Aplicação de regressão logística para identificação de miRNAs como biomarcadores potenciais em adenocarcinoma pulmonar

Bethina da Rocha Camargo[1], **Rainer Marco Lopez Lapa**[2], **Patricia Pintor dos Reis**[3], **Rogério Antonio de Oliveira**[4]

1 Introdução

O câncer de pulmão é um dos cânceres mais comuns e líder em mortalidade, com 1.8 milhões de casos novos e 1.3 milhões de óbitos por ano, no mundo (Globocan, 2018). No Brasil, as estimativas de incidência para 2018/2019 apontam o câncer de pulmão como o segundo mais frequente, com ocorrência de 31 mil novos casos, sendo a maioria em homens. Na Região Sudeste, a incidência de carcinomas de pulmão em 2018 foi a terceira mais frequente com 19,22 casos a cada 100.000 homens e 12,72 casos para cada 100.000 mulheres (INCA, 2018).O carcinoma pulmonar é geralmente diagnosticado em estadiamento avançado da doença, principalmente porque os sintomas não são percebidos nas fases iniciais de desenvolvimento da doença. O diagnóstico tardio é apontado como responsável pela baixa taxa de sobrevida dos pacientes, de 10% a 15% aos 5 anos após o diagnóstico (INCA, 2018). O carcinoma pulmonar de células não pequenas é o principal subtipo de câncer de pulmão, sendo que compreende 85% dos casos e é subdividido em vários subtipos histológicos, sendo o adenocarcinoma o mais frequentemente diagnosticado.

Nos últimos anos, têm-se evidenciado o papel de RNAs não codificadores, especialmente os miRNAS em diversas doenças, incluindo o câncer. Os miRNAS são RNAs pequenos, que não codificam proteínas, entretanto têm capacidade de regular a expressão gênica. Os miRNAs desempenham um papel regulatório importante em processos biológicos, tais como o desenvolvimento embrionário, a proliferação, diferenciação celulares e apoptose (morte celular).

Notavelmente a descoberta dos miRNAs como importantes moduladores da expressão gênica e seu envolvimento em doenças crônico-degenerativas como o câncer tem sugerido que estas moléculas podem constituir biomarcadores clinicamente aplicáveis, para melhorar o diagnóstico, prognóstico e o tratamento de pacientes com câncer. A identificação de alterações na expressão de miRNAs em adenocarcinoma pulmonar pode contribuir para o desenvolvimento de miRNAs como biomarcadores no câncer. Os biomarcadores moleculares podem ser úteis no desenvolvimento de estratégias para detecção precoce de doença, diagnóstico, determinação do prognóstico e para aplicação de terapêuticas mais precisas.

O modelo de regressão logística foi utilizado para identificar quais alterações na expressão génica dos miRNAs que poderiam estar relacionados ao processo de câncer das células de adenocarcinoma pulmonar, com o intuito de identificar e classificar com mais precisão as células normais de tumorais (Madadi et al., 2014).

[1]UNESP-Programa de Pós-graduação em Biometria. e-mail: *bethina.camargo@unesp.br*
[2]UNESP-Programa de Pós-graduação em Genética. e-mail: *reimco2@gmail.com*
[3]UNESP-Departamento de Cirurgia e Ortopedia. e-mail: *preis@fmb.unesp.br*
[4]UNESP-Departamento de Bioestatística. e-mail: *rogerio.oliveira@unesp.br*

2 Objetivo

Identificar alterações na expressão de miRNAs e a relação com o prognóstico de pacientes com adenocarcinoma pulmonar, utilizando análise de regressão logística múltipla.

3 Material e Métodos

Nesse estudo foram utilizados dados de sequenciamento de alto desempenho de miRNAs (miRNA-Seq) de 508 amostras de adenocarcinoma pulmonar e 45 amostras de tecido pulmonar normal. O banco de dado possui a expressão génica de 709 miRNAs quantificados em adenocarcinoma de células normais e tumorais de pulmão, considerando também status e localização do tumor, sexo, etnia e idade do paciente, coletados no portal Atlas Genômico do Câncer (TCGA - https://por- tal.gdc.cancer.gov). Esse banco de dados é útil para o desenvolvimento de diversas pesquisas científicas, sendo que os dados publicamente disponíveis são resultado de uma colaboração entre o Instituto Nacional do Câncer (NCI) e o Instituto Nacional de Pesquisa do Genoma Humano (NHGRI).

Foram calculados os valores mínimos, medianos e máximos da expressão de miRNAs. Devido à assimetria na distribuição dos valores observados, foi aplicado o teste não-paramétrico de Wilcoxon para verificar se as medianas dos grupos (tumores vs. normais) eram estatisticamente significativas, considerando um nível de significância de 5%.

Segundo Hosmer e Lemeshow, a regressão logística é uma análise de regressão para variáveis dicotômicas (1 - sucesso e 0 - fracasso), é uma técnica que nos permite estimar a probabilidade associada à ocorrência de determinado evento (Hosmer Jr, 1989), em face de um conjunto de variáveis explanatórias, comumente aplicadas nas áreas das ciências da saúde.

O modelo logístico apresenta algumas vantangens em sua utilização, como:

- Facilidade para lidar com variáveis independentes categórica;
- Fornece resultados em termos de probabilidade;
- Facilidade de classificação de indivíduos em categorias;
- Requer pequeno número de suposições;
- Alto grau de confiabilidade.

O modelo de regressão logistíca pode ser expresso:

$$\text{logito}(p_i) = \log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} = \beta' \mathbf{x_i},$$

em que p_i é a probabilidade da célula ser tumoral (Y=1), o vetor β são os coeficientes da regressão logística, estimados pelo método de máximo verossimilhança, e x_i são as variáveis independentes associadas a i-ésima célula.

Observa-se que a curva logística tem um comportamento probabilístico no formato da letra S, o que é uma característica da regressão logística.

- Quando $\beta' \mathbf{x_i} \to +\infty$, então $P(Y=1) \to 1$.

- Quando $\beta' \mathbf{x_i} \to -\infty$, então $P(Y = 1) \to 0$.

Por essa razão, ao interpretar os coeficientes da regressão logística, opta-se pela interpretação de e^{β} e não diretamente de β. Contudo, quando se utiliza o modelo logístico do ponto de vista de discriminação entre grupos, não há grande interesse na interpretação dos coeficientes. Para utilizar o modelo de regressão logística na discriminação de dois grupos, a regra de classificação é a seguinte:

- se $P(Y = 1) > 50\%$, então classifica-se $Y = 1$;
- em caso contrário, classifica-se $Y = 0$.

Para comparar a qualidade do ajuste do modelo regressão logística pode-se utilizar a estatística C (Linden, 2006). Os valores para esta medida variam de 0,5 a 1,0. Sendo que 0,5 indica que o modelo não é melhor que o acaso em fazer uma predição de associação em um grupo e o valor 1,0, indica que o modelo identifica perfeitamente aquelas células dentro de seu grupo. Os modelos de regressão logística são considerados razoáveis quando a estatística C é maior que 70% (Guo, 2018).

4 Resultados

Os dados correspondem a 554 amostras (508 tumores e 45 normais), sendo que foram avaliados 709 miRNAs expressos nessas amostras. A expressão de miRNAs foi correlacionada com os dados clínicos e histopatológicos dos pacientes. Em relação aos pacientes, tem-se 53,16% são mulheres e 76,31% são caucasianos. Em relação ao estadiamento do tumor, 55,5% correspondem ao estadio I, 23,99% corresponde ao estadio II e 20,51% corresponde aos estadios III e IV. O número de elevado de células no estágio I ocorre porque as células coletadas são provenientes de tumores retirados cirurgicamente. A idade média e o desvio-padrão são respectivamente 66,9 e 11,6 anos, sendo que 64,4% dos pacientes estavam vivos até a quantificação dos miRNAS que compoem a expressão gênica das células.

Os miRNAS estatisticamente significativos encontrados no teste não-paramétrico de Wilcoxon estão apresentados na tabela 1, as setas $\uparrow$ e $\downarrow$ denotam a expressão gênica aumentada ou diminuida respectivamente em relação aos valores observados para as células normais.

Um modelo de regressão logística multipla foi ajustada para os 709 miRNAS e as demais variáveis clínico-histopatológicas. Foram encontrados 18 miRNAS que explicam significativamente o aparecimento de células tumorais. As variáveis clinico-histopatológicas não foram estatisticamente significativas para explicar o aparecimento de células tumorais. O teste de adequação do modelo ao conjunto de dados não apresentou falta de ajuste (valor p = 0,939). A estatística C apresentou o valor de 99,9%, indicando qualidade na separação das células em tumorais e normais. Os miRNAS significativos estão apresentados na Tabela 1. Os parâmetros ajustados para o modelo de regressão logística estão apresentados na tabela 2. Os miRNAs com expressão génica desregulada, que foram estatísticamente significativos, na análise de regressão logística apontou alguns miRNAs iguais aos apresentados no trabalho de (Cinegaglia et al., 2016).

Tabela 1: miRNAS significativos encontrados pelo teste de Wilcoxon.

miRNAS[1]	**Célula Normal**			**Célula Cancerígena**			**Valor P**
	Mínimo	**Mediana**	**Máximo**	**Mínimo**	**Mediana**	**Máximo**	
hsa_miR_143_3p ↓	186991,0	600705,0	698155,0	207043,0	437562,5	768001,0	0,0001
hsa_miR_144_5p ↓	10000000	225813,0	359566,0	1347,0	99701,5	298277,0	0,0001
hsa_miR_20a_3p↑	0,000	9968,0	50670,0	0,000	40497,5	151138,0	0,0001
hsa_miR_21_5p ↑	92566,0	497239,0	664743,0	398236,0	746033,5	1128308,0	0,0001
hsa_miR_28_5p ↑	22191,0	85087,0	138715,0	30818,0	116794,0	242272,0	0,0001
hsa_miR_30a_5p ↓	144739,0	411687,0	578663,0	101842,0	293903,5	562749,0	0,0001
hsa_miR_320d ↑	0,000	20.000.000	49164,0	0,000	9968,0	96984,0	0,0020
hsa_miR_33b_5p ↑	0,000	10.000.000	61771,0	0,000	38233,0	123322,0	0,0001
hsa_miR_409_5p ↑	0,000	10.000.000	22575,0	0,000	37113,0	135105,0	0,0001
hsa_miR_429 ↑	0,000	55022,0	94970,0	20.000.000	101978,0	360635,0	0,0001
hsa_miR_486_5p ↓	38744,0	130364,0	254847	8290	63110,5	323843	0,0001
hsa_miR_490_3p ↓	0,000	16671,0	70943,0	0,000	0,000	162279,0	0,0001
hsa_miR_493_5p ↑	10000000	22575,0	71525,0	0,000	32540,5	115451,0	0,0005
hsa_miR_509_3_5p ↓	0,000	30.000.000	44677,0	0,000	0,000	84502,0	0,0004
hsa_miR_577 ↑	0,000	0,000	25625,0	0,000	28197,5	139129,0	0,0001
hsa_miR_629_3p ↑	0,000	9966,0	52193,0	0,000	40084,5	111298,0	0,0001

[1] ↓ e ↑ representam se o miRNA está subexpresso ou superexpresso respectivamente.

Tabela 2: Estimativas dos parâmetros do modelo de regressão logística.

miRNAS	**Estimativa**	**Erro Padrão**	**Teste de Wald**	**Valor P**
hsa_miR_143_3p	-0,000030	0,000015	4,2606	0,039
hsa_miR_144_5p	-0,000050	0,000016	8,3782	0,004
hsa_miR_20a_3p	0,000185	0,000092	4,0705	0,044
hsa_miR_21_5p	0,000051	0,000017	9,0938	0,003
hsa_miR_28_5p	0,000169	0,000067	6,3862	0,012
hsa_miR_30a_5p	-0,000060	0,000023	5,9562	0,015
hsa_miR_320d	0,000212	0,000101	4,3716	0,037
hsa_miR_33b_5p	0,000360	0,000143	6,3143	0,012
hsa_miR_362_3p	-0,000250	0,000102	6,1206	0,013
hsa_miR_369_5p	-0,000080	0,000032	6,3043	0,012
hsa_miR_409_5p	0,000420	0,000183	5,2417	0,022
hsa_miR_429	0,000220	0,000085	6,6892	0,001
hsa_miR_486_5p	-0,000110	0,000030	13,0855	0,003
hsa_miR_490_3p	-0,000230	0,000075	9,4291	0,002
hsa_miR_493_5p	0,000157	0,000052	9,2389	0,002
hsa_miR_509_3_5p	-0,000170	0,000071	5,8666	0,015
hsa_miR_577	0,000669	0,000209	10,1961	0,001
hsa_miR_629_3p	0,000200	0,000055	13,0560	0,003

5 Conclusão

Baseado dos resultados encontrados anteriormente, tem-se que os miRNAS:

- miR_143_3p, miR_144_5p, miR_30a_5p, miR_362_3p, miR_369_5p, miR_486_5p, miR_490_3p, miR_509_3_5p,

estão com expressão diminuída em relação aos valores medianos encontrados para as amostras normais. Os miRNAS:

- miR_20a_3p, miR_21_5p, miR_28_5p, miR_320d, miR_33b_5p, miR_409_5p, miR_429, miR_493_5p, miR_577, miR_629_3p,

estão com expressão aumentada em relação a mediana das amostras normais. A expressão aumentada ou diminuída dos miRNAs em amostras de adenocarcinoma pulmonar indica que os miRNAs identificados devem estar associados ao desenvolvimento e ou progressão tumoral para pacientes com adenocarcinoma de pulmão. Em trabalhos futuros, estes miRNAS serão investigados para verificar a sua importância no tempo de sobrevivência destes pacientes.

Agradecimentos

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

Referencias Bibliográficas

CINEGAGLIA, Naiara C. et al. Integrative transcriptome analysis identifies deregulated microRNA-transcription factor networks in lung adenocarcinoma. Oncotarget, v. 7, n. 20, p. 28920, 2016.

GLOBOCAN. Lung Cancer. Estimated Incidence, Mortality and Prevalence Worldwide in 2012. http://gco.iarc.fr/today/fact-sheets-populations, 2018. Accessed: 2019-28-02.

GUO, Changbin; YO, S.; JANG, Woosung. Evaluating predictive accuracy of survival models with PROC PHREG. 2018.

HOSMER JR, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. Applied logistic regression. John Wiley & Sons, 2013.

INCA. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2018: Incidência de Câncer no Brasil. Rio de Janeiro. https://www.inca.gov.br/publicacoes/livros/estimativa-2018-incidencia-de-cancerno-brasil, 2018. Accessed: 2019-28-02.

LINDEN, Ariel. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. Journal of evaluation in clinical practice, v. 12, n. 2, p. 132-139, 2006.

MADADI, Mahboubeh et al. Analyzing factors associated with women?s attitudes and behaviors toward screening mammography using design-based logistic regression. Breast cancer research and treatment, v. 144, n. 1, p. 193-204, 2014.