been reverse-transcribed given the exons known to be present in the retrocopy from exon-exon junction-spanning reads and read-pairs mapping to insertion sites were examined.

Estimating allele frequencies of retroCNVs

Because low coverage may cause our genotyping approach to undercall heterozygotes, and because we cannot distinguish homozygotes from heterozygotes using exon-exon junctions, we estimated the fraction of individuals containing each retroCNV (whether homozygous or heterozygous). This fraction, f, was calculated as the number of individuals with evidence of a retroCNV divided by the total number of individuals with evidence of either presence or absence of the retroCNV. We then estimated allele frequencies by assuming Hardy-Weinberg equilibrium: if f is the fraction of individuals with the retroCNV, $f = p^2 + 2pq$, and $1 - f = q^2$. Therefore, $q = (1-f)^{1/2}$ and $p = 1-(1-f)^{1/2}$. Note that retroCNVs with very high allele frequencies (i.e., with no individuals homozygous absent) will be incorrectly estimated as having an allele frequency of 1 although they are truly polymorphic with p approaching 1. Because we could not detect evidence of absence for retroCNVs with no detected insertion sites, we restricted allele frequency analyses to the 39 retroCNVs for which we could locate the insertion. These frequency estimates were used to compare allele frequencies of intronic and intergenic retroCNV insertions. Because exon-exon junction-spanning reads can produce evidence of retroCNV presence but not absence, potentially biasing allele frequency estimates, we repeated this comparison after omitting these data and verified that this bias did not qualitatively affect our results. In order to estimate the number of pairwise differences in retroCNV copy-number in the YRI subpopulation, we included retroCNVs genotyped by exon-exon junction spanning reads only, treating individuals with no evidence of retroCNV presence as homozygous absent, and calculating f as above, then estimating p and q and taking the summation of 2pq for each retroCNV.

Although it seems unlikely that any of these retroCNVs are caused by deletions of genes recently retrotransposed, we nonetheless polarized each of the 39 retroCNVs with a known insertion locus by using BLAT [61] to search for a retrocopy in the syntenic location of the chimpanzee genome as identified by liftOver [62]. Using this approach we confirmed that the presence of the retrocopy was indeed the derived allele for each of these 39 retroCNVs.

Experimental validation of retroCNVs

We attempted to validate all 39 retroCNVs with known insertion sites via PCR and DNA sequencing. For retroCNVs not present in the reference genome we designed primer pairs with one matching the parental gene sequence and one matching the insertion site sequence; this will yield a PCR product only when the retroCNV is present. We therefore cannot differentiate between false positives and cases where we could not amplify due to experimental difficulties. Indeed, two retroCNVs we attempted to amplify, *CACNA1B* and *FOXK2*, yielded numerous PCR products of different sizes and may lie within regions difficult to amplify with specificity and may not necessarily be false positives. Nonetheless, we conservatively report a false positive rate that assumes retroCNVs absent from the reference genome and yielding no clear PCR product are false positives. For retroCNVs present in the reference genome, we designed primers spanning the daughter (i.e. newly inserted) copy. In this case, both true and false positives should yield PCR products, and the sequence of the product is used to distinguish true positives from false positives. Thus, false positives are not confused with PCR failures. For larger retroCNVs, it is possible that primer pair spanning the insertion site may not reliably amplify across the retrocopy. In such cases, we designed an additional primer pair with one primer within the retrocopy and one primer in the flanking insertion sequence to identify retroCNV presence, while the primer pair spanning the insertion site was used to identify retroCNV absence. Primers for PCR were designed based on the reference genome sequence (hg19/GRCh37) using the Primer3 [63] and Oligotech (Oligos Etc., Eugene, OR) software packages. PCR reactions were carried out in a 25 μL reaction containing 50 ng of genomic DNA, 1× Taq DNA polymerase buffer (Invitrogen), 0.1 mM dNTP, 1 mM MgCl2, 1 unit Taq DNA polymerase (Invitrogen) and 6 pmol of each forward and reverse primer. Amplification conditions were: initial denaturation for 4 min at 94°C followed by 35 cycles of 45 sec at 94°C, 45 sec at 58°C and 1 min at 72°C and a final extension of 10 min at 72°C. PCR products were analyzed on 1% agarose gels and sequenced using the Big Dye Terminator kit (Applied Biosystems) and an ABI3100 Prism sequencer. The sequenced product was then examined to determine if it was consistent with the validation status indicated by the presence and/or size of the PCR product. The genomes used to validate these retroCNVs are listed in Table S3. These same genomes and methods were used to validate genotype calls for the *GNG10* and *DHFR* retroCNVs, using DNA from genomes listed in Table S3. DNA samples from all of these genomes were obtained from the Coriell Cell Repository (http://ccr.coriell.org).

Identification of chimeric transcripts containing retroCNVs

In order to detect chimeric expression of retroCNVs we downloaded paired-end alignments of RNA-Seq data from 60 European individuals (including 39 of the 41 Europeans in our data set) from ref. [45] and searched for read-pairs with unambiguous alignments where one read mapped to an exon of the retroCNV's parent gene (or the retrocopy itself if present in the reference genome) and the other read mapped to an exon of the gene in which the retroCNV was inserted. Only chimeric transcripts supported by 5 reads or more were considered, and only retroCNVs inserted into a known gene were included in this analysis.

We also tested for the expression of a chimeric transcript formed by the SKA3 retroCNV and its host gene, DDX10, using a pair of primers designed in SKA3 (5′ TCCCTCAGAAAAAGCTATGGTG 3′) and in DDX10 (5′ TCAAGGAGAGTGAT-

GATTC 3′). Total RNA was extracted using Trizol following the manufacturers' instructions (Invitrogen) and RNA integrity was analyzed using agarose gels. Reverse transcription was carried out using the Superscript III First Strand Synthesis Kit (Invitrogen). RT-PCR reactions were carried out in a 25 µl reaction mixture containing 1 µl of cDNA, 2.5 µl Taq DNA polymerase buffer, 0.1 mM dNTPs, 6.0 pmol of each, 1.0 mM MgCl2, and 1 U Taq DNA polymerase (Invitrogen). PCR conditions were as follows: 4 min at 94°C (initial denaturation), 35 cycles of 45 s at 94°C, 45 s at 58°C, and 1 min at 72°C, with a final extension step of 10 min at 72°C. RT-PCR products were analyzed on 8% silver-stained polyacrylamide gels. Sequencing reactions were carried out using DYEnamic (ET Terminator Cycle Sequencing Kit, Amersham Pharmacia) and an ABI 3130XL sequencer (Applied Biosystems). This experiment was performed in four cell lines: two from a single individual previously genotyped for the SKA3 retrogene [46], and two negative controls.

Estimating the false-negative rate of retroCNV discovery using paired-ends

In order to estimate an upper bound on the fraction of retroCNVs that we could not discover in the 17 genomes from the discovery set using paired-ends (AAC, SJS, and 15 individuals from the 1000 Genomes Project), we examined 10 fixed retrocopies present in the reference genome. Since these retrocopies are always homozygous present, we doubled the number of required read-pairs in order to detect a retroCNV as present (simulating the discovery of a heterozygous retroCNV). From these data we estimate the fraction of singletons (retroCNVs present in one of the 17 genomes, or 1/34 chromosomes, examined to discover retroCNVs with this method) our approach would fail to detect—a conservative upper bound on our false negative rate. This fraction can be used to estimate the fraction of retroCNVs present in i chromosomes in our discovery set by simply raising it to the i^{th} power.

Searching for positive selection around retroCNV insertions

In order to test for positive selection acting on retroCNVs, we first downloaded SNP genotype data for all SNPs within 100 kb of the insertion point for each retroCNV segregating in the CEU, YRI, and ASI subpopulations. Next, we inferred the haplotypic phase of each of these retroCNVs and their flanking SNPs by running fastPhase [64] with default parameters. RetroCNV genotype data from insertion sites were included as fastPhase input, with modifications in two cases involving retroCNVs absent from the reference genome. First, if a retroCNV was genotyped as homozygous absent in an individual from insertion site-spanning paired-end reads, but exon-exon junction spanning-read data from that same individual supported the presence of the retroCNV, the genotype was set to heterozygous for retroCNV presence. Second, if no paired-end reads were available for genotyping an individual and exon-exon junction data supported retroCNV presence, the individual was genotyped as having the retroCNV on one chromosome, and as having an unknown genotype on the other.

By examining the position homologous to insertion sites in the chimpanzee genome, we found that all of our insertions were derived. Our test for selection then asks whether there is a significantly lower value of π, the average number of pairwise differences per site, within the set of haplotypes having the retroCNV ($\pi_{derived}$) compared to the set of haplotypes lacking the retroCNV ($\pi_{ancestral}$), controlling for differences in allele frequencies [51]. We took the ratio of these measures, which we refer to as π_{der}/π_{anc}, as our test statistic. In order to determine if there was less nucleotide diversity in the set of haplotypes containing the retroCNV than is expected under neutrality, we performed 10,000 coalescent simulations using ms [50] with the same number of polymorphisms observed within 100 kb on either side of the retroCNV (plus one additional polymorphism taking the place of the retroCNV), and the same number of chromosomes as in the real sample. For these simulations, we assumed a single, flat recombination rate given by the region flanking the retroCNV insertions, as estimated from HapMap Phase II data [65]. For the CEU and ASI populations, a demographic model involving a bottleneck was used (using ms parameters -eN 0.05 0.5 -eN 0.15 1.5), and for YRI a recent population expansion was used (-eN 0.0 1.5). We then examined whether there was any polymorphism within the medial 25% of the simulated region having the same derived allele frequency as the retroCNV such that the ratio of π within haplotypes containing the derived allele to π within haplotypes containing the ancestral allele was less than or equal to the ratio calculated by partitioning the observed data according to alleles at the retroCNV. We calculated the P-value as the fraction of these simulated polymorphims meeting this criterion. This test was performed for each retroCNV segregating in each subpopulation in which at least two chromosomes contained the retroCNV and two chromosomes lacked it. We were able to test 17 retroCNVs in the CEU subpopulation, 16 in YRI, and 13 in ASI.

In order to determine whether candidate retroCNVs identified by this approach were also outliers compared to other polymorphisms segregating in humans, we compared the observed π_{der}/π_{anc} ratios to those calculated from non-overlapping 200 kb windows of SNPs from the 1000 Genomes data (http://www.1000genomes.org/). For each 200 kb window in each population, we calculated π_{der}/π_{anc} for up to one SNP lying within 10 kb of the center of the window and having a derived allele frequency landing in the same 5% bin as that of the retroCNV. We then calculated the fraction of these SNPs having π_{der}/π_{anc} less than or equal to that of the retroCNV for candidates for positive selection.

As an alternative method to search for evidence of positive selection in regions containing retroCNVs, we downloaded integrated haplotype scores (iHS) from ref. [48] and compared the density of high-|iHS| SNPs in regions containing retroCNVs to random genomic regions. Regions with a high density of high-|iHS| SNPs have previously been used as evidence of positive selection [48]. High-|iHS| SNPs were defined as those with iHS scores within either the upper or lower 2.5% tail of the empirical distribution of iHS scores from that same population. Within the retroCNV region, extended by 50 kb on each side, we counted the fraction of SNPs with high |iHS|, and calculated a χ^2 statistic comparing this fraction to the 0.05 expectation. We then repeated this test within 10,000 genomic regions of the same size, counting the fraction of these regions with a higher χ^2 statistic than in the retroCNV region.

Supporting Information

Figure S1 Nucleotide diversity on chromosome 11 among chromosomes containing and lacking the *GNG10* retroCNV in CEU. π is shown in 10 kilobase windows for chromosomes containing the *GNG10* retroCNV (red) and those lacking this retroCNV (black). The location of the retroCNV insertion is marked by an arrow. As with *DHFR*, there is a recombination hotspot distal to the retroCNV (data from ref. [65]).
(TIF)

Figure S2 Nucleotide diversity on chromosome 11 among chromosomes containing and lacking the *GNG10* retroCNV in YRI. π is shown in 10 kilobase windows for chromosomes containing the *GNG10* retroCNV (red) and those lacking this retroCNV (black).
(TIF)

Figure S3 Nucleotide diversity on chromosome 18 among chromosomes containing and lacking the *DHFR* retroCNV in ASI. π is shown in 10 kilobase windows for chromosomes containing the *DHFR* retroCNV (red) and those lacking this retroCNV (black).
(TIF)

Table S1 Genomes used to discover retroCNVs absent from the reference genome.
(XLS)

Table S2 Coordinates of retrotransposed genes and their insertion sites (hg19).
(XLS)

Table S3 Genomes used for experimental validation.
(XLS)

Table S4 RetroCNVs and genome sequences examined in each analysis.
(XLS)

Table S5 Movements of retroCNVs and fixed retrogenes originating on the X chromosome and originating on the autosomes.
(XLS)

Table S6 Movements of retroCNVs and fixed retrogenes to the X chromosome and to the autosomes.
(XLS)

Table S7 Movements of retroCNVs and fixed retrogenes originating on the X chromosome and originating on the autosomes, including retroCNVs with an unknown insertion site.
(XLS)

Table S8 Genotypes of two parent-offspring trios.
(XLS)

Acknowledgments

We thank Fernanda Koyama for assistance with experimental validation and Andrew Kern for helpful discussions about the positive selection analysis.

Author Contributions

Conceived and designed the experiments: DRS FCPN PAFG RBP AAC MWH SJdS. Performed the experiments: DRS FCPN RBP. Analyzed the data: DRS FCPN PAFG MWH SJdS. Contributed reagents/materials/analysis tools: AAC. Wrote the paper: DRS FCPN PAFG MWH SJdS.

References

1. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006) The evolution of mammalian gene families. PLoS ONE 1: e85. doi:10.1371/journal.pone.0000085
2. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704–712.
3. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, et al. (2012) Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. Cell 149: 912–922.
4. Iskow RC, Gokcumen O, Lee C (2012) Exploring the role of copy number variants in human adaptation. Trends Genet 28: 245–257.
5. Greenberg AJ, Moran JR, Fang S, Wu CI (2006) Adaptive loss of an old duplicated gene during incipient speciation. Mol Biol Evol 23: 401–410.
6. Long MY, Langley CH (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260: 91–95.
7. Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet 9: 938–950.
8. Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100: 605–617.
9. Girirajan S, Campbell CD, Eichler EE (2011) Human copy number variation and complex genetic disease. Annu Rev Genet 45: 203–226.
10. McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37–S42.
11. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med pp. 437–455.
12. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64.
13. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166–1174.
14. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
15. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–528.
16. Carreto L, Eiriz MF, Gomes AC, Pereira PM, Schuller D, et al. (2008) Comparative genomics of wild type yeast strains unveils important genome diversity. BMC Genomics 9: 524.
17. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science 320: 1629–1631.
18. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res 18: 2024–2033.
19. Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. Proceedings of the Royal Society B 277: 3213–3221.
20. Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007.
21. Schrider DR, Hahn MW (2010) Lower linkage disequilibrium at CNVs is due to both recurrent mutation and transposing duplications. Mol Biol Evol 27: 103–111.
22. Brosius J (1991) Retroposons - seeds of evolution. Science 251: 753–753.
23. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 3: e357. doi:10.1371/journal.pbio.0030357
24. Okamura K, Nakai K (2008) Retrotransposition as a source of new promoters. Mol Biol Evol 25: 1231–1238.
25. Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J (2008) Retrocopy contributions to the evolution of the human genome. BMC Genomics 9.
26. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 103: 3220–3225.
27. Bai YS, Casola C, Feschotte C, Betran E (2007) Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol 8: R11.
28. Betrán E, Thornton K, Long M (2002) Retroposed new genes out of the X in *Drosophila*. Genome Res 12: 1854–1859.
29. Emerson JJ, Kaessmann H, Betran E, Long MY (2004) Extensive gene traffic on the mammalian X chromosome. Science 303: 537–540.
30. Schrider DR, Stevens K, Cardeno CM, Langley CH, Hahn MW (2011) Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. Genome Res 21: 2087–2095.
31. Chiefari E, Iiritano S, Paonessa F, Le Pera I, Arcidiacono B, et al. (2010) Pseudogene-mediated posttranscriptional silencing of *HMGA1* can result in insulin resistance and type 2 diabetes. Nat Commun 1: 40.
32. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, et al. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465: 1033–1038.
33. Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, et al. (2011) Detection of structural variants and indels within exome data. Nat Methods 9: 176–178.
34. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
35. Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6: e80. doi:10.1371/journal.pbio.0060080
36. Diaz-Castillo C, Ranz JM (2012) Nuclear chromosome dynamics in the *Drosophila* male germ line contribute to the nonrandom genomic distribution of retrogenes. Mol Biol Evol 29: 2105–2108.
37. Galante PAF, Vidal DO, de Souza JE, Camargo AA, de Souza SJ (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. Genome Biol 8: R40.

38. Kojima KK, Okada N (2009) mRNA retrotransposition coupled with 5′ inversion as a possible source of new genes. Mol Biol Evol 26: 1405–1420.
39. Rogers RL, Hartl DL (2011) Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. Mol Biol Evol 29: 517–529.
40. Courseaux A, Nahon JL (2001) Birth of two chimeric genes in the *Hominidae* lineage. Science 291: 1293–1297.
41. Rogalla P, Kazmierczak B, Flohr AM, Hauke S, Bullerdiek J (2000) Back to the roots of a new exon - The molecular archaeology of a SP100 splice variant. Genomics 63: 117–122.
42. Jones CD, Custer AW, Begun DJ (2005) Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics 170: 207–219.
43. Wang W, Brunet FG, Nevo E, Long M (2002) Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 99: 4448–4453.
44. Jones CD, Begun DJ (2005) Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci U S A 102: 11373–11378.
45. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773-U151.
46. Galante PAF, Parmigiani RB, Zhao Q, Caballero OL, de Souza JE, et al. (2011) Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. Nucleic Acids Res 39: 6056–6068.
47. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
48. Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72. doi:10.1371/journal.pbio.0040072
49. Hudson RR, Kaplan NL (1986) On the divergence of alleles in nested subsamples from finite populations. Genetics 113: 1057–1076.
50. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
51. Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics 136: 1329–1340.
52. Anagnou NP, Antonarakis SE, Obrien SJ, Modi WS, Nienhuis AW (1988) Chromosomal localization and racial distribution of the polymorphic human dihydrofolate-reductase pseudogene (DHFRPI). Am J Hum Genet 42: 345–352.
53. McEntee G, Minguzzi S, O'Brien K, Ben Larbi N, Loscher C, et al. (2011) The former annotated human pseudogene dihydrofolate reductase-like 1 (DHFRL1) is expressed and functional. Proc Natl Acad Sci U S A 108: 15157–15162.
54. Cario H, Smith DEC, Blom H, Blau N, Bode H, et al. (2011) Dihydrofolate reductase deficiency due to a homozygous *DHFR* mutation causes megaloblastic anemia and cerebral folate deficiency leading to severe neurologic disease. Am J Hum Genet 88: 226–231.
55. Urlaub G, Chasin LA (1980) Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. Proc Natl Acad Sci U S A 77: 4216–4220.
56. Hawser S, Lociuro S, Islam K (2006) Dihydrofolate reductase inhibitors as antibacterial agents. Biochem Pharmacol 71: 941–948.
57. Huennekens FM (1994) The methotrexate story: A paradigm for development of cancer chemotherapeutic agents. In: Weber G, editor. Advances in Enzyme Regulation, Vol 34. pp. 397–419.
58. Cardenas-Navia LI, Cruz P, Lin JC, Rosenberg SA, Samuels Y, et al. (2010) Novel somatic mutations in heterotrimeric G proteins in melanoma. Cancer Biol Ther 10: 33–37.
59. Senderek J, Bergmann C, Weber S, Ketelsen UP, Schorle H, et al. (2003) Mutation of the *SBF2* gene, encoding a novel member of the myotubularin family, in Charcot-Marie-Tooth neuropathy type 4B2/11p15. Hum Mol Genet 12: 349–356.
60. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2012) Ensembl 2012. Nucleic Acids Res 40: D84–D90.
61. Kent WJ (2002) BLAT - The BLAST-like alignment tool. Genome Res 12: 656–664.
62. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34: D590–D598.
63. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: S KSaM, editor. Methods and Protocols: Methods in Molecular Biology. Totowa, NJ: Humana Press. pp. 365–386.
64. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
65. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Title: A genome-wide landscape of retrocopies in primate genomes

Fábio C. P. Navarro[1,2] and Pedro A. F. Galante[1,3]

[1] Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo, Brazil.

[2] Dep. de Bioquímica, Universidade de São Paulo, São Paulo, Brazil.

[3] Corresponding author

E-mail: pgalante@mochsl.org.br

Running title: Retrocopies in primate genomes

Keywords: Retrocopy, primate genomes, gene duplication, retrogene

ABSTRACT

The study of gene duplications contributes to the basic understanding of the evolutionary history, phenotypic characteristics and disease propensities of all living organisms. Despite the obvious importance of and the great availability of data necessary for the study of gene duplications, many species still remain to be further explored in terms of this issue. Here, we systematically analyzed mRNA retroposition, a class of gene duplication, in primate genomes. Analyzing seven anthropoid primates, we found a similar number of ~7,500 retroposition events (retrocopies) in Catarrhini (Old Word Monkeys [OWM], including human and other great apes), but a surprising large number of ~10,000 retrocopies in Platyrrhini (New World Monkeys [NWMs]), which seems to be a by-product of higher L1 sub-elements activity in these genomes. By analyzing retrocopy orthology, we dated most of primate retrocopies origin, estimated their fixation rate and catalogued retrocopies shared between murine rodents and primates, as well as species-specific retrocopies. Moreover, using RNAseq data, we reached a set of ~3,600 expressed retrocopies, some of which presenting tissue-specific or even species-specific expression. Taken together, our results provide further evidence for mRNA retroposition as an active mechanism in primates' evolution, and we highlight that retrocopies may not only introduce great genetic variability between lineages, but also create a large reservoir of potentially functional new genomic loci in the primate genomes.

INTRODUCTION

Gene duplication is one of the major contributors to the origin of adaptive evolutionary novelties (Ohno 1970; Long et al. 2003). Although complete genome duplications have had an important evolutionary role (Taylor and Raes 2004), it is the small-scale gene duplication that underlies the evolution of many novel phenotypic traits in many species (Conrad and Antonarakis 2007). Small-scale gene duplication events can be generated by chromosome segmental duplications, a DNA-mediated mechanism (reviewed in (Prince and Pickett 2002) and Marques-Bonet, T., Girirajan, S., & Eichler, E. E. (2009). The origins and impact of primate segmental duplications Trends in genetics : TIG, 25(10), 443–454. doi: 10.1016/j.tig.2009.08.002) or through reverse transcription of mature RNA intermediates, a mechanism called retroposition or retroduplication of mRNAs (Esnault, C., Maestre, J., & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes Nature genetics, 24(4), 363–367. doi:10.1038/74184). While the former mechanism has been extensively studied (Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18(6), 292–298. Elsevier. doi: 10.1016/S0169-5347(03)00033-8, Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., et al. (2005). Segmental duplications and copy-number variation in the human genome American journal of human genetics, 77(1), 78–88. doi:10.1086/431652), Conrad and Antonarakis 2007, , the impact and extent of retroduplication of mRNAs still deserves a deep and systematic investigation in many species (Kaessmann et al. 2009) .

In eutherian, mRNA retroduplication is carried out by two L1 (Long Interspersed Nuclear Element 1) proteins: one with reverse-transcriptase (Mathias et al. 1991) and

endonuclease (Feng et al. 1996) activities and a RNA-binding protein (Hohjoh and Singer 1997), which together hijack RNAs at the cytoplasm, synthesize (retro)copies and integrate the resultant transcripts into the nuclear genome (Esnault, C., Maestre, J., & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes Nature genetics, 24(4), 363–367. doi:10.1038/74184). Thereby, mRNA retrocopies usually contain only exonic sequences, lacking introns and the major regulatory regions from their parental genes (Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution Annual review of genetics, 19, 253–272. doi:10.1146/annurev.ge.19.120185.001345). However, despite the absence of regulatory regions, since late 80s (McCarrey and Thomas 1987) there is growing evidence that many retrocopies are in fact functional (usually called retrogenes), even those presenting non-coding transcripts (Poliseno et al. 2010; Tam et al. 2008; Fairbanks et al. 2012; Trembley et al. 2005; Hung et al. 2010; Baertsch et al. 2008)

Nowadays, for whole sequenced genomes, the detection of retrocopies relies on the finding of intronless duplications of multi-exonic genes (called parental genes). However, due to differences in the retrocopy screening strategy (Baertsch et al. 2008) there is no consensus for the number of retrocopies even in the human genome. Methods based on mRNA sequence alignments and accurate annotations have identified 7,000 to 13,000 retrocopies (Pei et al. 2012; Baertsch et al. 2008; Sakai et al. 2007). On the other hand, methods based on protein sequence alignments have reported 3,000 to 6,000 retrocopies (Vinckenbosch et al. 2006; Marques et al. 2005).

A remarkable feature of primate genomes is the proportion of retroposed insertions, adding up to ~45% for human (Venter et al. 2001; Lander et al. 2001), chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005), and gorilla (Scally et al. 2012).

Therefore, since mRNA retrocopies are a subclass of retroposed copies and a potential source of novel functional transcripts, it is reasonable to hypothesize that they might play key roles in the primate genome evolution. Nevertheless, although some studies have explored retrocopies in primates, many of their features remain to be elucidated (Kaessmann et al. 2009)

Here, we performed a systematic analysis of mRNA retrocopies in seven fully sequenced primates and two murine rodent genomes (our "outgroup"). Specifically, we catalogued their entire retrocopy repertoires, explored their retrocopies' origin, orthology and potentially expressed retrocopies. Overall, we believe that our results have brought new insights regarding retrocopies shaping and providing substrate for evolutionary innovations in primate genomes.

RESULTS

Retrocopies in primate genomes

In order to start our study, we first developed a set of pipelines to identify, select, and perform comparisons among retrocopies and their parental genes (for further information, see Material and Methods). Using our computational approach, we identified 57,212 loci originated from mRNA retrocopies in the seven studied primates (Table 1). A very similar number of events (~7.500 retrocopies, on average) were found in Catarrhini genomes (human, chimpanzee, gorilla, orangutan and rhesus), Table 1. Furthermore, Platyrrhini genomes (marmoset and squirrel monkey, NWMs) presented significantly more retrocopies (~10,000 events per species), approximately 50% more events, than other primates and murine rodents (Table 1; p-value < 2.2e-16, chi-square=449; d.f. = 1).

To further investigate the larger number of retrocopies in Platyrrhini genomes, we analyzed additional genomic features. In comparison to Catarrhini, no significant differences were found in terms of their genomic size, number of genes, number of transcripts, and percentage of genome composed of repetitive elements (Supplemental Table 1). However, we observed an intriguing difference: overall human, chimpanzee, gorilla, orangutan, and rhesus genomes have a similar composition of L1 sub-elements, but marmoset and squirrel monkeys presented an overrepresentation of L1PA7 and L1P3 (Figure 1A). These two L1 sub-elements correspond respectively to ~25% and ~5%, of the most frequent L1 elements in NWMs genomes, but they are significantly less frequent, ~5% (L1PA7; p-value< 2.2e-16, chi-square=50809; d.f. = 1) and ~1% (L1PA3; p-value=< 2.2e-16, chi-square=6913; d.f. = 1), in the Catarrhini genomes (Figure 1A). Analyzing the multiple alignment of L1PA7 ORF2p in the seven primate genomes, we also observed that, despite some similarities between Platyrrhini and Catarrhini L1PA7 content (suggesting an ancestral origin), a major number of L1PA7 copies were only found in Platyrrhini (Figure 1B), indicating a putative lineage specific expansion of this sub-element.

Ortholog retrocopies across primate and murine genomes.

In primate genomes, studies based on nucleotide substitutions found that most of mRNA retrocopies originated within primate lineage, 90-40 myr ago, in parallel to SINEs (Short Interspersed Nuclear Elements) expansion (Ohshima et al. 2003). To confirm and better explore this result, we took advantage of fully sequenced genomes from primates and murine rodents (our outgroup) to precisely identify their shared retrocopies. Due to the identical mechanism of insertion and the large size of primate/rodent genomes, it is

reasonable to expect that independent retroposition events will have distinct genomic insertion points. Consequently, a syntenic genomic locus, sharing the same retrocopied gene, must be the result of an ancestral retroposition event. By using this strategy (for details see Material and Methods and (Navarro and Galante 2013)) we identified 63 (less than 1%) retrocopies shared between murine rodents and primate, which probably originated before primate-rodent divergence, ~90-120 myr ago.

Next, by assuming that functional sequences are conserved for a long period of time (Charlesworth et al. 1995), we decided to further evaluate these 63 primates-rodent shared retrocopies. First, we found that a majority (51 out of 63 (80.0%)) of these retrocopies have an annotated RefSeq (Pruitt et al. 2014) transcript, Table S2. From these 51 retrocopies, 45 were classified as protein-coding genes (i. e., putative retrogenes) of which we found an enrichment of mRNA metabolic process, heat shock proteins and Zinc finger on INTERPRO terms. Additionally, four retrocopies were annotated as non-coding transcripts, and two were annotated as undergoing exonification (i. e., forming chimeric transcripts with other genes), see Table S2. Moreover, our RNA-seq analyses (see Material and Methods and next sections), confirmed that 50 (79%) of these retrocopies are expressed and, as expected for functional retrocopies (Kaessmann et al. 2009), most of them (96%) are expressed in testis, including 14 candidates with tissue-specific expression (Table S3).

Since purifying selection of genomic sequence represents powerful evidence for functionality (Lowe et al. 2007), we also evaluated the rate of non-synonymous/synonymous (Ka/Ks) distribution from these primate-rodent retrocopies. The 63 retrocopies presented a Ka/Ks distribution with a peak smaller than 0.5 (Figure S1; median 0.22), while 1000 random sets of 63 retrocopies presented a Ka/Ks centered between 0.5 and 1 (Figure S1, median

0.58). Such difference (p-value < 0.0001; Mann-Whitney U test) suggests that most of these retrocopies are subject to selective constraints and therefore potentially functional.

Additionally, we decided to investigate how many of the 63 primate-rodent shared retrocopies are related to the X chromosome (X), since some genes located in X (X-genes) 'export' retrocopies to autosomes (Emerson et al. 2004) to escape to the X-gene silencing during the haploid stages of spermatogenesis in males (Richler et al. 1992). In the human genome, we found 43% (27 out 63) of these retrocopies in accordance with this hypothesis, being both migrations out of (expected: 3 retrocopies; found: 13 retrocopies; p-value = 0.016) as well as to (expected: 2 retrocopies; found: 14 retrocopies, p-value = 0.0032) the X chromosome. In terms of comparison, only ~1% of all human retrocopies (excluding these 27 retrocopies) were inserted into or originated from genes located in the X chromosome.

Retrocopy orthology within primate genomes

Based on our results and data from others (Marques et al. 2005; Ohshima et al. 2003; Zhang et al. 2004), it is clear that most of primates' retrocopies have originated within their own lineage in the last 90 myr of our genome evolution. However, little is known about retrocopy orthology across primates and yet there is no consensus (Ohshima et al. 2003; Marques et al. 2005; Pei et al. 2012; Zhang 2013) whether they originated in a short period of time (during a mRNA retroposition burst in an ancestral organism, similar to segmental duplications (Marques-Bonet Nature 2009)) or diluted through the primate speciation period (Zhang 2013). In order to further investigate this question, we attempted to identify ortholog and species-specific retroposition across the primates.

Application of a similar aforementioned approach to identify murine-primate retrocopies (see Material and Methods) help us to we identified 4,168 retrocopies shared across primates (Figure 2A), i. e., these retrocopies' origin dates back to before the Platyrrhini-Catarrhini divergence, ~42 myr ago (Steiper and Young 2006). We also identified 5,662 retrocopies shared by Platyrrhini and 7,518 retrocopies shared by human and chimpanzees. (Figure 2A). We also evaluated chromosome 21, the only finished chromosome, regarding the percentage of shared retrocopies and found no significant deviation from other autosomes (chromosome 21 = 97.47% shared; autosomes median=96.37%; standard deviation of 1.33%). Next, in order to estimate the rate of retrocopies origin during the primate evolution, we performed a rough estimation of the number of retrocopies originated in each time period (Table 2). Overall, we found a continuous decrease of retrocopies' origin and fixation, beginning higher in the primate order (between 42 and 30 myr ago), with an average of ~142 (1707/12) retrocopies per million year (Table 2 and Figure 2), but hardly decreasing until great ape lineage (gorilla, chimpanzee and human), which presented ~68 retrocopies per million year. Curiously, the human lineage shows the smallest rate of retrocopies origin/ fixation (Table 2 and Figure 2). Otherwise, NWMs have a high rate of retrocopies origin and fixation, ~152 retrocopies per million years (Table 2 and Figure 2). We also investigated the overlap between the results from orthology and Ks analysis. We found that, especially for recent events Ks is not sensible enough to distinct human specific retrocopies, retrocopies specific to humans and chimpanzees and so on (Figure S2).

Next, we investigated the set of species-specific retrocopies. First, we identified candidate retrocopies specific to human, chimpanzee and gorilla: 127, 228, and 212 retrocopies, respectively (Figure 2B). A couple of the 127 human specific retrocopies are described as functional (such as NANOGP8 (Fairbanks et al. 2012) and CSNK2A3 (Wirkner

et al. 1992)) and others (11 events) that are still unfixed in the human population, as we described recently (Schrider et al. 2013). In contrast, larger sets of species-specific retrocopies were found in marmoset (3,980 events) and rhesus (1,623 events), Figure 2B. Even though it is likely that our set of species-specific retrocopies contains false-positive candidates (especially in rhesus and marmoset due to the lack of closely related species), the identification of this set of candidate genes may be an important starting point for further exploration to advance our understanding of species evolution.

Transcribed retrocopies in primates

It has been reporting an increasing number of protein coding and noncoding functional mRNA retrocopies (Poliseno et al. 2010; Tam et al. 2008; Fairbanks et al. 2012; Trembley et al. 2005; Hung et al. 2010; Baertsch et al. 2008). To be functional, a retrocopy needs to be transcribed (Kaessmann et al. 2009). Therefore, to escape transcriptional inability, retrocopies usually hijack regulatory elements from other transcribed regions adjacent to their insertion point (Vinckenbosch et al. 2006). Even though the ENCODE project has shed light on the stochasticity of the human genome transcriptional capacity, it also suggested that fractions of the expressed retrocopies are not transcriptional noise, but potentially functional (Pei et al. 2012). Therefore, in order to extend the set of expressed retrocopies, we used RNA-seq data (see Material and Methods) to identify expressed retrocopies in 6 healthy tissues (brain, cerebellum, testis, heart, liver and kidney) from five primates.

We identified a large set of expressed 3,562 candidate retrocopies in human (1,304), chimpanzee (1,500), gorilla (1,461), orangutan (846), and rhesus (1,324), Figure 3A. Interestingly, for most of primates, these retrocopies fitted the expected gene expression profile already described for human (Jongeneel et al. 2005): more diversified (higher number)

in testis and nervous tissues and less abundant in other highly specialized tissues, such as kidney, liver and heart, Figure 3B.

In order to understand how these retrocopies were expressed, we analyzed their closeness to regulatory regions. As expected (Vinckenbosch et al. 2006), a significant number of these retrocopies (71%; p-value < 2.2e-16; chi-square=308; d.f. = 2 – Permutation Test p-value < 0.0001 [Figure S3]) were located near or within known genes (Figure 3B). Since mobilization to another genomic location put the set of expressed retrocopies in a novel transcriptome regulatory context (Kalyana-Sundaram et al. 2012), we also evaluated the expression profiles of retrocopies and their parental genes. We found no correlation between retrocopies and their parental genes' expression (P=0.46; Spearman=-0.0241; Figure S4). However, we observed that these retrocopies presented a tissue-specific expression or were expressed in fewer tissues than their parental genes (Figure S5, p-value < 2.2e-16). We also found 310, 432, 486, 251 and 605 retrocopies presenting species-specific expression in human, chimpanzee, gorilla, orangutan and rhesus, respectively. Additional analyses will be required for an in-depth exploration to confirm that our set of transcribed retrocopies contains novel (functional) genes.

DISCUSSION

Several studies have pointed out mRNA retrocopies as a source of evolutionary novelty in several eukaryote species (Long et al. 2003; Ohno 1970; Kaessmann et al. 2009). Nevertheless, retrocopies still need to be deeply studied and therefore catalogued. Here, we performed a systematic analysis of retrocopies in seven primate genomes (human, chimpanzee, gorilla, orangutan, rhesus, marmoset and squirrel monkey, as well as two murine

rodents) and we showed, how abundant, active, and potentially expressed these mRNA retrocopies are.

To the best of our knowledge, we provide for the first time a most extensive catalogue of retrocopies for Old World and New World primates. In agreement with other authors (Pei et al. 2012; Baertsch et al. 2008; Balasubramanian et al. 2009), we found ~8000 retrocopies in the human genome. However, for chimp, orangutan and rhesus we found twice as many retrocopies than Zhang and colleagues described in a recent study (Zhang 2013). This difference emerges from what has already been noticed by Baertsch and colleagues (Baertsch et al. 2008): mRNA-based methodologies (such as we used) are more efficient to identify retrocopies containing mainly UTRs (untranslated regions) and/or short coding regions. On the other hand, retrocopy screening based on proteins (used by Zhang) usually reports ~2x less candidates. Moreover, due to the high similarity among primate genomes, a similar number of retrocopies between human and other primates is expected, such as we have found.

Platyrrhinis are the largest primate family. It contains ~150 species, most of them living in Central and South America (Groves 2001) and some becoming endangered. Furthermore, little is known about these monkeys: for example, we barely understand their origin in the New World, as well as details of their genome sequences (Jameson et al. 2012). Here, we not only described that marmoset and squirrel monkey (New World monkeys) have ~50% more mRNA retrocopies than Old Word monkeys, but we also suggested that this difference may be related to an extended L1 sub-element activity (L1PA7) into NWMs genomes. In line with our hypothesis, Ohshima et al. suggested that L1PA7 was one of the top three most probable L1 subfamilies involved in retrocopies' origin in ancestral primates 40-50 myr ago (Ohshima et al. 2003). In addition to these results, additional studies are needed for a

complete understanding of the reason for the high retrocopy content in Platyrrhini genomes, as well as the contribution of L1PA7 to retrocopy generation.

Taking the benefit of having access to a rich set of complete genome sequences for primates (and also for non-primates to be used as outgroups), we have identified retrocopies shared by primates and murine rodents (our outgroup) genomes. Thereby, we showed that more than 90% of primate and murine rodent retrocopies originated independently and after the split of their last common ancestrals. In agreement with our data, Marques et al. (Marques et al. 2005) and Zhang et al. (Zhang et al. 2004) have already suggested that most of human retrocopies were created after the last human-mouse split and Ohshima et al. suggested a burst of retrocopies (and Alus) formation in the genome of an ancestral primates, ~40-50 myr ago (Ohshima et al. 2003).

In addition, we also identified 63 retrocopies shared between primate and murine rodents. Most of these retrocopies yield indicators of functionality, such as a) they were already reported as transcribed genomic regions; b) they contain an annotated reference mRNA sequence; c) they seem to be under purifying selection; d) they are related to the X chromosome, by migrating out and to the X chromosome.

Recently, many studies have been reporting an increasing number of expressed and potentially functional retrocopies, most of them presenting not only protein coding (retrogenes), but also non-coding transcripts (Poliseno et al. 2010; Tam et al. 2008; Fairbanks et al. 2012; Trembley et al. 2005; Hung et al. 2010; Baertsch et al. 2008; Kalyana-Sundaram et al. 2012). As expected, a large fraction of these expressed retrocopies are thought to hijack regulatory regions or being inserted into transcribed region from coding genes (Vinckenbosch et al. 2006). In this manuscript we used RNAseq data and a well-refined gene expression pipeline to expand the set of transcribed retrocopies for primates through the identification of

~3,600 transcribed retrocopies in five primates, some of them showing a tissue specific and non-correlated expression to their parental genes. We also reported a set of intragenic retrocopies creating chimeric transcripts with their host genes, a mechanism to join protein domains, such as reported by Vinckenbosch et al (Vinckenbosch et al. 2006). We also identified sets of species- and/or tissue-specific retrocopies, which is an initial step in the track to functionalization (Bai et al. 2007; Vinckenbosch et al. 2006). Similarly to (Marques et al. 2005), we identified an enriched set of retrocopies expressed in brain and testis tissues, tissues essential to the evolutionary successful of all species.

Overall, we believe that our study has given at least three major contributions to the retrocopy field: first, we considerably expanded the catalog of mRNA retrocopies for primates, including the identification of large retrocopy sets in Platyrrhini genomes. We also suggested that part of retrocopy content in Platyrrhini would be related to an extra activity of L1 sub-elements; second, we have confirmed that most of primate and rodent retrocopies originated after their common ancestral. We outlined new details regarding retrocopy origin and conservation across primates and identified a small set of potentially functional retrocopies shared by primates and murine rodents; third, we described a large set of expressed retrocopies, which may contains many coding and non-coding functional retrocopies. In summary, all results presented here may help to unveil how retrocopies can contribute to shape, to create variability and novelties in the primate genomes.

METHODS

Data sources

The primate genome and transcriptome datasets were downloaded from the UCSC genome browser (http://genome.ucsc.edu) and the RefSeq database (Pruitt et al. 2014): version 49 (human [hg19], mouse [mm9] and rat [rn4]); version 50 (chimpanzee [panTro3]); version 51 (orangutan [ponAbe2, marmoset [calJac3], rhesus [rheMac2]); version 61 (squirrel monkey [SaiBol1.0]). Only gorilla transcripts were downloaded from ENSEMBL (http://www.ensembl.org; version 66). Genomic coordinates for: i) Transcription start site (TSS; GENCODE v12); ii) repetitive elements, polyadenylation (polyA) sites, and centromeric-telomeric regions were also obtained at UCSC Genome Browser and used in the retrocopy genomic context analysis. Finally, to investigate the expressed retrocopies, we used publicly available RNA-seq data [GEO: GSE30352] generated by (Brawand et al. 2011) for six tissues (brain, cerebellum, heart, liver, kidney and testis) of five primates (human, chimpanzee, gorilla, orangutan and rhesus).

Identifying retrocopies of protein coding genes

Since a main feature of retrocopies is that they are processed copies of multi-exonic genes, our pipeline relied on the identification of genomic intronless alignments from mature transcripts (mRNAs). First, all known coding gene transcripts mRNAs were aligned to their respective reference genome using BLAT (parameters: -mask=lower; -tileSize=12; -minIdentity=75; -minScore=100). Next, we selected alignments with identity greater than 75%, and either, more than 50% of the parental transcript or at least 120 nucleotides aligned. Alignments containing gaps larger than 15kb (putatively large introns) were excluded from further analysis. While this last

filter removed most of introns, it also allowed for a couple of repetitive elements (which are mainly <10 kb in length) insertions inside the putative retrocopy loci. Next, we selected the retrocopies by screening for parental exons in each putative retroduplication event and selecting only those candidates with, at least, two parental exons aligned (>50 nt each) adjacently. A random set of 200 human retrocopies (and their parental genes) was analyzed manually and a small fraction (<3%) of them was estimated as potentially false positive. For example, olfactory receptors (ORs) and other problematic transcripts were manually removed from the final dataset. More details about this pipeline, as well as additional information regarding primate retrocopies can be accessed in (Navarro and Galante 2013).

Characterization of the LINE1 family

To better understand the large number of retrocopies present in the marmoset and squirrel monkey genomes, we compared composition of LINE1 (L1) subfamilies, content and length of L1 elements from all the primate genomes using RepeatMasker data, version 3.3.0 (http://www.repeatmasker.org). Because of the high content of L1 only those subfamilies with more than 10,000 members in the seven primates were analyzed. In order to analyze L1PA7 and L1P3 expansion on NWM genomes, we initially selected L1PA7 elements with intact ORF2 regions in all primate genomes, and we conducted a multiple alignment of DNA sequence of their ORF2 using CLUSTALW2 (parameters: -type=dna -quicktree). Finally, we plotted the phylogenetic three coloring each leaf with a specie color, using iTOL (Letunic and Bork 2011).

Defining orthology of retrocopy events

To define retrocopy origin among primates, instead of using the number of non-synonymous mutations (Ohshima et al. 2003), which is an indirect evidence, we developed a strategy to select orthologous retroduplications events based on their syntenic genomic position. Taking the advantage of assess fully sequenced genomes and the ability to define flanking sequences of retroduplication events (Scally et al. 2012). We defined a flanking region as three thousand nucleotides adjacent to each retrocopy and composed by blocks with at least 150 nucleotides of non-repetitive sequences. To ensure that retrocopy segments were not included within the flanking regions, we started extracting flanking sequences 5,000 nucleotides up- and down-stream from each retrocopy event. Next, retrocopies and their flanking regions were aligned against all the other primate and murine rodents genomes using BLAT (parameters: -mask=lower; -tileSize=12; -minScore=50; -minIdentity=0). Events sharing the flanking regions and the containing the same parental retrocopies than the query genomes were classified as orthologous. This strategy was also previously applied to identify ortholog events in our retrocopy database, RCPedia (Navarro and Galante 2013).

Ka/Ks analysis

In order to perform the Ka/Ks analysis, first we extracted CDS (coding sequence) information from all retrocopies and their parental genes based on RefSeq

annotation. Next, we execute a multiple alignment between these retrocopy and parental gene sequences using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2). Finally, all sequence gaps were removed from the multiple alignments and we used in a BioPerl package (DNAstatistics; http://www.bioperl.org/) to calculate Ka and Ks of each multiple alignment. The DNAstatistics package implements Nei-Gojobori evolutionary pathway method and uses Jukes-Cantor method of nucleotide substitutions.

Identification of expressed retrocopies

Due to the high similarity between retrocopies and their parental genes, we developed two distinct strategies to reliably detect the set of expressed retrocopies: i) for those intragenic retrocopies, we searched for reads reporting chimeric transcripts merging host genes and their retrocopies; and ii) for all retrocopies (including those intragenic) we searched for reads with reliable alignments onto retrocopies. For either, we used the same RNAseq dataset from (Brawand et al. 2011).

To detect chimeric transcripts, reads from multiple tissues were aligned to their respective genomes using gsnap (Wu and Nacu 2010) (parameters: -t 30; -B 4; --nofails; -A sam; -m 2; -n 1). Next, we selected reads spanning exonic regions from either, host genes and their intragenic retrocopies. Finally, we selected only those alignments with at least five reads supporting the same chimeric event, alignments defining (putative) introns with canonical splice sites (GT-AG) and alignment quality higher than 40 (Phred scale). To detect all other expressed retrocopies we constructed a database containing the sequences and extra regions from mature transcripts of the parental genes. This database was created in order to eliminate false-positive

alignments from parental genes. Next, we aligned the reads against this database using bowtie2 (Langmead and Salzberg 2012) (--end-to-end; -p 63; -M 40; -D 20; -R 4;-N 0; -L 15; -i S,1,0.50; --ignore-quals) and only those reads aligned uniquely (and with alignment quality greater than 40) in the retrocopy regions were selected and used to the expression analysis.

Exploring the genomic context of expressed retrocopies

To understand the genomic context of the retrocopy datasets, we classified the events based their insertion point: i) intragenic or intergenic, based on the coordinates of RefSeq coding and non-coding transcripts; ii) polyA proximity (retrocopy insertion <15 kb of a polyA site); and iii) Transcription Start Site (TSS) proximity (retrocopy insertion <15 kb of a known TSS). Permutation test was performed creating 10,000 random groups of *locus* with equivalent length to the 1,304 expressed retrocopies in humans. Each *locus* was then classified as of as distant or intragenic/near. Finally we calculated the percentage of intragenic/near events for each random group and compared to the measured percentage.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank Anamaria A. Camargo, Maria D Vibranovski, Ludwig Christian Hinske, Luiz O. Penalva, Gustavo França, Andrei Rozansk, Robson F de Souza, and Luiz F. L. Reis

for valuable discussions and suggestions. This study was supported by FAPESP (Grant No. 2012/24731-1 to PAFG) and a fellowship from CAPES (to FCPN).

Figure legends

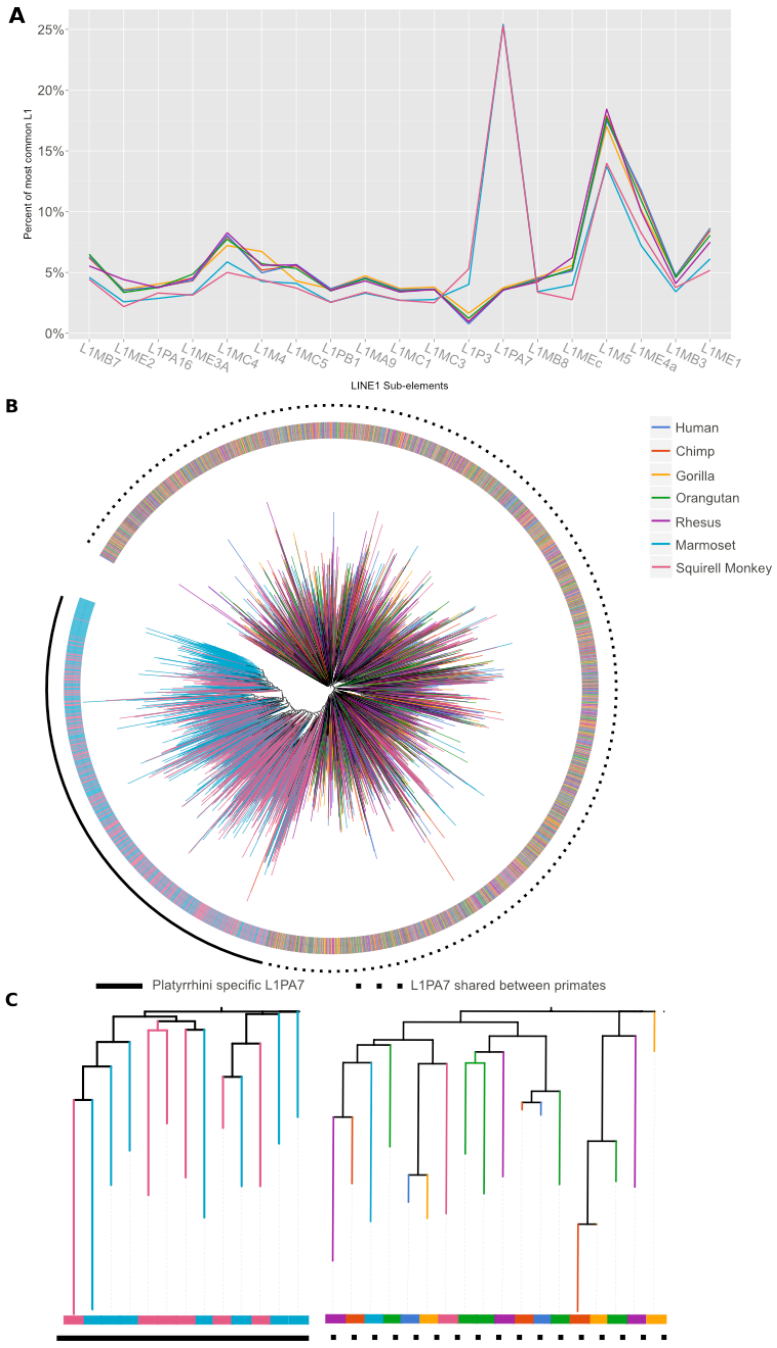


Figure 1. LINE1 sub-elements content in the primate genomes. a) Relative composition of the most frequent LINE1 sub-elements in the primate genomes. b) Phylogenetic tree generated by multiple alignment of intact L1PA7 ORF2 region. External ring and branch colors are defined by the species from which the sequences were extracted. c) small fragment of b) exemplifying a set of new world monkey specific L1PA7 (left) and L1PA7 shared between all primates (left)

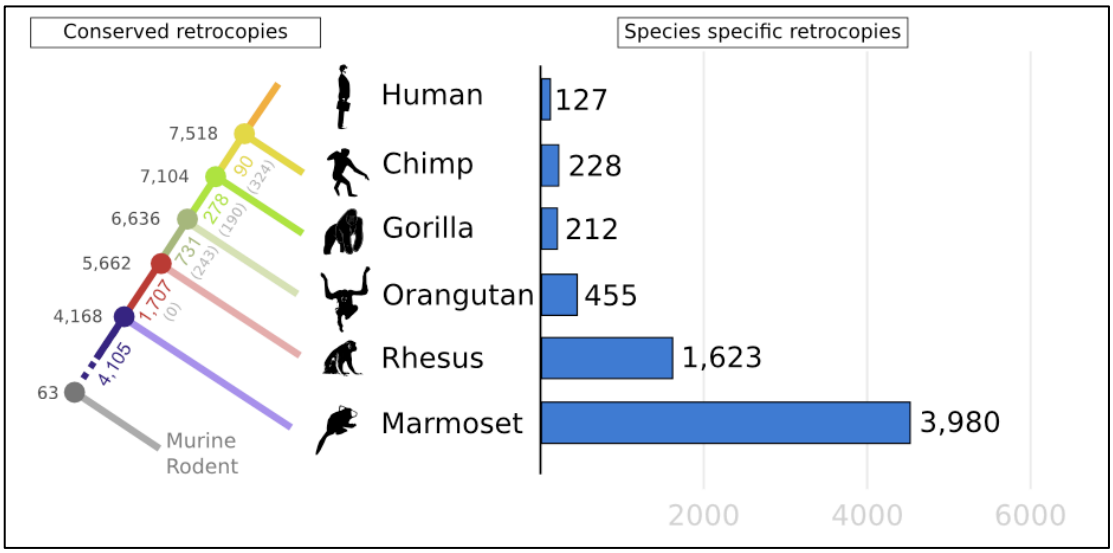


Figure 2. Conserved and species-specific retrocopies in primates. Dark gray numbers adjacent to colored circles presents conserved retrocopies. Colored numbers presents retrocopies originated per speciation period (eg.). Light gray numbers (in brackets) represent retrocopies conserved between some, but not among all respective primates.

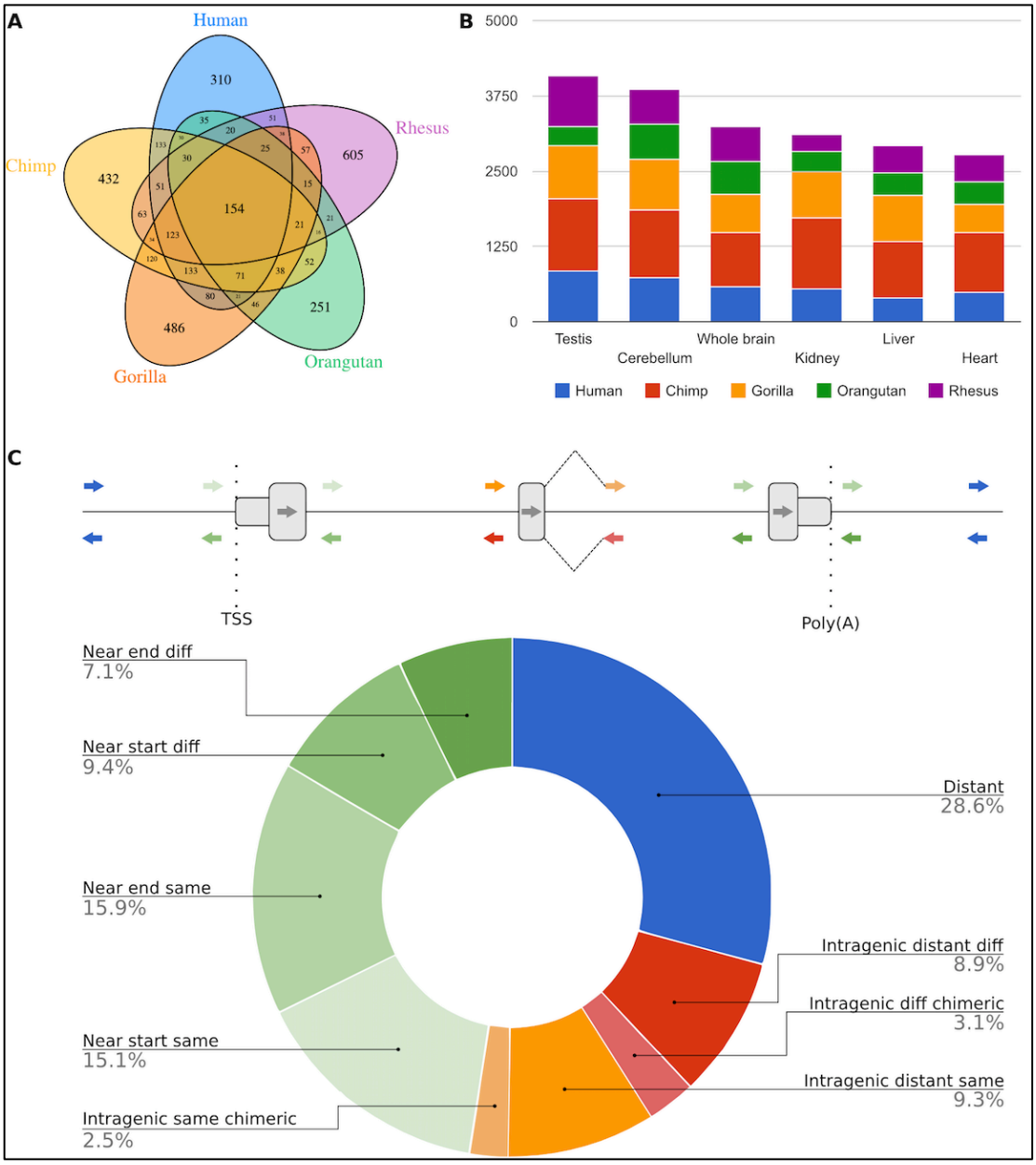


Figure 3. Expressed retrocopies and their genomic context. a) Venn diagram showing expressed retrocopies in human, chimpanzee, gorilla, orangutan, and rhesus. B) Bar plot showing retrocopy expression in the tissues. Retrocopies expressed in two or more tissues were quantified in all of them. C) The genomic context of expressed retrocopies. Retrocopies were classified as: chimeric transcript on the same or opposite strand of the host gene ("intragenic same chimeric" and "intragenic different chimeric", respectively); near TSS, on the opposite or same strand; near poly(A) site on the opposite or same strand; intragenic distance from TSS or poly(A) on opposite or same strand of the host gene; or distant from genes.

Table 1. Number of mRNA retrocopies and their parental genes identified per species.

Species	Number of retrocopies	Number of parental genes
Human	7,831	2,570
Chimpanzee	7,512	2,561
Gorilla	7,709	2,669
Orangutan	6,873	2,439
Rhesus	7,502	2,453
Marmoset	10,465	3,067
Squirrel monkey	9,320	2,864
Mouse	7,109	2,205
Rat	7,364	2,114

Table 2. Estimated rate of retrocopy origin/fixation during primate evolution.

Evolutionary Period	Branch number	Number of Retrocopies	Divergence Time	Retrocopies/Myr (average)
0 – 6 mya	1	127	6 myr	~21
6 – 8 mya	2	90	2 myr	~45
8 – 18 mya	3	278	10 myr	~28
18 – 30 mya	4	731	12 myr	~61
30 – 42 mya	5	1,707	12 myr	~142
0 – 42 mya	6	6,397	42 myr	~152
42 – 90 mya	7	4,105	48 myr	~85

Branches: #1: period after the last human/chimpanzee common ancestral; #2: period after the last gorilla/(chimpanzee, human) common ancestral and before the human/chimpanzee speciation. #3 period after the last orangutan/(gorilla, chimpanzee, human) common ancestral and before the gorilla/(human, chimpanzee) speciation; #4 period after the last rhesus/(orangutan, gorilla, chimpanzee, human) common ancestral and before the orangutan/(gorilla, chimpanzee, human) speciation; #5 in the OWMs lineage, the period after the last NWMs/OWMs common ancestral and before the rhesus/(orangutan, gorilla, chimpanzee, human) speciation; #6 in the marmoset lineage, the period after the last NWMs/OWMs common ancestral until now; #7 period after the last primate/rodent common ancestral and before the NWMs/OWMs speciation.

References

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

Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. *Genome Biol* **8**: R11.

Balasubramanian S, Zheng D, Liu Y-J, Fang G, Frankish A, Carriero N, Robilotto R, Cayting P, Gerstein M. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* **10**: R2.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2006. GenBank. *Nucleic Acids Research* **34**: D16–20.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* **141**: 1619–1632.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Conrad B, Antonarakis SE. 2007. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genom Human Genet* **8**: 17–35.

Emerson JJ, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303**: 537–540.

Fairbanks DJ, Fairbanks AD, Ogden TH, Parker GJ, Maughan PJ. 2012. NANOGP8: evolution of a human-specific retro-oncogene. *G3 (Bethesda)* **2**: 1447–1457.

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

Groves CP. 2001. *PRIMATE TAXONOMY*. Smithsonian Inst Press.

Hohjoh H, Singer MF. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J* **16**: 6034–6043.

Hung M-S, Lin Y-C, Mao J-H, Kim I-J, Xu Z, Yang C-T, Jablons DM, You L. 2010. Functional polymorphism of the CK2alpha intronless gene plays oncogenic roles in lung cancer. *PLoS ONE* **5**: e11418.

Jameson NM, Xu K, Yi SV, Wildman DE. 2012. Development and annotation of shotgun sequence libraries from New World monkeys. *Mol Ecol Resour* **12**: 950–955.

Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschild CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJG, et al. 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Research* **15**: 1007–1014.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**: 19–31. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19030023&retmode=ref&cmd=prlinks.

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu Y-M, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, et al. 2012. Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers. *Cell* **149**: 1622–1634.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

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

Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made

easy. *Nucleic Acids Research* **39**: W475–8.

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

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A* **104**: 8005–8010.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of Young Human Genes after a Burst of Retroposition in Primates. *Plos Biol* **3**: e357.

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

McCarrey JR, Thomas K. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* **326**: 501–505.

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

Ohno S. 1970. *Evolution by gene duplication*. Springer, New York.

Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N. 2003. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* **4**: R74.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biol* **13**: R51.

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

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3**: 827–837.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research* **42**: D756–63.

Richler C, Soreq H, Wahrman J. 1992. X inactivation in mammalian testis is correlated with inactive X-specific transcription. *Nat Genet* **2**: 192–195.

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

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.

Schrider DR, Navarro FCP, Galante PAF, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet* **9**: e1003242.

Steiper ME, Young NM. 2006. Primate molecular divergence dates. *Molecular Phylogenetics and Evolution* **41**: 384–394. http://books.google.com.br/books?id=_k60AAAAIAAJ&q=16815047&dq=16815047&hl=&cd=2&source=gbs_api.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**: 534–538.

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

Trembley JH, Tatsumi S, Sakashita E, Loyer P, Slaughter CA, Suzuki H, Endo H, Kidd VJ, Mayeda A. 2005.

Activation of pre-mRNA splicing by human RNPS1 is regulated by CK2 phosphorylation. *Mol Cell Biol* **25**: 1446–1457.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291**: 1304–1351.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* **103**: 3220–3225.

Wirkner U, Voss H, Lichter P, Weitz S, Ansorge W, Pyerin W. 1992. Human casein kinase II subunit alpha: sequence of a processed (pseudo)gene and its localization on chromosome 11. *Biochim Biophys Acta* **1131**: 220–222.

Zhang Q. 2013. The role of mRNA-based duplication in the evolution of the primate genome. *FEBS LETTERS* 1–8.

Zhang Z, Carriero N, Gerstein M. 2004. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet* **20**: 62–67.