Retrocópias - posição genômica e expressão em tecidos normais

Helena Beatriz da Conceição e Rafael Luiz Vieira Mercuri

Bioestatística (BIO5752) - Departamento de Genética e Biologia Evolutiva
Instituto de Biociências, Universidade de São Paulo

Resumo

Retrocópias são cópias de RNA mensageiro reverso transcritos e inseridos no genoma, resultado da atividade da maquinaria de LINE1. Como a endonuclease da maquinaria de LINE não apresenta tropismo por uma região específica (inter ou intragênica), retrocópias são inseridas de maneira aleatória no genoma. Graças às suas características (sendo a ausência de promotor de transcrição uma das principais), acreditou-se por algum tempo que essas retrocópias não poderiam ser expressas. Entretanto, sabe-se atualmente que elas são expressas e visto que elas podem ser encontradas em diversos pontos do genoma humano, o objetivo do trabalho é verificar se a expressão das retrocópias está ligada de alguma forma com a sua posição do genoma e em tecidos específicos, já que em regiões intragênicas as retrocópias poderiam se aproveitar dos promotores de seu gene hospedeiro e que esses genes hospedeiros podem ser expressos em tecidos específicos.

Introdução

O fenômeno de duplicação gênica inclui as vias mediadas por DNA - duplicação segmental cromossomal[1] - e RNA - transcrição reversa de intermediários de RNA maduros[2]. As cópias gênicas originadas por meio do mecanismo mediado por DNA mantêm as características dos genes a partir das quais foram geradas (genes parentais), incluindo éxons, íntrons e elementos reguladores de regiões promotoras para a transcrição[3]. Já as cópias gênicas criadas por meio do mecanismo mediado por RNA, chamadas de retrocópias, são caracterizadas pela conservação apenas dos éxons parentais, uma vez que o processo de duplicação ocorre a partir do RNA mensageiro maduro (sem íntrons e poliadenilado) - Figura 1.

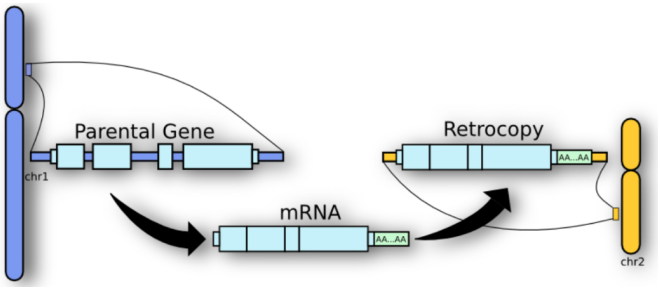

Figura 1. Origem e características das retrocópias. Representação esquemática de um evento de criação de uma retrocópia. Nesse exemplo, um gene parental (presente no cromossomo 1 (chr1)) é retroduplicado em uma retrocópia no cromossomo 2 (chr2).

A maquinaria enzimática necessária para a transcrição reversa do RNA mensageiro e a sua subsequente integração no genoma é fornecida majoritariamente por elementos transponíveis pertencentes às superfamílias de retrotransposons não-LTRs (Long Terminal

Repeats). Esses elementos codificam uma transcriptase reversa[4] e uma endonuclease/integrase[5], responsável por mediar a inserção de retrocópias no genoma. Em diversos mamíferos, a maquinaria de retrotransposição é codificada a partir de elementos LINE1 (long interspersed nuclear elements), que representam um conjunto de retrotransposons bem variado entre as espécies e é o mais abundante em humanos[6], por exemplo.

Até recentemente, todas as retrocópias eram classificadas como pseudogenes processados, devido à suposição de que a ausência de regiões regulatórias (para a transcrição) em mRNA seria um empecilho para a expressão das mesmas[7]. Contudo, análises de transcriptoma já revelaram que centenas de retrocópias em humanos e camundongos são capazes de gerar transcritos - alguns potencialmente funcionais[8]. A transcrição das retrocópias é possível a partir da obtenção de sequências regulatórias que, em geral, são provenientes de elementos regulatórios já existentes na vizinhança do ponto de inserção da retrocópia no genoma, como mostrado pelos trabalhos de Okamura e Nakai (2008)[9] e Fablet et al. (2009)[10].

As cópias originadas por meio da transcrição reversa de mRNA recebem a nomenclatura de retrocópias e, quando fixadas e funcionais, são identificadas como retrogenes[3]. Alguns exemplos de retrogenes incluem GLUD2, que apresenta função específica em tecidos neurais[3]; e TRIM5-CYPA que é uma quimera com papel na defesa antiviral[7][12]. Além disso, algumas retrocópias codificantes já foram associadas a diferentes tipos tumorais[6], tais como PTENP1[13] e BRAFP1[14]. Um outro exemplo de retrogene relacionado a câncer é o gene RHOB, um supressor da família Rho GTPase, que surgiu de um evento de retrotransposição nos primeiros estágios da evolução dos vertebrados[15]. Acredita-se que todos os vertebrados e sobretudo os mamíferos apresentam retrocópias, sendo que ~8.000 eventos fixados e algumas centenas de eventos polimórficos já foram identificados em humanos[11][16], por exemplo.

Objetivo

O objetivo geral deste projeto é investigar se a localização genômica de uma retrocópia afeta o seu nível de expressão. Para tanto, vamos: i) quantificar a expressão das retrocópias em múltiplos tecidos de indivíduos saudáveis; ii) classificar as retrocópias de acordo com a localização genômica (intergênica vs. intragênica); iii) aplicar testes estatísticos para verificar se existe diferença significativa de expressão de acordo com a localização genômica e com o tipo de tecido. Vamos aplicar dois testes de hipótese:

Hipótese 1:

H0: Diferente localizações de retrocópias (inter e intragênica) no genoma não interferem na sua expressão

H1: Diferente localizações de retrocópias no genoma interferem na sua expressão

Hipótese 2:

H0: Não existe diferença de expressão de retrocópias entre os tecidos

H1: Existe diferença de expressão de retrocópias entre os tecidos

Métodos

Iremos utilizar as retrocópias fixadas anotadas na RCPedia[16] - banco de dados de retrocópias desenvolvido pelo grupo de pesquisa do Dr. Galante e os dados de expressão gênica (RNA-Seq) de 53 tecidos do consórcio GTEx (https://www.gtexportal.org/)[17].

Partindo do conjunto de retrocópias fixadas (7.841) reportadas pela RCPedia[16] iremos verificar a sobreposição dessas retrocópias com as anotações do genoma humano no Gencode versão 26[18], utilizando a ferramenta bedtools intersect[19] para avaliar a sobreposição entre as anotações. Iremos verificar se há uma sobreposição de pelo menos 60 bases (metade do tamanho da menor retrocópia humana) e 90% de cobertura.

Para avaliar a expressão das retrocópias selecionadas em tecidos saudáveis, utilizaremos os dados do GTEx versão 8[17]. Iremos utilizar tecidos com pelo menos 70 amostras e excluindo amostras de cultura de células de fibroblasto. Para representar a expressão de um dado gene por tecido, iremos utilizar a mediana da expressão do gene nas amostras do tecido e iremos considerar a retrocópia expressa se a mediana for não nula.

Para determinar a localização genômica das retrocópias, iremos verificar a localização das retrocópias com relação aos genes codificadores anotados (Gencode v26) utilizando a ferramenta *bedtools intersect*.

Análise de Resultados

Para analisar os dados primeiramente vamos estabelecer a distribuição deste, se é normal ou não. Utilizaremos o teste de Shapiro-Wilk[20], para decidir quais testes aplicar (paramétricos ou não-paramétricos). Esse teste tem por hipótese nula que a distribuição dos dados é normal, e a sua hipótese alternativa de que não existe uma distribuição normal, se considerarmos o valor de $\alpha = 0.05$, iremos rejeitar a hipótese nula a partir de valores de $p \leq 0.05$. Entretanto, como estamos lidando com 46 tecidos, ou seja 46 amostras, temos que levar em conta o aumento do erro do tipo I (rejeitar a hipótese nula quando ela é verdadeira). Podemos usar correções para ajustar o nível de significância do experimento. Vamos utilizar a correção de Bonferroni [21], que é mais conservadora, utilizando então $\bar{\alpha} = \alpha/m = 0.05/46 \simeq 0.001087$.

Se tivermos como resultado que nosso dado tem uma distribuição não normal, aplicaremos a transformação em log2(TPM+1) para tentar normalizar os dados. Esse tipo de transformação é usual em dados de expressão gênica.

Ao fazer essa transformação realizaremos novo teste de normalidade e se o resultado do teste persistir em uma distribuição não normal, utilizaremos para a análise o teste de *Wilcoxon–Mann–Whitney*[22] para testar a diferença de expressão entre retrocópias inter e intragênicas e o teste de *Kruskal Wallis*[23] para a diferença entre os tecidos. O teste *Wilcoxon–Mann–Whitney* é um teste não paramétrico equivalente ao teste t de amostras independentes, que compara uma variável quantitativa em relação a dois grupos independentes pela organização de um ranking da variável nos dois grupos (Ex. **Tabela 1**) onde a hipótese nula desse teste é que a distribuição desse ranking nos grupos é igual.

Tabela 1: Exemplo de ranqueamento por expressão de retrocópias separadas por grupos de intra e intergênicas

Expressão	0.12	0.5	1.7	2.5	3.6	14.7	16	18.9	20.1	25.6
Grupo	Inter	Inter	Inter	Intra	Inter	Intra	Inter	Intra	Intra	Intra
Rank	1	2	3	4	5	6	7	8	9	10

O teste de *Kruskall Wallis* é um teste não paramétrico equivalente ao *One-Way ANOVA,* onde a expressão das retrocópias (variável quantitativa) nos diferentes tecidos (mais de dois grupos independentes) é ranqueada igualmente ao teste de *Wilcoxon–Mann–Whitney* [**Tabela 1**] e seu teste de hipótese segue a mesma linha: A distribuição desse ranking entre os diferentes grupos é igual. Entretanto, o teste não identifica qual ou quais amostras são estocasticamente dominantes. Test *post-hoc* podem ser aplicados, como o teste de Dunn e o teste de Conover-Iman.

Caso o resultado do primeiro ou segundo teste de normalidade aponte para uma distribuição normal, aplicaremos o teste t para amostras independentes e o *One-Way ANOVA*. Para comparar a expressão de retrocópias intra e intergênicas utilizaremos o teste t de amostras independentes, esse teste compara dois grupos independentes com a hipótese nula de que suas médias são iguais.

Para entender a relação entre a expressão por tecido das retrocópias usaremos o teste *One-Way ANOVA*, esse é um teste para dados com uma variável quantitativa (expressão do gene) e mais de dois grupos independentes que seguem um padrão de distribuição normal, onde a hipótese nula seria de que as médias dos grupos são iguais.

Resultados

A aplicação do teste *Shapiro-Wilk* nas distribuições de expressão de retrocópias intra e intergênicas em cada tecido do GTEx (em TPM) nos mostrou que essas distribuições não são normais. A aplicação de um segundo tipo de teste, *Anderson-Darling*, nos deu o mesmo resultado. Também aplicamos a transformação em log2(TPM+1), e as distribuições permanecem sem normalidade de acordo com os dois testes. A correção de *Bonferroni* também não altera o resultado.

Sabendo que nossos dados não seguem uma distribuição normal, seguimos para a aplicação do teste de *Kruskal-Wallis* para as distribuições intergênicas e intragênicas separadamente e com transformação em log2(TPM+1). O teste é significativo para as duas matrizes e para a transformação, indicando que existe diferença significativa de expressão de retrocópias entre os tecidos.

Por fim, aplicamos o teste de *Wilcoxon–Mann–Whitney* comparando a distribuição de expressão de retrocópias intragênicas e intergênicas para cada tecido. Foi feito o ajuste de *Bonferroni*, e encontramos 22 tecidos dos 46 que apresentam diferença significante (p-valor corrigido < 0.05). Os p-valores estão na **Tabela 2** e o gráfico da distribuição de expressão na **Figura 2**. Os dados e código utilizado para a análise deste trabalho estão em: https://github.com/helenabea/bioestat_helena_rafael

Tabela 2: Tecidos com valores de p≤0.05 no teste de *Wilcoxon–Mann–Whitney*

Tecidos	P valor ajustado	Tecidos	P valor ajustado
Minor Salivary Gland	1,81E-07	Vagina	9,14E-04
Skin Not Sun Exposed	8,58E-07	Colon Transverse	1,93E-03
Skin Sun Exposed	2,14E-06	Ovary	2,11E-03
Nerve Tibial	1,04E-05	Prostate	3,22E-03
Lung	5,42E-05	Adrenal Gland	5,73E-03
Adipose Subcutaneous	1,37E-04	Heart Left Ventricle	6,76E-03
Breast Mammary Tissue	1,54E-04	Liver	8,90E-03
Small Intestine Terminal Ileum	2,18E-04	Uterus	1,33E-02
Stomach	3,89E-04	Thyroid	1,50E-02
Esophagus Mucosa	5,16E-04	Heart Atrial Appendage	3,81E-02
Adipose Visceral	5,69E-04	Pancreas	4,09E-02

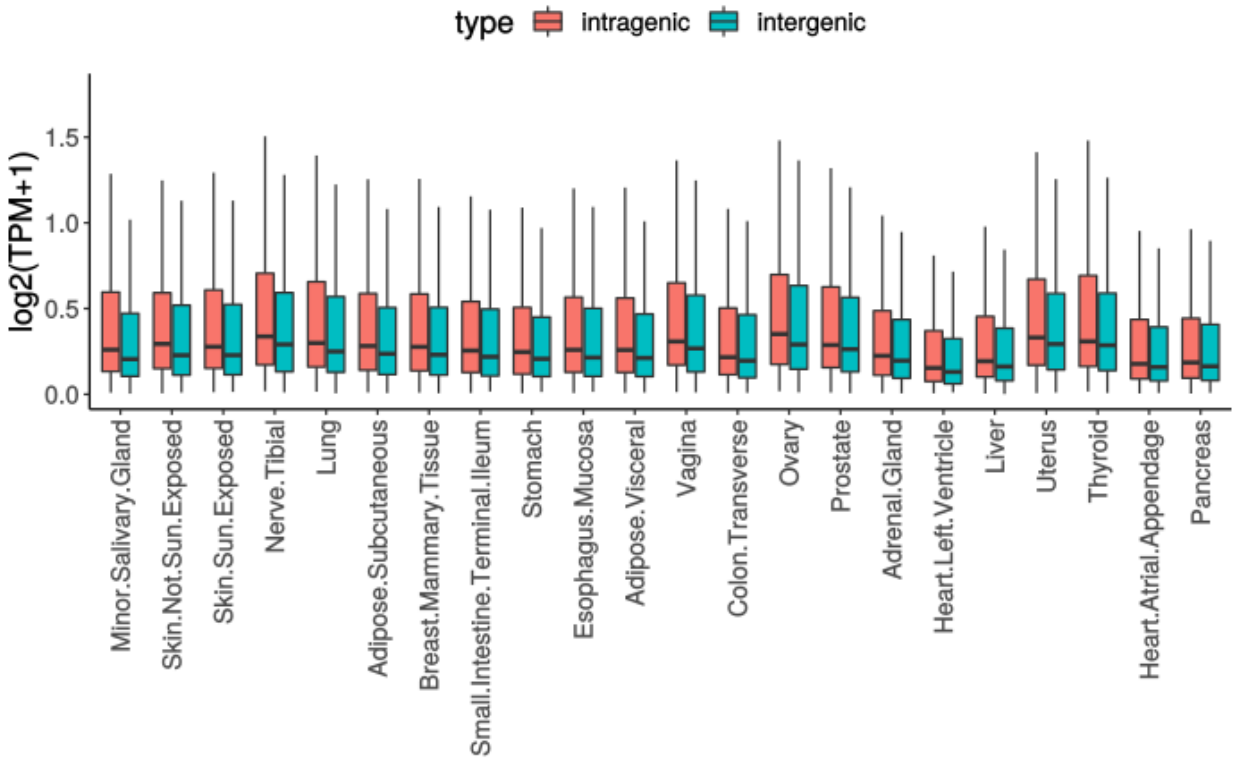

Figura 2: Expressão em log2(TPM+1) das retrocópias intragênicas e intergênicas nos Tecidos que apresentam diferença estatística significante entre os grupos. Ordenado pelo p-valor em ordem crescente.

Referências

[1] Prince, V. E., & Pickett, F. B. (2002, November 1). Splitting pairs: The diverging fates of duplicated genes. Nature Reviews Genetics. https://doi.org/10.1038/nrg928

[2] Esnault, C., Maestre, J., & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nature Genetics, 24(4), 363–367. https://doi.org/10.1038/74184

[3] Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanistic and evolutionary insights. Nature reviews Genetics. 2009;10(1):19-31. doi:10.1038/nrg2487.

[4] Mathias, S., Scott, A., Kazazian, H., Boeke, J., & Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science, 254(5039), 1808–1810. https://doi.org/10.1126/science.1722352

[5] Feng, Q., Moran, J. V., Kazazian, H. H., & Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell, 87(5), 905–916. https://doi.org/10.1016/S0092-8674(00)81997-2

[6] Kubiak, M. R. & Makałowska, I. Protein-Coding Genes' Retrocopies and Their Functions. Viruses 9, (2017).

[7] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Research. 2010;20(10):1313-1326. doi:10.1101/gr.101386.109.

[8] Casola, C. & Betrán, E. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? Genome Biology. 2017;9(6):1351-1373. doi:10.1093/gbe/evx081.

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

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

[11] Navarro FCP, Galante PAF. A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biology and Evolution. 2015;7(8):2265-2275. doi:10.1093/gbe/evv142.

[12] Kabza M, Kubiak MR, Danek A, et al. Inter-population Differences in Retrogene Loss and Expression in Humans. Gojobori T, ed. PLoS Genetics. 2015;11(10):e1005579. doi:10.1371/journal.pgen.1005579.

[13] R-K Li, J- Gao, L-H Guo, G-Q Huang & W-H Luo (2017). PTENP1 acts as a ceRNA to regulate PTEN by sponging miR-19b and explores the biological role of PTENP1 in breast cancer. Cancer Gene Therapy volume 24, pages 309–315 doi:10.1038/cgt.2017.29

[14] Florian A. Karreth (2015). The BRAF Pseudogene Functions as a Competitive Endogenous RNA and Induces Lymphoma In Vivo. Cell Volume 161, Issue 2, Pages 319-332. https://doi.org/10.1016/j.cell.2015.02.043

[15] Prendergast GC. Actin' up: RhoB in cancer and apoptosis. Nature Reviews Cancer. 2001; 1(2):162–8. PMID:11905808

[16] Navarro FCP, Galante PAF. RCPedia: a database of retrocopied genes. Bioinformatics. 2013;29(9):1235-1237. doi:10.1093/bioinformatics/btt104.

[17] The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on 09/18/2019.

[18] Frankish, Adam, et al. "GENCODE Reference Annotation for the Human and Mouse Genomes". Nucleic Acids Research, vol. 47, no D1, janeiro de 2019, p. D766–73. PubMed, doi:10.1093/nar/gky955.

[19] Quinlan, Aaron R., e Ira M. Hall. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". Bioinformatics (Oxford, England), vol. 26, no 6, março de 2010, p. 841–42. PubMed, doi:10.1093/bioinformatics/btq033.

[20] S. S. SHAPIRO, M. B. WILK, An analysis of variance test for normality (complete samples), Biometrika, Volume 52, Issue 3-4, December 1965, Pages 591–611, https://doi.org/10.1093/biomet/52.3-4.591

[21] Bonferroni, C. E., Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936

[22] Mann, Henry B.; Whitney, Donald R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". Annals of Mathematical Statistics. 18 (1): 50–60. doi:10.1214/aoms/1177730491

[23] Kruskal, William H.; Wallis, W. Allen (1 de dezembro de 1952). «Use of Ranks in One-Criterion Variance Analysis». Journal of the American Statistical Association. 47 (260): 583–621. doi:10.1080/01621459.1952.10483441