Universidade de São Paulo (USP)

Instituto de Física de São Carlos (IFSC)

JOÃO PAULO CASSUCCI DOS SANTOS

Análise da Retenção de Íntrons Mínimos em Câncer de Próstata

São Carlos

2021

João Paulo Cassucci dos Santos

Análise da Retenção de Íntrons Mínimos em Câncer de Próstata

Monografia apresentada ao Curso de Ciências Físicas e Biomoleculares, como requisito parcial para a obtenção do Título de Bacharel em Física Biomolecular, Universidade de São Paulo (USP).

Orientador: Prof. Ricardo De Marco (*in memorian*)
Prof. Andre Luis Berteli Ambrosio
M.Sc. Luíza Zuvanov

São Carlos

2021

Santos, Joao Paulo Cassucci dos
Analise da retençao de íntrons minimos em cancer de prostata. / Joao Paulo Cassucci dos Santos; orientador Ricardo De Marco (in memorian); co-orientador Andre Luis Berteli Ambrosio -- São Carlos, 2021.
32 p.

Trabalho de Conclusão de Curso (Bacharel em Ciências Físicas e Biomoleculares) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2021.

1. Retençao intrônica. 2. Introns minimos. 3. Cancer de prostata. 4. RNA Seq.. I. De Marco (in memorian), Ricardo , orient. II. Ambrosio, Andre Luis Berteli, co-orient. III. Título.

RESUMO

Neste projeto, buscou-se analisar o fenômeno da retenção intrônica em tecidos cancerosos da próstata, com ênfase nos íntrons caracterizados como mínimos (<= 112 pares de bases). Já é amplamente conhecido que células cancerosas têm uma maior taxa de íntrons retidos de forma geral quando comparadas com células saudáveis originadas do mesmo tecido. Experimentos de RNA-Seq de amostras extraídas de pacientes (tumor e normal adjascente) foram alinhados ao genoma humano, seguidos de um filtragem de baixas coberturas de leitura e depois obtido um parâmetro indicativo de retenção intrônica conhecido por *IRratio*. As distribuições dos limiares para o tamanho de íntrons mínimos e para o seu conteúdo GC (guanina-citosina, em %) foram obtidas através de métodos de regressão gaussiana e *Kernel Density Estimation* respectivamente, permitindo a classificação dos resultados em três populações distintas: HMiG (genes com introns mínimos e alto %GC), LMiG (genes com introns mínimos e baixo %GC) e woMiG (genes sem introns mínimos). Em seguida, por meio de análises de distribuição através de *boxplots* e testes de Qui Quadrado, verificou-se que as populações possuem níveis de retenção intrônica distintos, que a retenção ocorre com maior frequência nas amostras cancerosas e que o nível de retenção para a população HMiG ficou abaixo do esperado. Como última análise, os termos de *Gene Ontology* dos genes que tiveram uma retenção diferencial maior nos tecidos cancerosos foram colocados em tabelas e tiveram suas implicações biológicas analisadas. Chegou-se a conclusão que retenção intrônica em câncer está associada com vias metabólicas responsáveis por *splicing* e síntese de mRNAs, além de questões estruturais envolvendo lamininas e questões de sinalização e motilidade por meio de interferências em semaforinas. Uma correlação positiva foi observada entre o aumento de RI e genes que codificam para processos de aumento na transcrição, tradução e enovelamento proteicos. Este trabalho abre novas perspectivas de estudos sobre os efeitos e as adaptações celulares que resultam do aumento de RI em câncer.

Palavras-chave: Retenção intrônica. Íntrons mínimos. Câncer de próstata. RNA-Seq.

SUMÁRIO

1 INTRODUÇÃO

O fenômeno da retenção de íntrons (RI) nas moléculas de RNA mensageiro (mRNA) é um tipo de *splicing* alternativo que tem como função biológica, majoritariamente, a regulação da expressão gênica através de degradação mediada por mutação sem sentido (NMD). (1, 2) *Splicings* alternativos desta natureza ocorrem em seres humanos (2), mas estudos buscando entender os efeitos deste fenômeno são relativamente recentes para vertebrados em geral. (3, 4)

O fenômeno de RI está também relacionado com diversas doenças. Em câncer, por exemplo, a RI é a forma de *splicing* alternativo mais comum. Ademais, é visto que em células cancerosas, a RI altera transcritos de genes responsáveis pelo processamento de RNA e exportação nuclear de moléculas. (5–7) A RI foi também relacionada a grande variedade genética encontrada em diferentes cânceres (8), cuja característica torna o tratamento desta doença tão complexo.

Particularmente, a retenção de pequenos íntrons (chamados de íntrons mínimos) é de grande interesse nesse projeto. Estes íntrons aprimoram o processo de exportação de moléculas do núcleo para o citoplasma. Além disso, há evidências de que não se distribuem igualmente entre os genes humanos, estando presente principalmente em genes de *housekeeping*, fosforilação e operações de tráfego de moléculas. (9, 10) Outra característica conhecida dos íntrons mínimos é em relação ao conteúdo GC (conteúdo de guanina-citosina). Íntrons mínimos possuem uma distribuição bimodal, cujos valores podem ser usados como marcadores capazes de distinguir diferentes populações de genes em humanos. Desta forma, genes humanos podem ser classificados entre aqueles que possuem íntrons mínimos de baixo conteúdo GC, alto conteúdo GC, e também há aqueles com nenhum íntron mínimo. (8) É visto também que íntrons mínimos de baixo GC possuem pouca RI em condições normais, já íntrons mínimos de alto GC possuem valores de RI elevados. Curiosamente, a população de genes com íntrons mínimos de baixo GC está relacionada a processos de divisão celular e enriquecida com oncogenes. (8)

O foco deste estudo está direcionado para retenção de íntrons mínimos nas células cancerosas da próstata. Sabe-se que a severidade e a resistência de um tumor da próstata podem estar associadas à frequência de RI, com tumores mais agressivos e resistentes retendo mais. (11) Em 2020, o câncer de próstata foi o tipo de maior incidência em homens

no Brasil (29,2 %, https://www.inca.gov.br/en/node/2244) e o segundo em mortalidade (13,1 %). O câncer de próstata foi o segundo tipo mais comum de câncer não-cutâneo em homens em 2018. Além disso, é a quinta maior causa de mortes por câncer em homens globalmente. (12)

Neste projeto, utilizou-se bancos de dados de RNA-Seq de diferentes tipos de amostras teciduais da próstata, tanto cancerosas quanto saudáveis de tecido adjacente, de modo a averiguar se de fato ocorre a retenção de íntrons mínimos em maior quantidade nos tecidos cancerosos. Para determinar de modo mais preciso o tamanho de corte em número de pares de base dos íntrons mínimos em humanos, utilizou-se um método de regressão por funções gaussianas. Para o conteúdo GC, uma análise de distribuição com os dados dos íntrons mínimos obtidos tornou possível a divisão destes em 2 grupos: HMi (High Minimal Introns) e LMi (Low Minimal Introns) conforme descrito por (8). A retenção de cada um destes grupos foi avaliada por meio de boxplots e testes de qui-quadrado. Por último, os genes que continham íntrons com maior retenção foram analisados através de análise de enriquecimento de termos de ontologia gênica (GO). (13)

Este projeto teve como objetivo principal estudar a relação entre retenção de íntrons e os seus efeitos sobre o câncer de próstata, em particular os íntrons mínimos foram escolhidos devido a suas importantes implicações biológicas e também como um possível discriminador de funções gênicas.

2 MATERIAIS E MÉTODOS

2.1 Banco de dados e *IRratio*

Para realização das análises de retenção de íntrons mínimos em células cancerosas de pacientes, a partir de biópsias (tumores primários), selecionamos dados de sequenciamento de mRNA (RNA-seq) do projeto PRJNA128733 disponibilizado no banco de dados ENA (14). Ao todo, 30 experimentos de RNA-Seq distintos estavam disponíveis. Destes, selecionamos inicialmente aqueles que possuíam (i) experimentos tanto de células cancerosas quanto de células saudáveis adjacentes oriundas de em um mesmo paciente e (ii) profundidade de sequenciamento significativa (>6 milhões de leituras para os pares). Assim, quatro pares (saudável e tumoral) foram selecionados: pacientes 13, 15, 19 e 23.

Cada experimento de RNA-Seq foi baixado do banco de dados no formato FASTQ para ser processado pelo software IRFinder. (15) IRFinder utiliza as anotações conhecidas sobre genes codificantes de proteínas e realiza o alinhamento das sequências curtas de RNA-Seq (reads) sobre as fitas de mRNA de modo a verificar a quantidade de sequências que se alinham sobre regiões reconhecidas como íntrons. O resultado de interesse desta análise é o valor *IRratio* que descreve a quantidade de leituras do mRNA que possuem o íntron em sua composição sobre o total de leituras do mRNA *(IRratio = IntronDepth / (max(splices right , splices left) + IntronDepth).*(15)

A anotação do genoma humano utilizada como base para o alinhamento do RNA-Seq foi o arcabouço GRCh38 do *Human Genome Resources* no NCBI . Uma base de referência foi construída utilizando o software STAR para alinhamentos (16) e o conjunto de *scripts* bedtools2 para processamento de arquivos tipo *.bed.*(17)

Antes de passarem pelo processo de alinhamento com o genoma humano, os experimentos de RNA-Seq tiveram sequências de baixa qualidade removidas através do software Trimmomatic (18), que utilizou o parâmetro padrão Phred33 para selecionar as melhores sequências. Adicionalmente, o programa realizou uma remoção dos *primers* advindos do método Illumina Genome Analyzer II, de modo a evitar enviesamento nos alinhamentos.

Os experimentos de RNA-Seq em questão foram subsequentemente analisados utilizando o software FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) que demonstrou que mais de 95 % dos reads possuíam uma qualidade acima de 30 (no

parâmetro Phred), mostrando que o corte de sequências de baixa qualidade foi efetivo. Além disso, as sequências adaptadoras ILLUMINA responsáveis pela iniciação da síntese dos *reads* foram removidas com sucesso. Os experimentos selecionados tinham ainda como característica uma síntese de leitura não-direcionada. Portanto, os *outputs* do software IRFinder foram todos do tipo "non-dir". Destes *outputs*, foram removidos os íntrons que possuíam na coluna *Warning* o aviso *LowCover*, pois podem conter um baixo número de alinhamentos, podendo gerar ruído ao resultado final. (15)

2.2 Caracterização e classificação dos íntrons

Antes de avançar especificamente na retenção de íntrons mínimos, foram feitas análises levando em conta íntrons de todos os tamanhos para verificar se ocorre maior retenção em células cancerosas, um fenômeno já observado nestes tipos celulares. (5–7)

A distinção entre íntrons mínimos e íntrons longos é algo bastante evidente na maioria dos eucariotos, porém não há um consenso sobre qual tamanho, em pares de base, determina o limiar entre íntron mínimo e longo. Neste projeto, o tamanho foi utilizado como parâmetro para distinguir íntrons mínimos de longos com a utilização de análises matemáticas já desenvolvidas para classificação de íntrons com base no tamanho. (19)

Em relação aos íntrons mínimos de baixo e alto conteúdo GC, foi utilizado um método não paramétrico de estimativa de densidade por Kernel (KDE) para avaliar a separação entre as distribuições destas duas populações. O vale entre os os dois picos foi determinado como sendo o valor de referência para a separação. Este método já foi utilizado anteriormente para alcançar esta distinção. (8)

Conforme previamente descrito, a presença e o conteúdo GC dos íntrons mínimos podem ser utilizados como parâmetros classificadores de genes. (8) Desta maneira, nosso estudo envolveu conjuntos amostrais determinados tanto a nível de íntron como de genes classificados de acordo com os seguintes critérios:

- Se o íntron é mínimo ou longo (classificado por número de pares de base)
- Se o íntron mínimo é de baixo ou alto conteúdo GC (medido em %GC)
- Se o íntron é derivado de genes com íntrons mínimos de alto conteúdo GC (HMiG) ou baixo conteúdo GC (LMiG) ou genes sem íntrons mínimos (woMiG)

- Dentro dos grupos HMiG e LMiG, os íntrons também foram divididos entre aqueles que são mínimos (MiHMiG ou MiLMiG) ou longos (LiHMiG ou LiLMiG)

2.3 Análise dos parâmetros obtidos

Para avaliar com mais clareza a distinção entre as células saudáveis, calculamos a diferença entre o valor de *IRratio* de íntrons derivados das células tumorais e seus respectivos valores encontrados nas células saudáveis, obtendo um parâmetro que será chamado de *IRchange* que indica a variação na retenção de íntrons entre os tipos celulares. Para verificar as distribuições dos valores encontrados, foram utilizados gráficos do tipo boxplot para os valores de *IRratio*, que nos permitiu avaliar o grau da distinção entre as células. O teste de distinção entre grupos amostrais Mann-Whitney U (MWW) foi utilizado para avaliar a confiança desta diferença. Entretanto, os bancos de dados utilizados possuem um alto número de *outliers*, o que contribui para tornar a medida do teste MWW insuficiente para análise, pois a distribuição é não-gaussiana e o teste avalia a média. Portanto, uma regressão quantil dividindo o conjunto amostral em 50%/50% foi utilizada em conjunto com o teste MWW para avaliar se a distinção entre as medianas das distribuições é significativa também.

Para análises de proporção dos diferentes grupos de genes contendo íntrons com RI acima de 10% encontrados em pares de células tumoral/saudável, foi utilizado o teste qui quadrado do tipo *goodness of fit*, no qual os valores encontrados para células tumorais foram determinados como observados e para as saudáveis como valores esperados.

2.4 Análise de enriquecimento de termos de ontologia de gênica

Após realizar as filtragens nos resultados do output do IRFinder, buscou-se analisar a função dos genes cujo valor de IRchange de pelo menos um de seus íntrons fosse acima de 0.1, indicando maior retenção na célula cancerosa. Para realizar esta análise, utilizou-se os bancos de dados do *Gene Ontology Resources* (13), estabelecendo uma análise comparativa das divisões já descritas de HMiG, LMiG e woMiG entre os tecidos com sequenciamento profundo. O resultado foi apresentado em tabelas com os termos seguido da significância do p-valor obtido utilizando o teste exato de Fisher e a correção de Bonferroni. Em casos onde o número de termos encontrado é grande demais, resultados menos específicos ou redundantes foram retirados para melhor visualização e entendimento.

3 RESULTADOS E DISCUSSÕES

3.1 Banco de dados e *IRratio*

A fim de obter resultados de análises de RI com menor viés amostral, os dados de sequenciamento foram filtrados com base no tamanho da biblioteca de acordo com o limiar mínimo de seis milhões de bases. Desta forma, foram selecionados os pares de sequenciamento provenientes de tecido tumoral e tecido saudável adjacente: C13xN13, C15xN15, C19xN19 e C23xN23. Cada par é derivado de uma célula cancerosa com um valor de *Gleason Score* distinto. Devido a alta profundidade dos sequenciamentos escolhidos, é esperado que estes bancos de dados contenham uma maior quantidade de íntrons que não possuam problemas como baixa cobertura devido a falta de leituras mapeadas. A tabela 1 estabelece uma comparação entre os íntrons reconhecidos como tendo cobertura adequada para as células cancerosas e saudáveis.

A figura 1 mostra os detalhes do alinhamento dos arquivos FASTQ após a remoção das sequências de baixa qualidade além do *Gleason Score* que indica a severidade do tumor através da sua morfologia comparada às células saudáveis. Todos os experimentos de RNA-Seq são do tipo paired-end e não-direcionados.

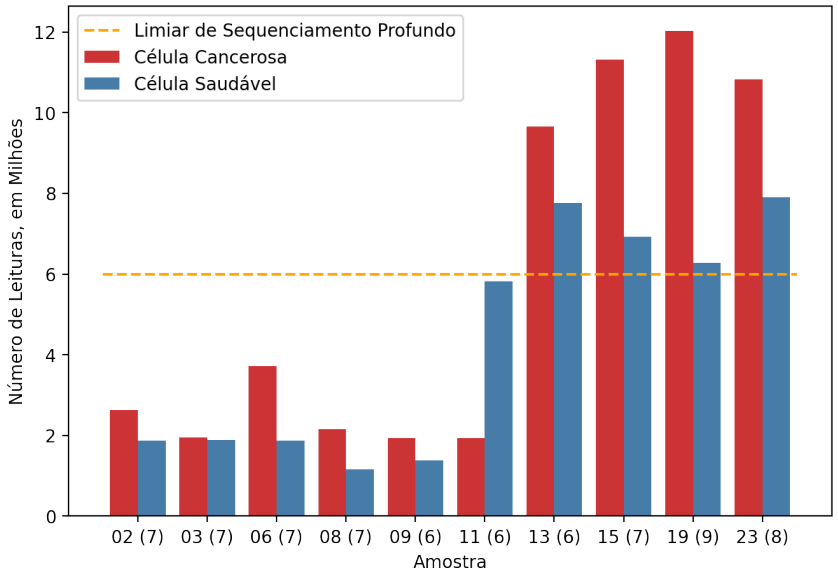


Figura 1 – Comparação do número de leituras em cada conjunto amostral pareado do experimento, o valor em parênteses indica o *Gleason Score* do tumor
Fonte: Elaborada pelo autor

Pela tabela, podemos verificar uma diferença na profundidade do sequenciamento

Tabela 1 – Introns de genes codificantes identificados pelo IRFinder como possuindo cobertura adequada para classificar o valor da retenção intrônica como significativa, o número total de íntrons identificados pelo IRFinder no genoma humano GChr38 foi 250248. Dados do tecido 02 são representativos das demais amostras com baixa profundidade de sequenciamento

Introns com cobertura adequada			
Tipo Celular	Íntrons no Canceroso	Introns no Saudável	Sobreposição
02	3889	2469	1623
13	20910	15239	8477
15	23008	14958	9851
19	23511	13659	8942
23	22426	16696	10464

Fonte: Elaborada pelo autor

entre células cancerosas e saudáveis. Enquanto o tecido tumoral possui em torno de 9% de íntrons com uma cobertura adequada, este valor fica em torno de 6% no tecido saudável.

Análises comparativas entre os tecidos saudáveis e doentes deverão então seguir uma condição de apenas comparar os íntrons que contém uma cobertura adequada nos dois tecidos para não enviesar os resultados finais devido ao tamanho da biblioteca.

3.2 Caracterização e classificação dos íntrons

A primeira distinção a ser feita é em relação ao tamanho do íntron. Para isso, diversas espécies de eucariotos foram escolhidas para servirem como organismos modelo para averiguar se o método escolhido é adequado para este processo. A distribuição de densidade de íntrons do genoma analisado em função do tamanho do íntron em escala logarítmica é um método bastante usado em estudos de tamanho intrônico. (10) Em nosso estudo, usamos a definição de que o ponto de intersecção entre as duas componentes gaussianas derivadas da distribuição obtida serve como valor de referência para classificação de íntrons em duas populações conforme tamanho: íntrons mínimos e íntrons longos (Figura 2).

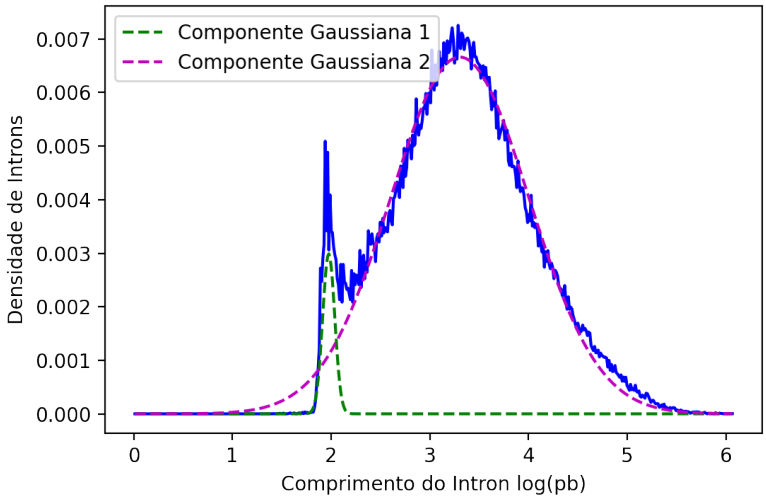


Figura 2 – Componentes gaussianas em uma distribuição log-normal para os íntrons referentes à espécie *Homo sapiens*, os íntrons foram retirados do genoma GChr38 do NCBI (o logaritmo no eixo x está em base 10).

Fonte: Elaborada pelo autor

O mesmo processo foi realizado para sete outras espécies, sendo elas: *Caenorhabditis elegans*, *Caenorhabditis intestinalis*, *Danio rerio*, *Galus galus*, *Mus musculus*, *Pseudonaja textilis* e *Xenopus tropicalis*. O valor da intersecção em cada espécie pode ser visto no histograma da figura 3.

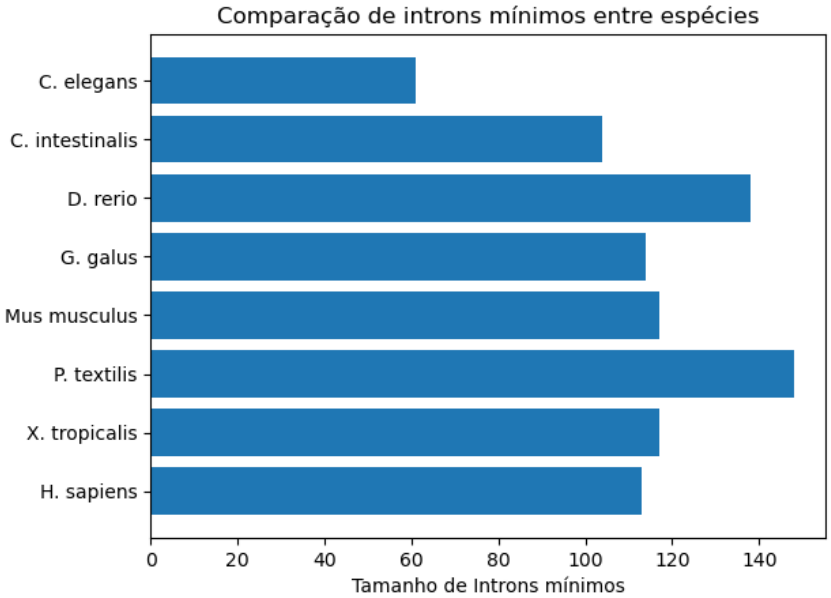


Figura 3 – Histograma mostrando os valores de íntrons mínimos para cada espécie modelo escolhida, foram escolhidos 5 espécies pertencentes aos vertebrados e 2 invertebrados.

Fonte: Elaborada pelo autor

O histograma mostra que o método é efetivo para realizar uma classificação de íntrons em mínimos e longos, sendo condizente com o resultado de outros trabalhos. (9) Desta maneira, foi determinado o valor de 112pb para o comprimento máximo de um íntron mínimo humano.

Tendo obtido um valor de corte para o comprimento, é possível agora estabelecer uma divisão entre os íntrons mínimos que possuem baixo conteúdo GC e alto conteúdo GC. Nesta etapa, utilizamos uma métrica já determinada pelo nosso grupo de pesquisa que observou uma relação de dependência de conteúdo GC de íntrons mínimos e o fenômeno de RI. (8) Por meio da análise de distribuição de íntrons conforme conteúdo GC pelo método KDE, as populações de baixo e alto conteúdo GC foram determinadas, respectivamente, de acordo com o valor do vale entre os picos de 43% (Figura 4).

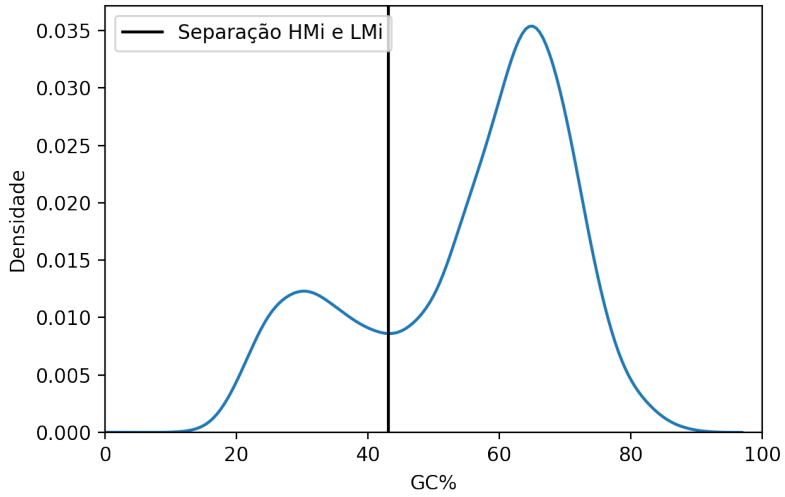


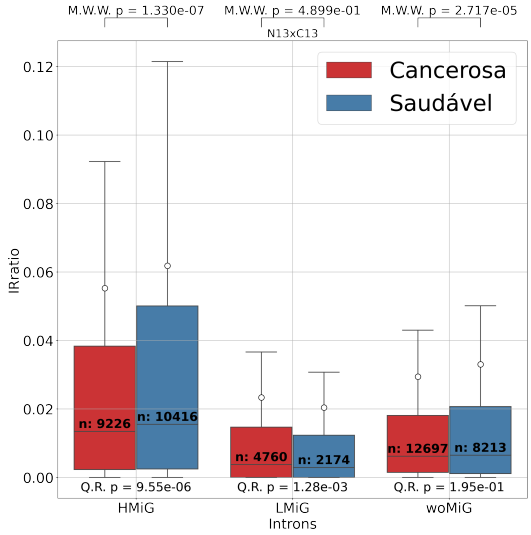
Figura 4 – Gráfico em KDE mostrando onde ocorre a divisão entre a população dos íntrons de baixo GC e alto GC.

Fonte: Elaborada pelo autor

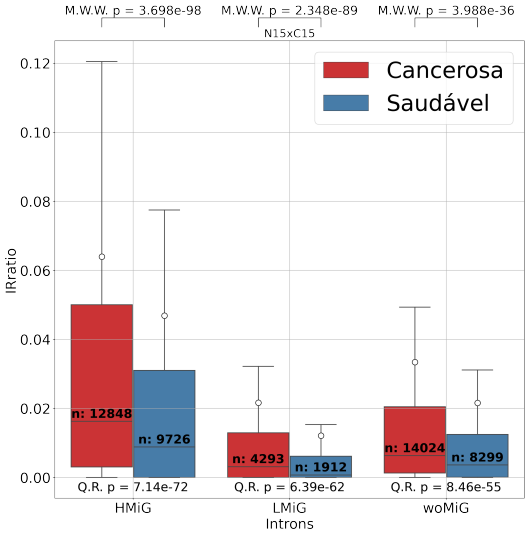
3.3 Análise dos parâmetros obtidos

A primeira análise realizada foi feita com a classificação de íntrons com base no seu gene de origem, se o gene do qual ele é derivado contém: íntrons mínimos de alto conteúdo GC (HMiG), íntrons mínimos de baixo conteúdo GC (LMiG), ou se não possui nenhum íntron mínimo (woMiG). (8) Aqueles íntrons derivados de genes que possuem tanto íntrons mínimos de baixo GC quanto de alto GC não foram levados em consideração, eles compõem uma parcela pequena do total de íntrons humanos (apenas 1,5% de todos

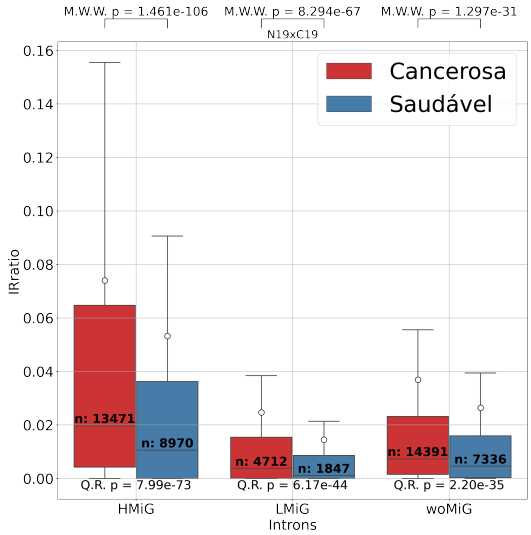
os genes possuem íntrons mínimos de baixo e alto conteúdo GC ao mesmo tempo).



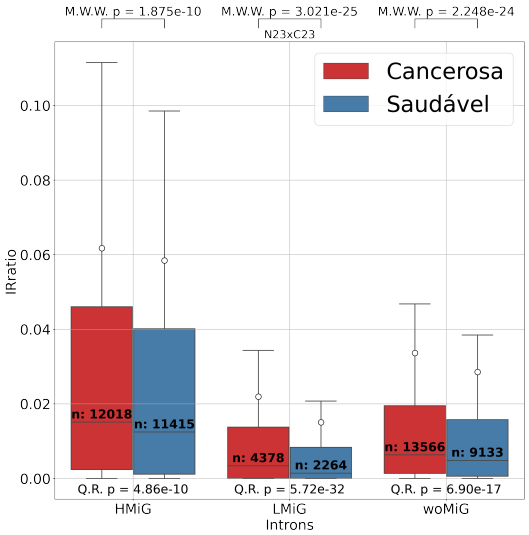
(a) Tecido Nº13



(b) Tecido Nº15



(c) Tecido Nº19



(d) Tecido Nº23

Figura 5 – Histogramas para os 4 tecidos com sequenciamento profundo. No gráfico, estão apresentados tanto o teste de Mann-Whitmann U quanto o teste de regressão quantil para avaliar a distribuição dos íntrons e a diferença entre as suas medianas respectivamente.

Fonte: Elaborada pelo autor

Pelos boxplots, podemos observar o aumento generalizado da RI em células cancerosas conforme já descrito em outros trabalhos. (5–7) Curiosamente, no entanto, nossa classificação de genes em três grupos permitiu a observação mais detalhada do fenômeno

de IR ao verificarmos que há uma diferença significativa de nível de RI entre os três grupos de genes. Os íntrons derivados de LMiGs tendem a ter menores valores de RI, os derivados de woMiGs valores intermediários e os de HMiGs maiores valores. Esta observação se manteve tanto entre as células tumorais como entre as células saudáveis.

O teste de regressão quantil e MWW foram aplicados dois a dois, de modo a comparar as células saudáveis e cancerosas. Dentro dos tecidos analisados, dados do paciente 13 apresentaram maior quantidade de resultados não-significativos ($p > 0.05$) tanto para MWW, no caso do grupo LMiG, quanto para a regressão quantil, no caso do grupo woMiG. A célula 13 é a que possui o *Gleason Score* mais baixo entre todas, e é sabido que a quantidade de íntrons retidos em um tumor está associada também a sua agressividade. (11) Desta maneira, suponhamos que a baixa agressividade da célula cancerosa faz com que não haja uma variação significativa na retenção destes íntrons no geral.

Em todos os boxplots existe uma diferença considerável entre as medianas e as médias dos valores, isto indica que a distribuição não segue uma função normal, tendo uma grande quantidade de *outliers*. Além disso, os valores de RI indicados no eixo y mostram que a grande maioria dos íntrons possui um valor de retenção abaixo de 0.05, o que é baixo para ser biologicamente significativo. (20) Para lidar com estes tipos de dados, decidimos avaliar a mudança de RI entre os grupos de genes contendo íntrons com *IRratio* maior que 0.1 para os pares tumor/saudável por meio de análise de proporção verificada por teste qui-quadrado do tipo *goodness of fit*. As proporções encontradas na célula saudável foram usadas como valores esperados com os quais os valores encontrados em células tumorais foram comparados (Tabela 2).

A tabela 2 nos mostra que a retenção de íntrons se demonstra maior do que o esperado nas células cancerosas nos grupos LMiG e woMi em todos os tipos celulares, enquanto que para o grupo HMiG o nível de íntrons retidos é sempre menor do que o esperado. Ademais, a contribuição do grupo LMiG para o resultado do teste é a maior entre os grupos de genes para todas as amostras analisadas. Desta forma, esta observação é indício de que a mudança de RI em íntrons de LMiG possa estar relacionada com vias importantes nessa doença.

Dois dos três grupos avaliados anteriormente (HMiG e LMiG) podem ser novamente divididos, desta vez em relação ao tamanho do íntron e não apenas de que gene ele

Tabela 2 – Tabelas de teste de Qui-quadrado para os 4 tecidos com sequenciamento profundo, todas as tabelas obtiveram p-valores significativos demonstrando que existe uma diferença entre a linha dos cancerosos e dos saudáveis.

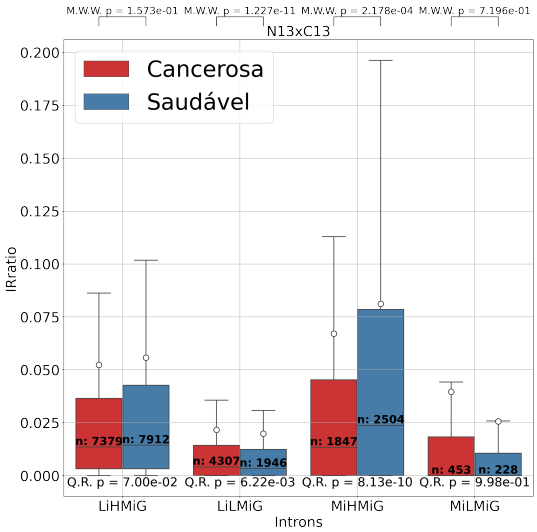
Tabela para o tecido 13 (p-valor = 5.96e-88)				
	HMiG	LMiG	woMi	Soma
Observado	1109	194	668	1971
Esperado	1423.65	64.34	483.01	1971.00
Contribuição	69.54	261.29	70.85	401.68
Tabela para o tecido 15 (p-valor = 1.24e-46)				
	HMiG	LMiG	woMi	Soma
Observado	2012	172	943	3127
Esperado	2314	78	734	3127
Contribuição	39.52	112.61	59.27	211.40
Tabela para o tecido 19 (p-valor = 1.64e-55)				
	HMiG	LMiG	woMi	Soma
Observado	2514	227	1124	3865
Esperado	2876	105	883	3865
Contribuição	45.66	141.13	65.51	252.30
Tabela para o tecido 23 (p-valor = 4.86e-70)				
	HMiG	LMiG	woMi	Soma
Observado	1735	183	877	2795
Esperado	2082	71	641	2795
Contribuição	57.97	174.49	86.74	319.20

Fonte: Elaborada pelo autor

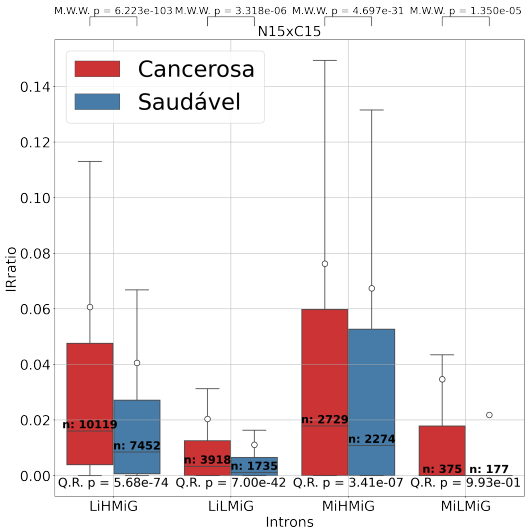
é derivado (se ele é um íntron mínimo ou longo). Então será realizada novamente uma análise de boxplot desta vez levando em consideração apenas as divisões feitas em relação a estes dois grupos.

Os boxplots da figura 6 nos permite verificar onde especificamente a retenção está sendo mais intensa. Mais uma vez, a célula 13 foi a que apresentou maior quantidade de valores não-significativos em relação aos p-valores dos testes (em relação aos grupos LiHMiG e MiLMiG), além disso o grupo MiHMiG reteve mais na célula saudável do que na célula cancerosa de forma significativa. Nos outros tipos celulares, a retenção ocorreu mais intensamente em todos os grupos nas células cancerosas, como de esperado, mas o grupo MiLMiG não mostrou diferença com $p < 0.05$ para a regressão quantil, provavelmente devido ao baixo número de íntrons analisados para ele.

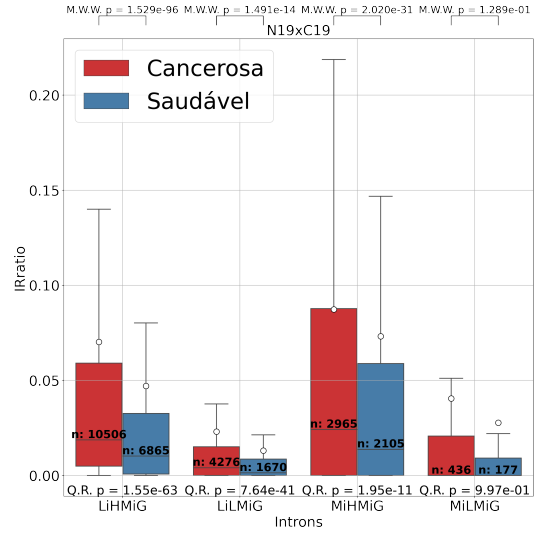
Mais uma vez, foram feitas tabelas com teste Qui-quadrado para avaliar a significância biológica da retenção intrônica. A tabela 3 busca analisar se há uma variação de



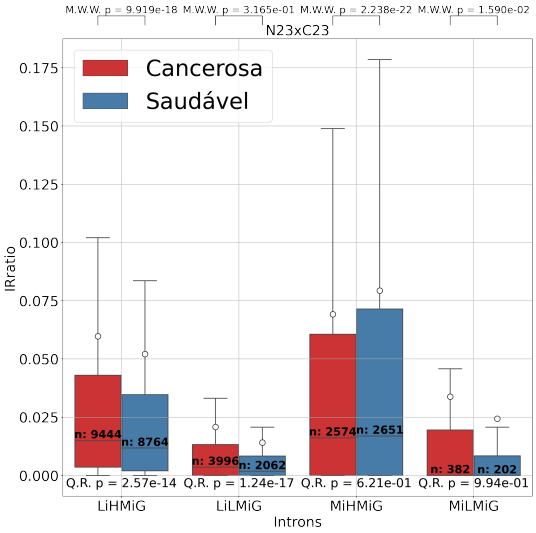
(a) Tecido Nº13



(b) Tecido Nº15



(c) Tecido Nº15



(d) Tecido Nº23

Figura 6 – Histogramas para os 4 tecidos amostrais com sequenciamento profundo, os mesmos testes foram realizados mas desta vez incluindo a distinção do tamanho dos íntrons dividindo os grupos HMiG e LMiG em LiHMiG/MiHMiG e LiLMiG/MiLMiG respectivamente.

Fonte: Elaborada pelo autor

retenção quando o tamanho dos íntrons do grupo HMiG é levado em consideração, então os grupos analisados foram MiHMiG e LiHMiG. Foi observada uma diferença significativa para todas as tabelas, com os íntrons longos tendo uma retenção maior do que a esperada. Isto pode estar associado à diferença no mecanismo de *splicing* destes íntrons. Existem evidências que éxons flanqueados por íntrons longos têm os sítios de reconhecimento de *splicing* dentro de suas próprias sequências, enquanto aqueles flanqueados por íntrons pequenos possuem estes sítios nos íntrons. (21) Isto indica que o maquinário de *splicing* dos íntrons mínimos difere dos longos, e, portanto, eles podem estar sendo afetados de maneiras distintas.

Tabela 3 – Tabelas do teste Qui-quadrado para os 4 tecidos com sequenciamento profundos desta vez envolvendo divisões no grupo HMiG, Todos os tecidos obtiveram um p-valor significativo, com maior retenção do que o esperado para LiHMiG.

Tabela para o tecido 13 (p-valor = 1.25e-10)			
	MiHMiG	LiHMiG	Soma
Observado	246	863	1109
Esperado	345	764	1109
Contribuição	29	13	41
Tabela para o tecido 15 (p-valor = 3.46e-26)			
	MiHMiG	LiHMiG	Soma
Observado	474	1538	2012
Esperado	700	1312	2012
Contribuição	73	39	112
Tabela para o tecido 19 (p-valor = 2.89e-11)			
	MiHMiG	LiHMiG	Soma
Observado	640	1874	2514
Esperado	795	1719	2514
Contribuição	30	14	44
Tabela para o tecido 23 (p-valor = 4.53e-09)			
	MiHMiG	LiHMiG	Soma
Observado	413	1322	1735
Esperado	525	1210	1735
Contribuição	24	10	34

Fonte: Elaborada pelo autor

Uma terceira tabela com teste de Qui-quadrado foi realizada, desta vez focando na diferença de tamanho dentro do grupo LMiG, mas desta vez não foram obtidos resultados significativos, com a retenção sendo bem próxima do que se esperava. O maquinário de *splicing* também varia de acordo com o conteúdo GC dos íntrons, com íntrons longos tendo

um conteúdo GC similar aos éxons que ele flanqueia, enquanto enquanto íntrons pequenos tendem a ter conteúdo GC menor que seus éxons adjacentes (21), o baixo conteúdo GC destes íntrons de um modo geral pode ser uma evidência de que seus maquinários não são distintos, mesmo considerando a diferença de tamanho, mas o número de íntrons (de 183 a 227) avaliados para este caso é muito baixo e estudos futuros precisam ser realizados para confirmar esta hipótese.

Tabela 4 – Tabelas do teste Qui-quadrado para os 4 tecidos com sequenciamento profundos desta vez envolvendo divisões no grupo LMiG. Nenhum dos tecidos obteve um p-valor significativo.

Tabela para o tecido 13 (p-valor = 0.06)			
	MiLMiG	LiLMiG	Soma
Observado	30	164	194
Esperado	22	172	194
Contribuição	3.03	0.38	3.41
Tabela para o tecido 15 (p-valor = 0.84)			
	MiLMiG	LiLMiG	Soma
Observado	20	152	172
Esperado	21	151	172
Contribuição	0.03	0.0	0.03
Tabela para o tecido 19 (p-valor = 0.35)			
	MiLMiG	LiLMiG	Soma
Observado	33	194	227
Esperado	28	199	227
Contribuição	0.75	0.11	0.86
Tabela para o tecido 23 (p-valor = 0.05)			
	MiLMiG	LiLMiG	Soma
Observado	25	158	183
Esperado	17	166	183
Contribuição	3.47	0.36	3.83

Fonte: Elaborada pelo autor

3.4 Análise de enriquecimento de termos de ontologia gênica

Analisando os termos da tabela 5, fica evidente que os tecidos 13 e 23 demonstraram possuir resultados mais significativos, especialmente na questão de componentes celulares. A razão para tantos termos não terem significância para os tecidos 15 e 19 pode ser devido ao baixo número de íntrons derivados do grupo LMiG contendo pelo menos 10% de RI

diferencial entre o par tumor/saudável.

Tabela 5 – Termos de Ontologia de Genes para íntrons derivados do grupo LMiG, a cor amarela denota o componente celular, a verde denota função molecular, a azul processo biológico e vermelho via metabólica do banco "Reactome". ns: p-valor > 0.05; *:0.05 > p-valor > 0.001; **:0.001 > p-valor > 0.0001; ***:0.0001 > p-valor > 0.00001; ****:0.00001 > p-valor

Termos de Ontologia de Genes				
Termo Ontológico	13 P-Valor	15 P-Valor	19 P-Valor	23 P-Valor
nuclear speck	ns	ns	ns	*
nuclear protein-containing complex	ns	ns	ns	*
nucleoplasm	***	ns	ns	**
nuclear lumen	**	ns	ns	**
organelle lumen	*	ns	ns	*
intracellular organelle lumen	*	ns	ns	*
membrane-enclosed lumen	*	ns	ns	*
nucleus	*	ns	ns	***
intracellular membrane-bounded organelle	ns	ns	ns	*
intracellular organelle	ns	ns	ns	*
protein binding	ns	ns	ns	*
regulation of RNA splicing	*	ns	ns	ns
mRNA Splicing - Major Pathway	ns	*	ns	ns
mRNA Splicing	ns	*	ns	ns

Fonte: Elaborada pelo autor

Os termos nos mostram que grande parte dos genes está com sua localização principal em regiões do interior nuclear das células (nucleoplasm, nuclear lumen, etc.) e os tecidos 13 e 15 demonstraram estar associados a termos de regulação de *splicing* de moléculas de mRNA, além de terem a função de ligação a proteína. Em especial, o componente celular *nuclear speck* mostrou-se significativo para a célula 23 e ele é conhecido por ser uma região do nucléolo celular responsável pela síntese de fatores de *splicing*. Isto indica que estas vias estão sendo particularmente afetadas pela retenção intrônica e, como estão associadas a processos de *splicing*, podem acabar por gerar um efeito em cascata que faz com que a retenção de íntrons aumente ainda mais, afetando outras vias e comprometendo o metabolismo celular como um todo.

Termos bastante similares aos do grupo LMiG surgem para os íntrons pertencentes ao grupo woMiG, indicando que eles também estão associados com questões de *splicing*.

Tabela 6 – Termos de Ontologia de Genes para íntrons derivados do grupo woMiG, a cor amarela denota o componente celular, a verde denota função molecular, a azul processo biológico e vermelho via metabólica do banco *Reactome*. ns: p-valor > 0.05; *:0.05 > p-valor > 0.001; **:0.001 > p-valor > 0.0001; ***:0.0001 > p-valor > 0.00001; ****:0.00001 > p-valor

Termos de Ontologia de Genes				
Termo Ontológico	13 P-Valor	15 P-Valor	19 P-Valor	23 P-Valor
spliceosomal complex	*	ns	ns	ns
nuclear speck	***	ns	ns	ns
early endosome	*	ns	ns	ns
nuclear body	**	ns	*	*
nuclear protein-containing complex	*	ns	ns	ns
chromatin	ns	ns	ns	*
nucleoplasm	***	ns	****	**
transcription coregulator activity	ns	ns	ns	*
RNA binding	****	ns	*	**
nucleic acid binding	*	ns	ns	***
transcription regulator activity	ns	ns	ns	**
RNA splicing, via transesterification reactions	**	ns	ns	ns
negative regulation of nucleobase-containing compound	ns	ns	ns	****
regulation of RNA metabolic process	*	ns	ns	****
mRNA 3'-end processing	***	ns	ns	ns
RNA Polymerase II Transcription Termination	**	ns	ns	ns
mRNA Splicing - Major Pathway	**	ns	ns	ns

Fonte: Elaborada pelo autor

Como este grupo é mais numeroso, os resultados são mais significativos em geral, porém ainda há muita diferença entre os tecidos, em especial a célula 15 não mostrou nenhum termo significativo. Isto provavelmente ocorre devido a natureza errática da expressão em células cancerosas, podendo afetar vias metabólicas consideravelmente distintas uma das outras, dificultando o reconhecimento de padrões em diferentes tecidos e também o seu tratamento.

O grupo HMiG (tabela 7) foi o mais distinto de todos, embora ele também envolva termos relacionados a metabolismo de ácidos nucleicos, desta vez, ao invés de *splicing*, há

Tabela 7 – Termos de Ontologia de Genes para íntrons derivados do grupo HMiG, a cor amarela denota o componente celular, a verde denota função molecular, a azul processo biológico e vermelho via metabólica do banco *Reactome*. ns: p-valor > 0.05; *:0.05 > p-valor > 0.001; **:0.001 > p-valor > 0.0001; ***:0.0001 > p-valor > 0.00001; ****:0.00001 > p-valor

Termos de Ontologia de Genes				
Termo Ontológico	13 P-Valor	15 P-Valor	19 P-Valor	23 P-Valor
extracellular vesicle	*	ns	ns	ns
vesicle	****	ns	ns	ns
cytosol	****	****	**	****
intracellular anatomical structure	****	****	***	****
GTPase regulator activity	*	ns	ns	**
transferring phosphorus-containing groups	ns	****	*	ns
nucleoside phosphate binding	ns	**	ns	*
semaphorin-plexin signaling pathway	*	ns	ns	ns
supramolecular fiber organization	*	ns	ns	ns
vesicle-mediated transport	**	ns	ns	ns
nitrogen compound metabolic process	****	****	****	**
Laminin interactions	*	ns	ns	**

Fonte: Elaborada pelo autor

mais termos relacionados a síntese de moléculas de mRNA, com termos como transferases de grupos contendo fósforo e GTPases. Outros termos importantes também estão relacionados com questões de sinalização mediada por semaforinas, em especial SEMA classe 3 que estão retidas em todas as células e estão relacionadas com questões de motilidade de células do câncer de próstata. (22)

Por último, também foram encontrados termos nos tecidos 13 e 23 que evidenciam possíveis efeitos sobre a matriz extracelular das células, como é o caso do termo de organização supramolecular em processos biológicos e também interação com lamininas. Estes termos estão relacionados com questões como adesão celular e, caso sejam afetados negativamente, podem levar ao desprendimento das células de sua matriz de origem iniciando uma metástase. Já é sabido que células tumorais do câncer de próstata expressam os mRNAs característicos de proteínas como laminina 5, mas defeitos pós-transcricionais impedem o seu funcionamento adequado, e a retenção intrônica é indicada como possível

causa deste problema. (23)

4 CONCLUSÕES E CONSIDERAÇÕES FUTURAS

O projeto demonstrou que a retenção de íntrons em tecidos cancerosos é significativamente maior que seus correspondentes saudáveis, confirmando o que já era conhecido dos transcriptomas destes tipos celulares. Além disso, foi possível validar um método de discrição entre íntrons mínimos e longos através do ponto de intersecção entre duas regressões gaussianas para diversas espécies de vertebrados (19), incluindo *Homo sapiens*. No final, o tamanho de 112 pares de base foi adequado para os objetivos do projeto.

Outro parâmetro validado foi a distinção entre as populações de conteúdo GC baixo e alto presente dentro dos íntrons mínimos, com 43% sendo o valor de corte mais adequado. A partir deste e do parâmetro de tamanho, 3 populações foram capazes de ser claramente distinguidas: íntrons mínimos de alto conteúdo GC, baixo conteúdo GC, e íntrons longos (8).

Ficou claro também, através das análises de boxplots, que a presença de um íntron mínimo em um gene mostrou-se importante para a retenção de seus íntrons. Dividiu-se então as populações em HMiG, LMiG e woMiG.

Em relação a distribuição das retenções em tecidos cancerosos comparada a distribuição esperada dos tecidos normais ficou claro que o grupo HMiG reteve significativamente menos do que os grupos LMiG e woMiG. Além disso, dentro do grupo HMiG a retenção dos íntrons longos ocorreu mais do que o esperado. A origem evolutiva dos íntrons longos e também íntrons mínimos de baixo GC pode explicar o porquê disto ocorrer (24).

A análise dos termos GO esclareceu que íntrons pertencentes aos grupos woMiG e LMiG retidos em maior quantidade nas células cancerosas estão envolvidos com os mecanismos de *splicing*, sendo uma possível fonte da retenção generalizada que se observa em câncer. Dentro dos íntrons retidos do grupo HMiG termos relacionados à síntese de moléculas de RNA foram encontrados, além de vias como interação de laminina, fibras supramoleculares e sinalização por meio de semaforinas. Estas vias afetadas já são conhecidas por terem suas expressões e produtos moleculares modificados em células cancerosas da próstata, ocasionando perda de motilidade, adesão celular e replicação celular descontrolada (22, 23).

As análises deixaram bastante evidente que a RI é um fenômeno com grande in-

fluência no metabolismo de células cancerosas e também saudáveis, mas pouco foi averiguado em relação aos produtos moleculares destas retenções e os seus destinos biológicos. É sabido que a íntrons retidos podem ter uma gama de diferentes destinos (20) com o principal sendo NMD, mas também pode haver a geração de proteínas trucadas ou fora de fase que podem ter ou não uma função biológica distinta. Estudos futuros poderão estudar por meio de análises de marcadores como códons de parada o destino destes mRNAs com RI e as suas implicações biológicas e patológicas para o câncer de próstata.

Referências

1 GE, Y.; PORSE, B. T. The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. **BioEssays**, Wiley, v. 36, n. 3, p. 236–243.

2 JACOB, A. G.; SMITH, C. W. J. Intron retention as a component of regulated gene expression programs. **Human Genetics**, Springer Science and Business Media LLC, v. 136, n. 9, p. 1043–1057.

3 PLEISS, J. A. *et al.* Rapid, transcript-specific changes in splicing in response to environmental stress. **Molecular Cell**, Elsevier BV, v. 27, n. 6, p. 928–937.

4 SYED, N. H. *et al.* Alternative splicing in plants – coming of age. **Trends in Plant Science**, Elsevier BV, v. 17, n. 10, p. 616–623.

5 DVINGE, H.; BRADLEY, R. K. Widespread intron retention diversifies most cancer transcriptomes. **Genome Medicine**, Springer Science and Business Media LLC, v. 7, n. 1.

6 JUNG, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. **Nature Genetics**, Springer Science and Business Media LLC, v. 47, n. 11, p. 1242–1248.

7 SMART, A. C. *et al.* Intron retention is a source of neoepitopes in cancer. **Nature Biotechnology**, Springer Science and Business Media LLC, v. 36, n. 11, p. 1056–1058.

8 FARIA, L. Z. de. **Study of evolution and architecture of minimal introns**. 164 p. Dissertação (Mestrado) — Instituto de Fisica de Sao Carlos, Universidade de Sao Paulo, 2020.

9 JIAYAN, W. *et al.* Systematic analysis of intron size and abundance parameters in diverse lineages. **Science China Life Sciences**, Springer Science and Business Media LLC, v. 56, n. 10, p. 968–974.

10 YU, J. Minimal introns are not "junk". **Genome Research**, Cold Spring Harbor Laboratory, v. 12, n. 8, p. 1185–1189.

11 ZHANG, D. *et al.* Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. **Nature Communications**, Springer Science and Business Media LLC, v. 11, n. 1.

12 BRAY, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, Wiley, v. 68, n. 6, p. 394–424.

13 MI, H. *et al.* PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. **Nucleic Acids Research**, Oxford University Press (OUP), v. 47, n. D1, p. D419–D426.

14 KANNAN, K. *et al.* Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 108, n. 22, p. 9172–9177.

15 MIDDLETON, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. **Genome Biology**, Springer Science and Business Media LLC, v. 18, n. 1.

16 DOBIN, A. *et al.* STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, Oxford University Press (OUP), v. 29, n. 1, p. 15–21.

17 QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, Oxford University Press (OUP), v. 26, n. 6, p. 841–842.

18 BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for illumina sequence data. **Bioinformatics**, Oxford University Press (OUP), v. 30, n. 15, p. 2114–2120.

19 LIM, L. P.; BURGE, C. B. A computational analysis of sequence features involved in recognition of short introns. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 98, n. 20, p. 11193–11198.

20 GRABSKI, D. F. *et al.* Intron retention and its impact on gene expression and protein diversity: A review and a practical guide. **WIREs RNA**, Wiley, v. 12, n. 1.

21 AMIT, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. **Cell Reports**, Elsevier BV, v. 1, n. 5, p. 543–556.

22 ALTO, L. T.; TERMAN, J. R. Semaphorins and their signaling mechanisms. *In*: **Methods in Molecular Biology**. .: New York: Springer. p. 1–25.

23 HAO, J. *et al.* Investigation into the mechanism of the loss of laminin 5 ($\alpha 3\beta 3\gamma 2$) expression in prostate cancer. **The American Journal of Pathology**, Elsevier BV, v. 158, n. 3, p. 1129–1135.

24 GELFMAN, S. *et al.* Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. **Genome Research**, Cold Spring Harbor Laboratory, v. 22, n. 1, p. 35–50.