



**UNIVERSIDADE FEDERAL DA BAHIA**

**INSTITUTO MULTIDISCIPLINAR  
EM SAÚDE**

**PROGRAMA DE PÓS-GRADUAÇÃO EM BIOCIÊNCIAS**

**FERNANDA CASTRO PIRES DOS SANTOS**

**IDENTIFICAÇÃO E CLASSIFICAÇÃO FUNCIONAL DE  
GENES DIFERENCIALMENTE EXPRESSOS  
COMPARTILHADOS EM ESTUDOS DE CÂNCER DE  
PULMÃO**

Vitória da Conquista, BA  
13 de Novembro de 2018

**FERNANDA CASTRO PIRES DOS SANTOS**

**IDENTIFICAÇÃO E CLASSIFICAÇÃO FUNCIONAL DE  
GENES DIFERENCIALMENTE EXPRESSOS  
COMPARTILHADOS EM ESTUDOS DE CÂNCER DE  
PULMÃO**

Dissertação apresentada ao Programa de Pós-Graduação em  
Biotecnologia, Universidade Federal da Bahia, como requisito  
para obtenção do título de Mestre em Biotecnologia.

Orientador: Prof. Dr. Sávio Torres de Farias  
Universidade Federal da Paraíba, João Pessoa, PB

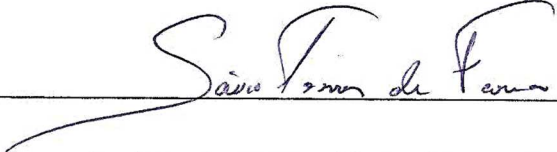
Co-orientador: Prof. Dr. Leandro Martins Freitas  
Universidade Federal da Bahia, Vitória da Conquista, BA

Vitória da Conquista, BA  
13 de Novembro de 2018

**FERNANDA CASTRO PIRES DOS SANTOS**

ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA À OBTENÇÃO DO GRAU DE MESTRE EM  
BIOCIÊNCIAS E APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-  
GRADUAÇÃO EM BIOCIÊNCIAS, UNIVERSIDADE FEDERAL DA BAHIA.

VITÓRIA DA CONQUISTA – BA, 13/11/2018.




DR. SÁVIO TORRES DE FARIAS (ORIENTADOR)  
(UNIVERSIDADE FEDERAL DA PARAÍBA)



DR. LEANDRO MARTINS DE FREITAS (CO-ORIENTADOR)  
(UNIVERSIDADE FEDERAL DA BAHIA)



PROF. DR. BRUNO LOPES BASTOS (EXAMINADOR)  
(UNIVERSIDADE FEDERAL DA BAHIA)



PROF.<sup>a</sup> DR<sup>a</sup> CÍNTIA RODRIGUES MARQUES (EXAMINADORA)  
(UNIVERSIDADE FEDERAL DA BAHIA)

Dedico essa dissertação aos meus pais, razão de toda minha trajetória nos estudos. Dedico a minha mãe, que se esforçou todo esse tempo para manter meus estudos, você sempre será exemplo de garra, determinação e luta. Dedico ao meu pai que não teve a oportunidade de graduar, tão pouco pós-graduar, e mesmo assim, se empenhou para conhecer a Biotecnologia, você sempre será inspiração. Todos esses anos me dediquei a colocar em prática todos os conselhos dos meus pais, aprendi a valorizar, reconhecer, apreciar, dedicar, escutar, silenciar e estudar. Sempre com muita seriedade, honestidade e compromisso. Dedico, esse estudo ao meu irmão que sempre me apoiou no sonho de ser pesquisadora. Dedico ao meu namorado que sempre me deu forças para continuar os estudos. Dedico a todo cidadão de bem, principalmente aos que fizeram o melhor diante das adversidades.

## **AGRADECIMENTOS**

Aos professores Leandro Martins de Freitas e Sávio Torres de Farias pelas orientações e ensinamentos em torno da Biotecnologia. Também agradeço toda a compreensão durante esse tempo. A banca por ter aceitado o convite para a avaliação desse trabalho. Aos professores do programa Biociências e aos que não fazem parte do programa, por ter contribuído de forma importantíssima no meu conhecimento científico. Ao grupo GENBIO pela amizade e compartilhamento de conhecimento. Aos novos amigos do mestrado e aos que já conheciam foi muito bom compartilhar ideias e conhecimentos com vocês. A FAPESB pelo apoio financeiro. Aos amigos da graduação que sempre me deram apoio nesses anos de estudo, em especial Patricia, Manassés, Joice, Jessica e Poliana. Aos meus familiares que de longe deram força. Aos meus pais Pedro Pires dos Santos e Leidimar Pereira de Castro Santos por todo incentivo, compreensão nos momentos que não estive presente fisicamente, pelo apoio financeiro para me manter em outra cidade por anos, ao meu irmão Rafael Castro Pires dos Santos pelo apoio, amizade e companheirismo. Ao meu namorado José Antônio por toda paciência, amor, amizade e companheirismo.

“O conhecimento é o nosso melhor bem comum”  
Jacques Dubochet  
Nobel de Química 2017

SANTOS, Fernanda Castro Pires dos. Identificação e classificação funcional de genes diferencialmente expressos compartilhados em estudos de câncer de pulmão. 79 f.: il. 2018. Dissertação (Mestrado) – Instituto Multidisciplinar de Saúde, Universidade Federal da Bahia, Vitória da Conquista, 2018.

## RESUMO

O câncer de pulmão não pequenas células (CPNPC) é o câncer com efeito adverso, mais letal. O mecanismo específico desta doença não é totalmente compreendido. Para analisar sistematicamente os genes associados com o subtipo adenocarcinoma de pulmão, dados de genes foram extraídos utilizando inicialmente o Gene Expression Omnibus conjunto de dados (GEO), em seguida foi feita para as proteínas análise de interação proteína-proteína (PPI) usando banco de dados APID. O gene ontology (GO) e a análise de enriquecimento da via foi realizado para os genes diferencialmente expressos. A rede PPI gerada por APID foi visualizada pelo Cytoscape. As regiões funcionais foram analisadas utilizando-se plugin Golorize. Na meta-análise 213 genes diferencialmente expressos foram identificados, 49 genes superexpressos e 164 subexpressos. A análise de enriquecimento gênico retornou 10 termos GO enriquecidos. GO revelou 2 processos biológicos de relevância para os genes superexpressos, divisão celular (GO: 0051301) e regulação positiva do processo apoptótico (GO: 0043065). Foram associados com resposta de subexpressão 8 processos biológicos, transdução de sinal (GO: 0007165), adesão celular (GO: 0007155), resposta imune (GO: 0006955), resposta imune inata (GO: 0045087), apoptótica (GO: 0006915), resposta inflamatória (GO: 0006954), regulação negativa da proliferação celular (GO: 0008285), regulação positiva da célula proliferação (GO: 0008284). A expressão dos genes foi usada para classificar as amostras em normal ou tumoral usando *heatmap* e análise de componentes principais (PCA). A classificação das amostras revela a separação dos dois grupos com taxa verdadeiro positivo de 97% e taxa verdadeiro negativo de 99%. Este estudo utiliza de metodologias computacionais para revelar a expressão gênica do adenocarcinoma de pulmão e identificar a assinatura gênica do tumor. A confirmação de 13 genes diferencialmente expressos por imuno-histoquímica indica esses genes como críticos para desenvolvimento do câncer de pulmão. Esses genes podem ser usados para o desenvolvimento de teste molecular e auxiliar no diagnóstico.

**Palavras-chave:** *microarray*, adenocarcinoma de pulmão, expressão de genes.

SANTOS, Fernanda Castro Pires dos. Identification and functional classification of differentially expressed genes shared in lung cancer studies. 79 f.: ill. 2018. Dissertation (Master degree) - Multidisciplinary Institute of Health, Federal University of Bahia, Vitória da Conquista, 2018.

## ABSTRACT

Lung cancer is not small cells (NSCLC) is the cancer with adverse, more lethal effect. The mechanism in question is not fully understood. Abstract related to genes associated with the subtype of adenocarcinoma of the lung, genes extracted from the Gene Expression expression (GEO), were then presented as protein expression cell (PPI) cells, using the database. APID data. O gene ontology (GO) and a pathway enrichment analysis was performed for differentially expressed genes. The PPI network generated by APID was viewed by Cytoscape. Regions were included using the Golorize plugin. In the meta-analysis 213 differentially expressed genes were identified, 49 genes upexpressed and 164 downexpressed. A gene enrichment analysis returned 10 enriched GO terms. 2 biological processes of relevance for the superexpressive genes, cell cycle (GO: 0051301) and positive control of the apoptotic process (GO: 0043065). (GO: 0007165), cellular statement (GO: 0007155), immune response (GO: 0006955), innate immune response (GO: 0045087), apoptotic (GO: 0006915), inflammatory response (GO: 0006954), negative regulation of cell proliferation (GO: 0008285), positive control of cell proliferation (GO: 0008284). Gene expression was used to classify samples into normal or tumor using heatmap and principal component analysis (PCA). The classification of the samples reveals the clustering of the two groups with a true positive rate of 97% and a true negative rate of 99% TNR. This study uses computational methods to reveal the gene expression of lung adenocarcinoma and to identify the gene signature of the tumor. The confirmation of 13 differentially expressed genes by immunohistochemistry indicates these genes are critical for lung cancer development. These genes can be used for the development of molecular test and help in the lung cancer diagnosis.

**Keywords:** *microarray*, lung adenocarcinoma, gene expression.



## LISTA DE ILUSTRAÇÕES

Figura 1	Desenvolvimento do câncer de pulmão .....	18
Figura 2	Modelo de câncer baseado em <i>priming</i> epigenético de tumor .....	19
Figura 3	Representação da regulação do EGFR WT (selvagem) e EGFR mutante por ERFFI1 ....	23
Figura 4	Presença de mutações nas isoformas de Ras.....	25
Figura 5	Via de sinalização de Ras.....	26
Figura 6	Ativação da via RAS/RAF/MAPK, pelo ligante EGF ao receptor EGFR.....	27
Figura 7	Comparação da expressão de um <i>microarray</i> .....	35

## LISTA DE ILUSTRAÇÕES DO ARTIGO

Figure 1	Protein-protein interaction of differentially expressed genes upregulated .....	56
Figure 2	Protein-protein interaction of differentially expressed genes downregulated.....	57
Figure 3	The GO biological processes network upregulated .....	58
Figure 4	The GO biological processes network downregulated.....	59
Figure 5	Heatmap and hierarchical clustering of samples using gene expression values .....	61
Figure 6	Principal component analysis of healthy and tumor samples .....	62
Figure 7	Level of protein expression of 38 upregulated genes, of total 49 genes.....	63
Figure 8	Immunohistochemical staining of PAICS (A), SOX4 (B) and TOP2A (C).....	64
Figure 9	Level of protein expression of 133 downexpressed genes, of total 164 genes. ....	65

## LISTA DE TABELAS

Tabela 1	Lista de dados do câncer de pulmão selecionados pelo GEO/NCBI.....	36
----------	--	----

## LISTA DE TABELAS ARTIGO

Table 1	List of lung cancer datasets selected in GEO/NCBI .....	48
Table 2	Number of DEG up and downregulated in the six microarray studies used .....	51
Table 3	Frequency and Scores of the lung cancer DEG upregulated and disease in the samplings using the Malacards method.....	52
Table 4	Frequency and Scores of the lung cancer DEG downregulated and disease in the samplings using the Malacards method.....	53
Table 5	Pathway KEGG of Genes UP.....	54
Table 6	Pathway KEGG of Genes DOWN .....	54
Table 7	Hierarchical classification .....	61
Table 8	Binary classification .....	61

## LISTA DE ABREVIATURAS E SIGLAS

AKT	Serine/Threonine Kinase 1
APID	Agile Protein Interaction DataAnalyzer
DEG	Differentially Expressed Genes
ELK1	ETS Transcription Factor
ERBB1	ERBB Receptor Feedback Inhibitor 1
FDR	False discovery rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GS	Gene Signature
IL-6	Interleukin 6
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAPK	Mitogen-Activated Protein Kinase
MIG6	Mitogen-Inducible Gene 6 Protein
mTOR	Mechanistic Target Of Rapamycin Kinase
NSCLC	Non-small-cell lung cancer
PCA	Principal Component Analysis
PI3K	Phosphoinositide 3-kinase
PPI	Protein – Protein Interaction
RAF	Proto-Oncogene, Serine/Threonine Kinase
SCLC	Small-cell lung cancer
STAT	Signal Transducer And Activator Of Transcription
TKI	Tyrosine Kinase Inhibitor

## LISTA DE ABREVIATURAS E SIGLAS

AC	Adenocarcinoma
BASCs	Células-tronco bronquioalveolares
CCA	Células de Clara
CPCE	Carcinomas de pulmão de células escamosas
CPGC	Carcinomas de pulmão de grandes células
CPNPC	Câncer de pulmão de não pequenas células
CPPC	Câncer de pulmão de pequenas células
DAVID	Banco de Dados de Anotação, Visualização e Descoberta Integrada
EGF	Fatores de crescimento epidérmico
EGFR	Receptor do fator de crescimento epidérmico
GEF	Fatores de troca de nucleotídeos de guanina
HRas	Proto-Oncogene
INCA	Instituto Nacional de Câncer
KRAS	Proto-Oncogene
NRAS	Proto-Oncogene
OMS	Organização Mundial da Saúde
RNA	Ácido ribonucleico

## SUMÁRIO

INTRODUÇÃO	14
REVISÃO DA LITERATURA	17
JUSTIFICATIVA	33
OBJETIVOS	34
GERAL	34
ESPECÍFICO	35
MATERIAL E MÉTODOS	35
AMOSTRAS DE ESTUDO	35
ANÁLISE ESTATÍSTICA	36
KEGG E ANOTAÇÃO FUNCIONAL	36
REDE PPI E REDE DE PROCESSOS BIOLÓGICOS	37
CLUSTERIZAÇÃO DE AMOSTRAS (HEATMAP E PCA)	37
IMUNOHISTOQUÍMICA (PROTEIN ATLAS)	37
REFERÊNCIAS	38
CAPITULO 1	44
CONSIDERAÇÕES FINAIS	79

## INTRODUÇÃO

O câncer de pulmão é uma patologia heterogênea na origem celular e desenvolvimento. Os carcinomas pulmonares são geralmente classificados como câncer de pulmão de pequenas células (CPPC) e câncer de pulmão de não pequenas células (CPNPC) (WONG et al., 2017). O CPNPC é histopatologicamente e clinicamente distinto do CPPC, e é ainda subcategorizado como adenocarcinoma (AC, 40%), carcinomas de pulmão de células escamosas (CPCE, 25%), carcinomas de pulmão de grandes células (CPGC 10%), carcinoma adenoescamoso e outros subtipos mais raros dos quais os adenocarcinomas são os mais comuns (WONG et al., 2017; SHIAU; TSAO, 2017). O adenocarcinoma de pulmão passou por uma nova classificação de subtipos, e atualmente a Organização Mundial da Saúde (OMS) reconhece vários subtipos morfológicos de adenocarcinoma, são eles: papilar, micropapilar, acinar, sólido e lepidico (CHIRIEAC; ATTANOOS, 2018; GESTHALTER; BILLATOS; KATHURIA, 2017). A sintomatologia clínica do adenocarcinoma e os sinais físicos são identificados como sendo inúmeros, principalmente na doença já avançada (NANAVATI; REED, 2018). Podem incluir tosse persistente, hemoptise (expectoração de sangue), dor torácica, alterações vocais como rouquidão, piora da falta de ar, pneumonia recorrente ou bronquite (NANAVATI; REED, 2018). O adenocarcinoma pode ainda levar à doença metastática, síndromes paraneoplásicas, endocrinopatias, disfunção hematológica e características sistêmicas genéricas (CHIRIEAC; ATTANOOS, 2018).

O câncer de pulmão é a principal causa de morte por câncer em todo o mundo. O CPPC é responsável por volta de 15% dos casos e o CPNPC por aproximadamente 85%, com estimativa de 221.200 novos casos e 158.040 mortes antecipadas em 2015 nos Estados Unidos (CHIRIEAC; ATTANOOS, 2018). O tabagismo é responsável por cerca de 90% dos cânceres de pulmão e diminui a sobrevida (LANDI et al., 2008). O adenocarcinoma, é considerado subtipo predominante de câncer de pulmão, porque ocorre tanto em indivíduos sem história de tabagismo, quanto em fumantes (LANDI et al., 2008). Em muitos países, o adenocarcinoma superou o carcinoma escamoso como o subtipo histológico mais comum do câncer de pulmão e também é o subtipo histológico mais comum em mulheres, asiáticas e nunca fumantes (SELAMAT et al., 2012).

O carcinoma do pulmão provoca a morte de mais de 150.000 pessoas todos os anos nos Estados Unidos, superando assim a mortalidade combinada de câncer de mama e próstata (SIEGEL, 2017). Apesar dos estudos na compreensão de sua patogênese molecular e avanços terapêuticos, o câncer de pulmão é a causa mais comum de óbitos vinculada ao câncer em todo

o mundo (KRADIN et al, 2017). O tabagismo é considerado a principal causa de câncer de pulmão, mas as intensificações do câncer de pulmão em indivíduos não fumantes tem aumentado consideravelmente, o que sugere interações genéticas e outros fatores ambientais interferindo no desenvolvimento do câncer de pulmão (GESTHALTER; BILLATOS; KATHURIA, 2017). Em 2017 foi estimado 600.920 americanos com chance de vir a óbito por câncer, correspondendo a cerca de 1.650 mortes por dia (SIEGEL, 2017). As causas mais comuns de morte por câncer incluem mortes por câncer de pulmão, o que correspondem à 26% dos casos (SIEGEL, 2017). Apesar do alto número de mortes, as taxas de incidência do câncer de pulmão continuam a diminuir cerca de duas vezes mais rápido nos homens que nas mulheres (SIEGEL, 2017). As diferenças entre os sexos nas tendências do câncer de pulmão refletem diferenças históricas no uso do tabaco (SIEGEL, 2017). A ocorrência desse tipo de câncer segue com declínio nos homens, a partir da década de 1980 até os dias atuais, porém nas mulheres a diminuição ocorre a partir de meados dos anos 2000 (NANAVATI; REED, 2018).

Carcinógenos existentes na fumaça do cigarro são responsáveis por 85% dos cânceres de pulmão, e 15 a 25% dos casos encontram-se em “não fumantes” (MENDELSON et al.,2015). Esses indivíduos que nunca fumaram, são categorizados como aqueles que ao longo da vida tiveram exposição em menos de 100 cigarros (MENDELSON et al.,2015). Apesar de não fumarem, eles são apontados como os que tiveram exposição à alguns fatores ambientais como o fumo passivo de cigarro ou *cannabis*, poluição do ar interno por fumaça do uso doméstico de carvão ou gás de cozinha, poluição do ar externo por amianto, radônio, cromo, arsênico, cádmio, sílica e níquel, bem como doença pulmonar e fatores da dieta alimentar (GESTHALTER; BILLATOS; KATHURIA, 2017; CHIRIEAC; ATTANOOS, 2018).

Dentre os indivíduos que nunca fumaram, geralmente prevalecem nas mulheres a presença de câncer de pulmão (GESTHALTER; BILLATOS; KATHURIA, 2017). Os Estados Unidos, vem apontando que o câncer de pulmão segue com destaque nos óbitos por câncer, e também vem sendo a causa mais comum de câncer em todo o mundo (NANAVATI; REED, 2018). Em indivíduos do sexo feminino, 26% dos óbitos por câncer é por causa do câncer de pulmão (NANAVATI; REED, 2018). Enquanto em indivíduos do sexo masculino, por volta de 28% dos óbitos de câncer são causados por câncer de pulmão (NANAVATI; REED, 2018). Entre todos os óbitos por câncer, 19,4% foram óbitos por câncer de pulmão, além de ser o segundo câncer mais frequentemente diagnosticado em homens e mulheres (SHIAU; TSAO 2017). O índice de sobrevivência em 5 anos é baixo, mas melhorou de 16,4% em 1975 para 21,4% em 2010 (NANAVATI; REED, 2018). Pacientes que apresentam estágio clínico avançado

exibem taxas de sobrevivência de 7 a 9% para o estágio IIIB e 2 a 13% para o estágio IV em comparação com apresentações em estágio inicial (50 a 73% para o estágio IA e 43 a 58% para o estágio IB) (SHIAU; TSAO 2017). A chance de desenvolver câncer de pulmão ao longo da vida é de 1:13 para homens e 1:16 para mulheres (NANAVATI; REED, 2018).

No Brasil, há aproximadamente 27.000 novos casos de câncer de pulmão a cada ano (FIGUEREDO et al., 2015). Assim como no resto do mundo, no Brasil o câncer de pulmão é o segundo tipo mais comum de câncer no sexo masculino; com 17.330 novos casos estimados em 2016 (ARAÚJO et al., 2018). Além disso, o câncer de pulmão é o quarto tipo mais comum de câncer encontrado em mulheres; com 10.890 novos casos notificados no mesmo ano (ARAÚJO et al., 2018). O Instituto Nacional de Câncer (INCA) estimou um total de 596.000 novos casos de câncer em 2016 no Brasil, 28.220 (4,7%) dos quais foram malignidades primárias do pulmão (ARAÚJO et al., 2018). O adenocarcinoma é o principal tipo histológico de CPNPC presente no Brasil (COSTA et al., 2016). Além disso, destaca a prevalência progressiva de casos de CPNPC entre mulheres (COSTA et al., 2016).

Estudos mostraram que 12,7% dos homens e 8,0% das mulheres com 18 anos ou mais eram fumantes em 2016, essas proporções foram de 43,3% e 27% em 1989 (ARAÚJO et al., 2018). Os principais componentes das políticas brasileiras de combate ao tabaco incluem proibição de fumar em locais públicos, impostos mais altos para produtos de tabaco e rótulos de advertência sobre cigarros (ARAÚJO et al., 2018). Estima-se que, até 2020, a população com câncer de pulmão associada ao tabagismo no Brasil será de 83,3% entre os homens e 64,8% entre as mulheres (ARAÚJO et al., 2018).

O Brasil intensificou o programa antitabagismo através da restrição das propagandas na criação da lei 9.294/1996 (WONG et al., 2017). No Brasil, o número de mortes em mulheres devido ao câncer de pulmão está entre os que mais crescem no mundo, junto com Espanha e Chipre, mostrando que existe uma necessidade de campanhas dirigidas a esse público (WONG et al., 2017). Avaliação da mortalidade provocada pelo câncer de pulmão no Brasil mostra o aumento do número de mortes (ARAÚJO et al., 2018). O número de mortes em homens passou de 10,6 mortes/100.000 habitantes para 31,1 mortes/100.000 habitantes entre 1979 a 2004 (ARAÚJO et al., 2018). Entre as mulheres a incidência passou de 3,0 mortes/100.000 habitantes para 5,4 mortes/100.000 habitantes no mesmo período (ARAÚJO et al., 2018). O Brasil possui uma das maiores tributações sobre o preço do cigarro (76,3%), mas o preço final ainda é baixo comparado com outros países devido a produção apresentar baixo custo (SILVA et al., 2014). A legislação brasileira contra o tabagismo foi criada em 1996 (lei 9.294/1996) e modificada e complementada em 2000 (lei 10.167/2000) (ARAÚJO et al., 2018). A lei ainda regula sobre



propagandas na internet desde 2011 (lei 12.546/ 2011) (SILVA et al., 2014).

Para entender as mudanças moleculares envolvidas no adenocarcinoma de pulmão, utilizamos dados de adenocarcinomas de pulmão e comparamos com amostras de tecidos normais usando *microarray*<sup>1</sup>; com o objetivo de investigar seu padrão característico de expressão gênica e vias de sinalização. Analisamos seis conjuntos de dados de *microarray* e verificamos os genes comuns diferencialmente expressos entre esses dados. Os dados de expressão gênica mostraram 49 genes superexpressos e 164 subexpressos pelos seis estudos.

## REVISÃO DA LITERATURA

Um ponto crucial na história de um tumor é a transição de uma célula normal para um estado maligno (VICENTE-DUEÑAS et al., 2018). Durante o desenvolvimento embrionário, o processo de diferenciação celular confere comprometimento de uma célula a um destino específico e terminalmente diferenciado (ROY ANDHEBROK, 2015). Tradicionalmente, foi aceito que, uma vez que uma célula tenha concluído seu caminho de diferenciação em direção a um destino específico, esse estado é permanente e irreversível (GURDON; MELTON, 2008). Esta visão da maturidade celular como um estado imóvel foi desafiada, e atualmente é pertinente e aceito a ideia da plasticidade celular de células maduras, células totalmente diferenciadas de adultos podem ser geneticamente reprogramadas para induzir células-tronco pluripotentes (iPS) (GURDON; MELTON, 2008; TAKAHASHI; YAMANAKA, 2006).

O câncer surge de alterações genéticas e epigenéticas mudando o estado da célula normal para uma célula tumoral com replicação acelerada e que não desempenha mais as funções normais. Uma representação do câncer de pulmão com diferenciação das células normais que crescem rapidamente e impedem o funcionamento do órgão (figura 1).

---

<sup>1</sup> O desenvolvimento do chamado *chip* de DNA ou tecnologia de *microarray* permite investigar a expressão de centenas ou milhares de genes em uma dada amostra usando uma reação de hibridização.

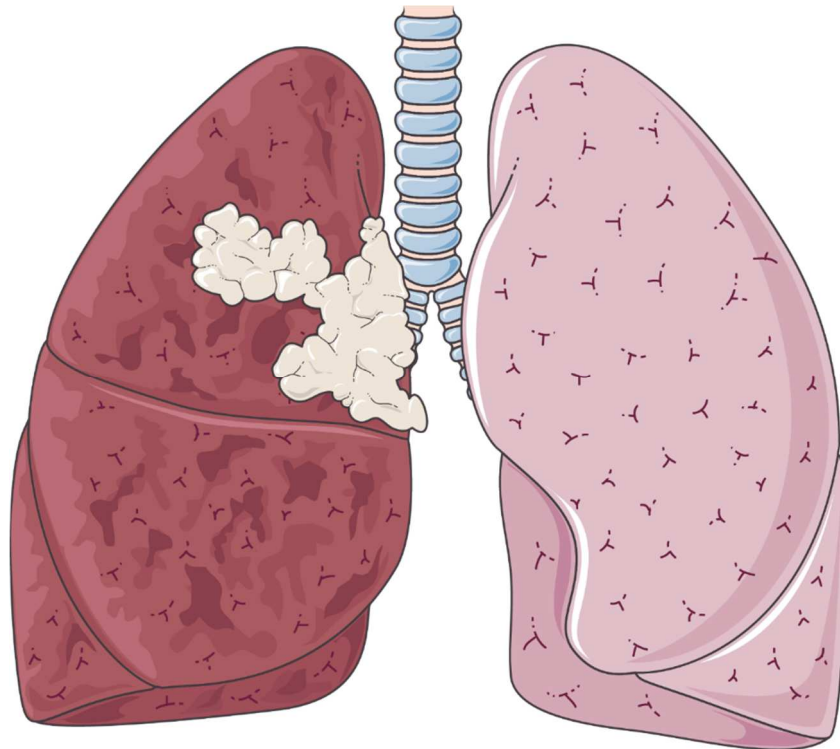


Figura 1. Desenvolvimento do câncer de pulmão. Danos genéticos alteram algumas células do pulmão, promovendo condições para o crescimento descontrolado, multiplicação e invasão no tecido pulmonar (células em branco representando crescimento e invasão de células cancerosas no lobo superior direito do pulmão).

O estado de diferenciação final de células também foi observado durante a progressão maligna (PURI et al., 2014). Células maduras com maior plasticidade têm a capacidade de adquirir algumas das características genotípicas e fenotípicas de um estado semelhante ao progenitor ou adotar um estado distinto diferenciado (PURI et al., 2014). No desenvolvimento do câncer, a mutação nos principais genes reguladores é um dos principais impulsionadores do aumento da plasticidade (PURI et al., 2014). A expressão oncogênica<sup>2</sup>, é necessária nos estágios iniciais do desenvolvimento do câncer e para a progressão (HANAHAN; WEINBERG, 2011). A injúria oncogênica "redefine" o epigenoma e/ou transcriptoma de uma célula normal (a célula de origem do câncer) e dá origem a um programa de diferenciação patológica, que finalmente leva ao desenvolvimento de tumores, novo modelo de câncer baseado em *priming*<sup>3</sup> epigenético de tumor (figura 2) (KRIZHANOVSKY; LOWE, 2009; PEREZ-CARO, et al., 2009). Após uma lesão oncogênica uma célula madura pode sofrer perda de identidade celular em seu caminho para a neoplasia e manter essa plasticidade anormal através dos estágios malignos (PURI et al., 2014).

<sup>2</sup> É a denominação dada aos genes relacionados com o surgimento de tumores.

<sup>3</sup> É a denominação dada a uma nova identidade de células tumorais durante o desenvolvimento do câncer. Que pode permanecer latente até ser desencadeada por estímulos endógenos ou ambientais.

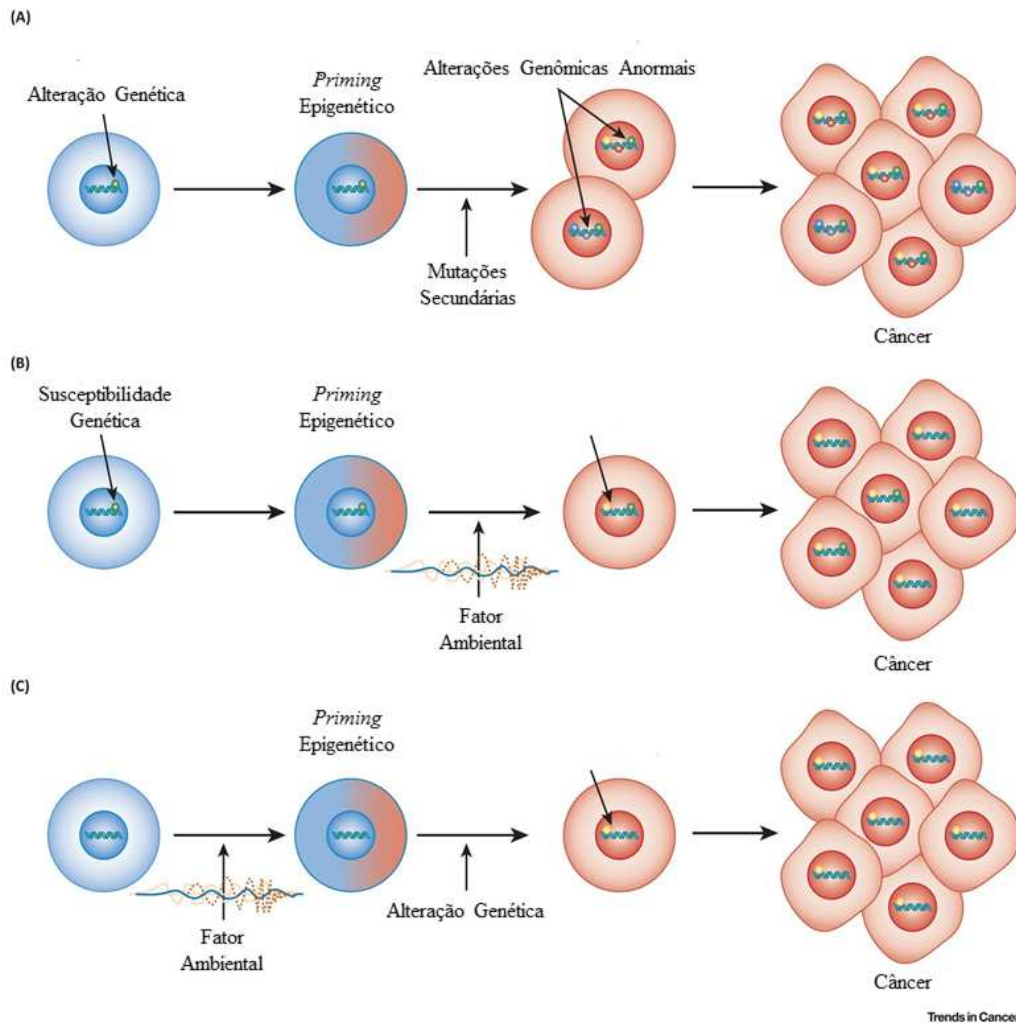


Figura 2. Modelo de câncer baseado em *priming* epigenético de tumor. (A) Interações gene-ambiente podem causar eventos secundários aleatórios e desencadear uma progressão tumoral. (B) No entanto, um determinado *priming* epigenético pode ser particularmente sensível a um determinado fator ambiental (interação gene-ambiente), de modo que a exposição a esse fator favorecerá o surgimento de uma segunda lesão que, por sua vez, impulsionará o desenvolvimento do câncer. (C) Em outros casos, a exposição a um determinado fator ambiental causa reprogramação epigenética que aumenta a suscetibilidade ao câncer (VICENTE-DUEÑAS et al., 2018).

O surgimento do câncer ocorre pelo acúmulo de alterações genéticas somáticas que resultam em proliferação celular anormal e bloqueiam a morte celular (BEROUKHIM, et al., 2010; TRAN; KWON, 2018). Essas alterações podem ser do tipo substituições de base única, translocações, alterações no número de cópias e posteriormente alterações citogenéticas (BEROUKHIM, et al., 2010). Novas ferramentas tecnológicas foram criadas para a caracterização genômica (BEROUKHIM, et al., 2010). Essas ferramentas têm possibilitado a identificação das alterações em amostras de câncer humano possibilitando melhor diagnóstico,

prognóstico e tratamento (BEROUKHIM, et al., 2010). A identificação de tais alterações no genoma pode fornecer dados relevantes sobre os defeitos celulares que permitem o surgimento do câncer e sugerir potenciais estratégias terapêuticas (BEROUKHIM, et al., 2010). A análise genômica de amostras de tumores, por meio de estudos citogenéticos e, mais recentemente, por perfis baseados em *microarray*, identificou alterações recorrentes interligadas a tipos particulares de câncer (BEROUKHIM, et al., 2010). O diagnóstico precoce e a terapia personalizada se baseiam em *insights* de vias de sinalização molecular relevantes, bem como genes relacionados ao câncer (TRAN; KWON, 2018).

Para auxiliar o diagnóstico imediato no estágio inicial do desenvolvimento do câncer de pulmão, é importante a descoberta de biomarcadores da patologia e um estudo profundo nas alterações do ambiente pulmonar, já que está exposto as mudanças do ambiente externo (KOTTON; MORRISEY, 2014). O tecido pulmonar apresenta uma estrutura altamente especializada, revestido com um epitélio complexo, pseudo-estratificado, predominando células ciliadas, porém também apresenta células secretoras, populações raras, com a função de célula progenitora, proteção e reparo da via aérea de prováveis lesões (HOGAN et al. 2014). Já os alvéolos apresentam revestimento por tipos de células diferentes, conhecidas como células alveolares tipo I escamosas e células epiteliais alveolares tipo II cuboides (HANNA; ONAITIS; 2013). São responsáveis respectivamente pela interação capilar e por secreção de lipídios e proteínas que reduzem a tensão superficial durante a ventilação (HANNA; ONAITIS; 2013). As células alveolares também podem atuar como precursores de células do tipo I durante a reparação alveolar (HANNA; ONAITIS; 2013).

O ponto que marca o início celular do câncer de pulmão mantém-se em grande parte inexplorado (GIANGRECO; GROOT; JANES, 2007). Considera que diferentes subtipos histopatológicos de tumores surgem de células de origem distintas, e localização em microambientes definidos (GIANGRECO; GROOT; JANES, 2007). Sobretudo, devido ao padrão de distribuição proximal-distal do pulmão, sabe-se que em localizações mais distais é onde está presente o adenocarcinoma (GIANGRECO; GROOT; JANES, 2007). Possíveis células candidatas para origem dos tumores, incluem as células-tronco teciduais porque sua longa vida *in vivo* permitiria acumulação de mutações genéticas conduzindo a tumorigênese (GIANGRECO; GROOT; JANES, 2007). Outras células também estão inclusas como candidatas para a formação do câncer, como as células diferenciadas e progenitores (SMALLEY; ASHWORTH, 2003). Isso seria possível uma vez que suas taxas de rotatividade, bem como combinações de modificações genéticas a conferisse capacidade de autorrenovação (GIANGRECO; GROOT; JANES, 2007; SMALLEY; ASHWORTH, 2003; SUTHERLAND

et al., 2011).

Anatomicamente nas vias aéreas mais distais do pulmão, existem as células Clara que revestem os bronquíolos (RAWLINS et al., 2009). Essas demonstraram a capacidade de gerar células ciliadas e capacidade de autorrenovar em resposta à lesão epitelial (RAWLINS et al., 2009). É possível observar nesse compartimento mais distal do pulmão, os ductos alveolares ou espaços aéreos de troca gasosa, que são compostos por alvéolos tipo I e células alveolares do tipo II, sendo estas últimas consideradas as principais células tronco, progenitoras do epitélio alveolar com base na sua capacidade tanto de originar células alveolares tipo I como de se autorrenovar (RAWLINS et al., 2009; GIANGRECO; REYNOLDS; STRIPP, 2002).

Embora numerosas mutações tenham sido descritas, o estado de mutação permanece desconhecido em mais de 50% dos casos. Atualmente, podemos identificar alvos terapêuticos em apenas 20% dos cânceres de pulmão (CHIRIEAC; ATTANOOS, 2018). Genes associados ao câncer de pulmão e o processo de formação do câncer são elementos que têm demonstrado ser complicados (GESTHALTER; BILLATOS; KATHURIA, 2017). Ocorrem diversas mutações nos genes do câncer de pulmão durante a carcinogênese (GESTHALTER; BILLATOS; KATHURIA, 2017). No adenocarcinoma de pulmão, esses genes alvos inclui genes como: EGFR (GESTHALTER; BILLATOS; KATHURIA, 2017).

Os primeiros relatos de uma mutação do receptor do fator de crescimento epidérmico (EGFR) foram de um subconjunto de pacientes com adenocarcinoma do pulmão (KRADIN et al, 2017). Mutações de EGFR aparecerem em 10 e 15% dos pacientes com câncer de pulmão (KRADIN et al, 2017). As mutações de EGFR são mais comumente associadas aos não fumantes assim como, em pacientes com CPNPC com histologia de Adenocarcinoma (SHIAU; TSAO 2017). A prevalência da mutação do EGFR em estudos populacionais da Ásia Oriental alterna entre 30% e 50% em comparação com populações caucasianas que apontam uma prevalência de 10 a 20% (LINDEMAN et al., 2013; ZHOU et al., 2011). A sinalização de EGFR anormal leva à ativação de componentes de sinalização na cascata celular, como AKT, que está associada ao aumento da proliferação celular e ao desenvolvimento de câncer (LINDEMAN et al., 2013; ZHOU et al., 2011).

Células expressando EGFRs mutantes, tem hiperfosforilação acontecendo e um dos candidatos da hiperfosforilação é a proteína codificada pelo gene ERFF1 (também conhecido como MIG6), um gene de resposta imediata que é induzido por fatores de crescimento epidérmico (EGF) e estímulo de estresse (MAITY et al., 2015). Vários estudos relataram que o

RNA<sup>4</sup>, de ERRFI1 é aumentado em linhagens celulares de adenocarcinoma de pulmão com mutação em EGFR (MAITY et al., 2015). Estas observações levantam as questões sobre se ERRFI1 é um supressor de tumor para o adenocarcinoma de pulmão dirigido por EGFR mutante e, em caso afirmativo, como o EGFR mutante induz adenocarcinomas de pulmão na presença de ERRFI1 (MAITY et al., 2015).

Sabe-se que o ERRFI1 inibe a atividade da EGFR e promove a degradação do WT EGFR (tipo selvagem), por isso é considerado um supressor de tumores com sinalização ativa do WT EGFR (MAITY et al., 2015). O papel do ERRFI1 na regulação de EGFRs mutantes do câncer de pulmão apresenta fosforilação constitutiva de ERRFI1 em tirosina na posição Y394/Y395 em células de adenocarcinoma de pulmão (MAITY et al., 2015). A fosforilação nestes locais é inibida por erlotinib em células sensíveis a Tyrosine Kinase Inhibitor (TKI), mas não resistentes, sugerindo que os EGFR mutantes fosforilam diretamente e possivelmente regulam ERRFI1 (MAITY et al., 2015). Examinando ainda as consequências funcionais da fosforilação observa-se que aumenta a ligação de proteínas ERRFI1 a EGFR mutantes, mas, ao contrário do EGFR do tipo selvagem, a interação aumentada não direciona o EGFR mutante para a via de degradação (MAITY et al., 2015). Os achados de fosforilação de ERRFI1, Y394/Y395, em linhagem celulares de adenocarcinoma de pulmão e o consequente aumento da ligação de ERRFI1 com EGFRs mutantes leva à estabilidade de EGFR mutante. O ERRFI1 fornece fortes evidências de que os EGFR mutantes podem contornar parcialmente a inibição por ERRFI1 em células de adenocarcinoma de pulmão através da fosforilação de tirosina de ERRFI1 em resíduos chave. No entanto, a atividade inibidora residual de ERRFI1 ainda é supressora de tumor em modelos de tumor de pulmão mutados dirigidos por EGFR, uma vez que a deficiência de ERRFI1 reduz a sobrevivência de camundongos devido à tumorigênese acelerada. O modelo de atividade inibitória de ERRFI1 contra WT (selvagens) e mutante de EGFR é mostrado na figura 3 (MAITY et al., 2015).

---

<sup>4</sup>Ácido ribonucleico ou RNA, é um tipo de ácido nucleico, uma molécula polimérica linear formada por unidades menores chamadas nucleotídeos. Intervém em várias funções biológicas importantes como a codificação genética, e a descodificação durante a tradução de proteínas, regulação e expressão dos genes.

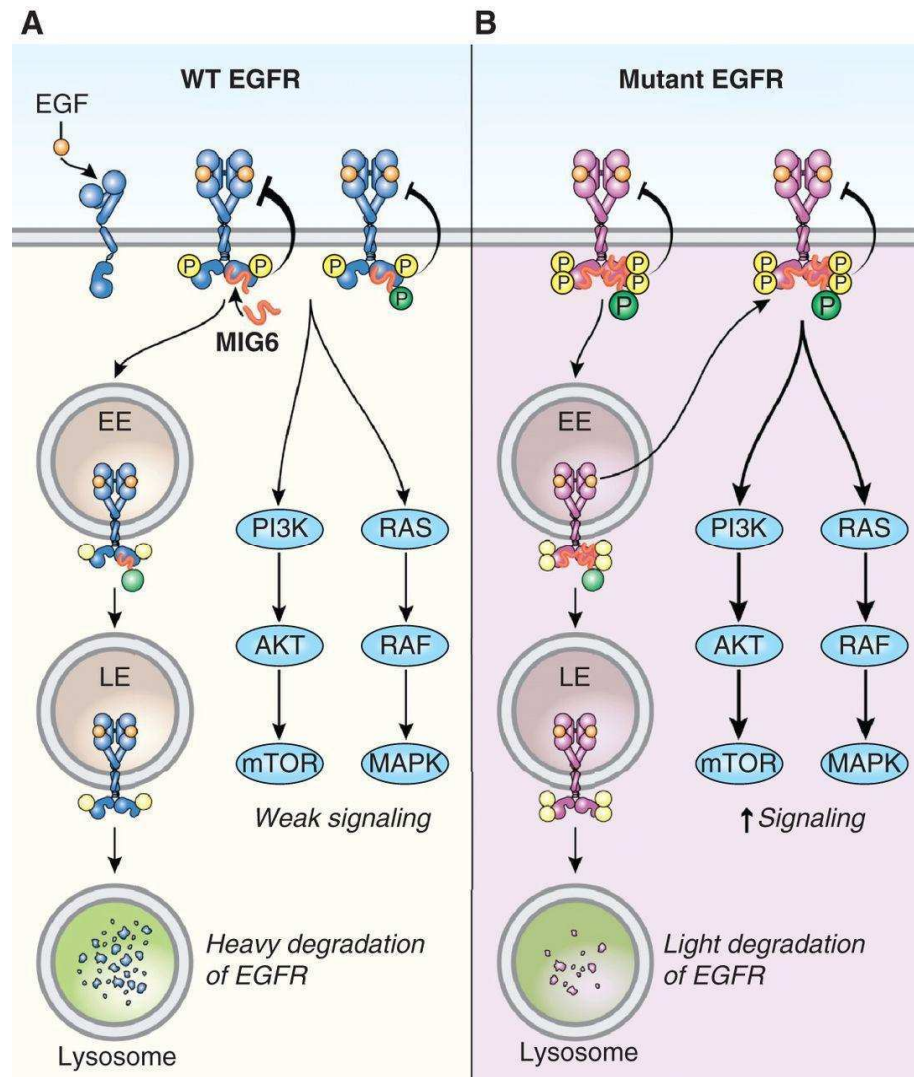


Figura 3. Representação da regulação do EGFR WT (selvagem) e EGFR mutante por ERRFI1. (A) A proteína do gene ERRFI1 (MIG6) se liga, inibe a atividade quinase do WT EGFR ativo e promove o transporte WT EGFR para a degradação em endossomos iniciais (EE) e endossomos tardios (LE) até chegar ao Lisossomo. Ao passo que esse processo acontece, um mecanismo de feedback pode está acontecendo. Ou seja, para reverter o papel inibitório do gene ERRFI1 no WT EGFR, o WT EGFR ativo, fosforila os resíduos Y394/Y395 da proteína traduzida pelo gene ERRFI1, para aumentar a ligação a ERRFI1 e diminuir a inibição da quinase de EGFR de WT. (B) EGFR mutante e ativo, resulta em fosforilação aumentada nos resíduos Y394/Y395 da proteína do gene ERRFI1. Esse mecanismo aumenta a interação da proteína com o EGFR mutante, além disso, diminui a inibição da quinase de EGFR mutante, e também promove degradação do EGFR mutante. A inibição é atenuada, mas incompleta, e suficiente para o ERRFI1 (MIG6) funcionar como um potente supressor de tumores para a iniciação e progressão da tumorigênese (MAITY et al., 2015).

Entre os proto-oncogenes<sup>5</sup>, a família do fator de crescimento epidérmico (EGF) desempenha um papel importante no crescimento do tumor local (MEERT et al., 2003). Este fenómeno requer proteínas reguladoras do crescimento, tais como o EGFR (MEERT et al., 2003). O EGF, pertence a um grupo de moléculas que estimulam quatro receptores de tirosina quinases (RTKs) da família ERBB e juntos eles estimulam uma complexa rede de sinalização que é essencial para a organogênese e para a determinação de várias linhagens celulares (YARDEN; PINES, 2012). Os membros da família de receptores ERBB são expressos em muitas células das linhagens epiteliais, mesenquimais e neuronais, nas quais desempenham diversos papéis no desenvolvimento, proliferação, diferenciação, migração celular, metabolismo e sobrevivência (TEBBUTT; PEDERSEN; JOHNS, 2013; AVRAHAM; YARDEN, et al., 2011). A ativação inapropriada de EGFR e ERBB2 no câncer pode ocorrer por meio da superexpressão (geralmente devido à duplicação do gene), mutações pontuais, deleções parciais e estimulação autócrina do ligante-receptor, o que pode levar à ativação independente de ligante aos receptores, bem como ao aumento da ativação após o acoplamento com o ligante (TEBBUTT; PEDERSEN; JOHNS, 2013).

Em tumores, EGFR estimula muitas proteínas e vias de sinalização intracelular, como os fatores de transcrição MAPK, PI3K/Akt e mTOR, e STAT (AVRAHAM; YARDEN, et al., 2011; HYNES; MACDONALD, 2009). Considerando o câncer de pulmão, as vias PI3K/Akt e STAT parecem ser hiperativas no contexto de EGFRs mutantes (YARDEN; PINES, 2012). A via PI3K/Akt é controlado diretamente pelo EGFR (YARDEN; PINES, 2012). No entanto, dados recentes sugerem que a ativação de STAT é indireta e mediada pela elevação da produção de IL-6 em células tumorais (HYNES; MACDONALD, 2009).

Os TKIs de EGFR de primeira geração, como o erlotinib e o gefitinib, são inibidores reversíveis que competem pela ligação do domínio quinase com o ATP endógeno, impedindo assim a sua atividade de fosforilação da tirosina e bloqueando a sinalização na cascata (TEBBUTT; PEDERSEN; JOHNS, 2013). Em pacientes com CPNPC tratados com gefitinibe, sugere que os tumores primários com alta atividade de PI3K/Akt eram mais responsivos ao bloqueio de EGFR (HYNES; MACDONALD, 2009).

---

<sup>5</sup>É um gene normal que se torna um oncogene devido a uma mutação ou ao aumento de expressão gênica.



As proteínas RAS são proto-oncogenes frequentemente mutadas nos cânceres humanos (PRIOR, LEWIS, MATTOS, 2012; QUINLAN; SETTLEMAN, 2009). Elas são codificadas por 3 genes: HRAS, KRAS e NRAS. Estas proteínas são GTPases que funcionam como interruptores moleculares nas vias reguladoras que são responsáveis pela proliferação e sobrevivência celular (QUINLAN; SETTLEMAN, 2009). A sinalização através da RAS/RAF/MAPK regula uma variedade de funções celulares que são responsáveis para a tumorigênese. A função desta via é a transdução de sinais do meio extracelular para o núcleo celular, onde genes específicos são ativados para o crescimento, divisão e diferenciação celular. A via RAS/RAF/MAPK também está envolvida na regulação do ciclo celular, cicatrização de feridas e reparação de tecidos, sinalização de integrinas e migração celular (MOLINA; ADJEI, 2006). As proteínas Ras são rigorosamente reguladas por fatores de troca de nucleotídeos de guanina (GEF) que promovem a dissociação do GDP e a ligação de GTP, e proteínas ativadoras da GTPase (GAP) que estimulam a atividade GTPase intrínseca de RAS para desligar a sinalização (QUINLAN; SETTLEMAN, 2009). Todas as isoformas de RAS compartilham identidade de sequência (figura 4) em todas as regiões que são responsáveis pela ligação GDP/GTP (PRIOR, LEWIS, MATTOS, 2012).

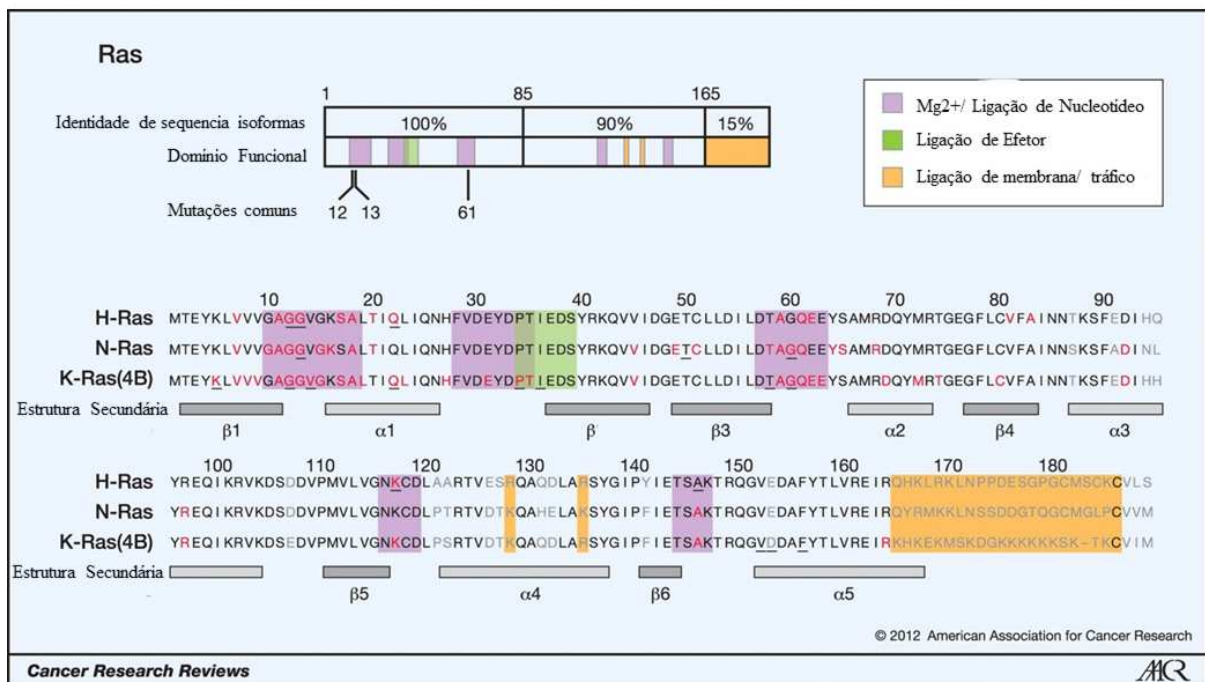


Figura 4. Presença de mutações nas isoformas de RAS. São observadas as mutações pontuais nas três isoformas de RAS e as principais mutações estão presentes em regiões idênticas nas isoformas. As mutações estão presentes nos códons 12, 13 e 61, são 44 mutações pontuais. Representadas em vermelho, são os nucleotídeos mutados, em cinza são os nucleotídeos mutados que variam entre as isoformas, e os sublinhados são os nucleotídeos mutados em distúrbios do desenvolvimento (PRIOR, LEWIS, MATTOS, 2012).

As proteínas RAS estão estrategicamente posicionadas na superfície interna da membrana plasmática. Os EGFR, como o fator de crescimento epidérmico prototípico, atuam principalmente para promover essa alteração conformacional na RAS, a fim de iniciar uma cascata de sinalização ao citoplasma e, eventualmente, ao núcleo (GORDON; BAKSH, 2011). A ativação da via começa quando um sinal se liga ao EGFR, pelo EGF (figura 5), que estimula a autofosforilação dos domínios catalíticos citoplasmáticos do EGFR e o recrutamento de proteínas adaptadoras, Grb2 e Sos (MOLINA; ADJEI, 2006; GORDON; BAKSH, 2011). A sinalização estimulada pela ligação do EGF ao receptor EGFR, recruta por meio da fosforilação dos domínios, fatores de troca de nucleotídeos de guanina (GEFs), são esses os GEFs: SOS-1 ou CDC25. O GEF torna-se capaz de interagir com as proteínas RAS na membrana celular para promover uma mudança conformacional e a troca do GDP pelo GTP (MOLINA; ADJEI, 2006). Após a ativação de RAS, a RAF é recrutada para a membrana celular. A ativação da RAF estimula uma cascata de sinalização pela fosforilação da MAPK que sucessivamente fosforila e ativa as proteínas MAPK3 e MAPK1. A ativação de MAPK3 e MAPK1 fosforilam e ativam uma variedade de fatores de transcrição nuclear e quinases, incluindo ELK-1 (MOLINA; ADJEI, 2006).

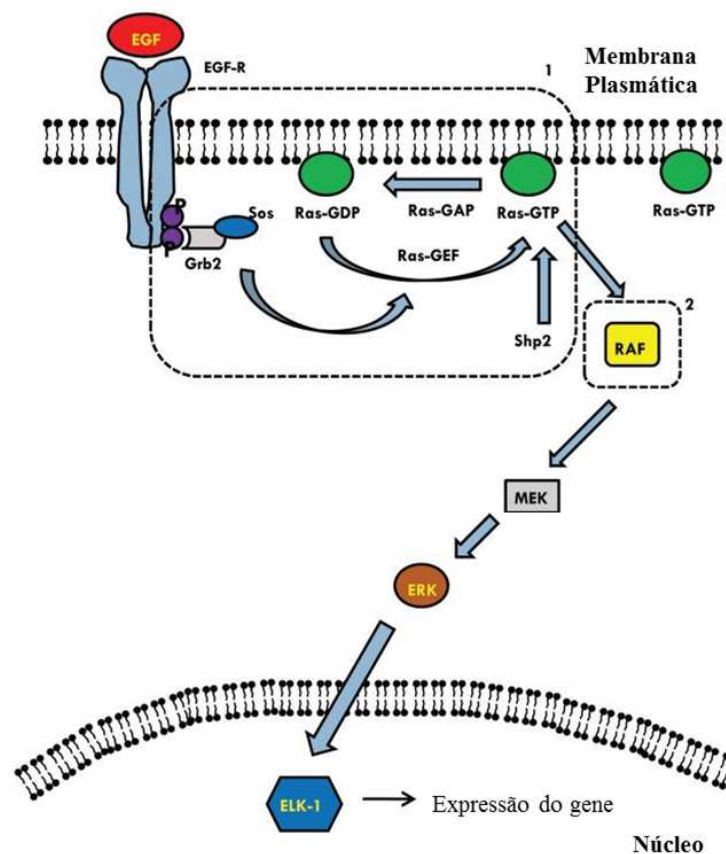
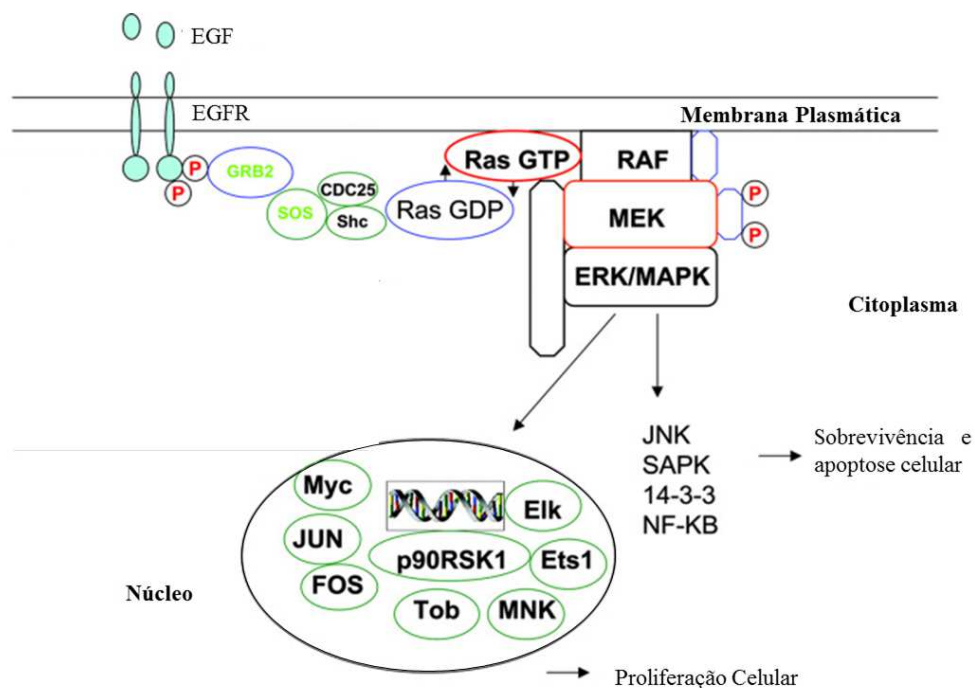


Figura 5. Via de sinalização de RAS (GORDON; BAKSH, 2011). (1) Ao EGF ligar ao EGFR uma cascata de sinalização e mecanismos acontecem, como a autofosforilação dos domínios citoplasmáticos de EGFR, recrutamento de proteínas adaptadoras para a membrana plasmática (Sos e Grb2), mudança conformacional de RAS através da troca de GDP para GTP. (2) Recrutamento de RAF, e uma cascata de sinalização por meio da fosforilação acontece em MEK (MAPKK), ERK (MAPK1 e 3) para enfim sinalizar para o fator de transcrição ELK-1 ativando a transcrição e conseqüentemente a proliferação celular.

Uma vez ativado pela fosforilação, o ELK-1 adquire a competência de ligação ao DNA<sup>6</sup> para ativar a transcrição gênica e impulsionar a proliferação. ELK-1 é apenas um dos muitos fatores de transcrição modulados pela sinalização de RAS (figura 6) (GORDON; BAKSH, 2011). A MAPK, que em mamíferos também é chamada de MEK, é uma serina/treonina quinase ativada em resposta a múltiplos sinais, incluindo fatores de crescimento e citocinas, para promover a sobrevivência e apoptose celular através de vários mediadores como JNK, SAPK, 14-3-3 e NF-KB (MOLINA; ADJEI, 2006). RAF e MAPK não são os únicos alvos a jusante de KRAS, outros efetores a jusante de KRAS incluem a via de sobrevivência de células PI3K, as pequenas proteínas de ligação a GTP Rac e Rho e a via quinase de proteína ativada por estresse (MOLINA; ADJEI, 2006).



<sup>6</sup>Molécula de ácido desoxirribonucleico, faz parte do dogma central da biologia molecular e é onde está toda a informação genética.

Figura 6. Ativação da via RAS/RAF/MAPK. Quando o ligante EGF se acopla ao receptor EGFR acontece uma mudança de posição dos domínios catalíticos citoplasmáticos do EGFR, permitindo a autofosforilação e atividade quinase nesse local. Desta forma as proteínas adaptadoras GRB2 consegue se ligar aos domínios do EGFR, e recruta as proteínas SOS e CDC25. As proteínas SOS e CDC25 são capazes de interagir com a proteína RAS a ponto de promover uma mudança conformacional, com troca de GDP-GTP. Após o processo de ativação de RAS, o próximo passo é a ativação de RAF, que sinalizará toda uma cascata de fosforilação MAPK, para o recrutamento de fatores de crescimento, citocinas (JNK, SAPK, 14-3-3 E NF-KB) da sobrevivência e apoptose celular e enfim sinalizar proliferação celular (MYC, JUN, FOS, etc.). (MOLINA; ADJEI, 2006).

No adenocarcinoma de pulmão a principal célula responsável são as células epiteliais do tipo II alveolar, controlados por KRAS (HOGAN et al. 2014). Sendo assim, adenocarcinomas de pulmão em camundongos induzidos por mutação em *KRas*, revelou a presença de uma população rara de células duplamente positivas (DPCs)<sup>7</sup> para células alveolar tipo II (AT2) e para células de Clara (KIM et al., 2005). Além disso, são quiescentes na homeostase pulmonar normal, mas se autorrenovam e dão origem a células bronquiolares e alveolares após lesão. Sugerindo características semelhantes às células progenitoras e apoiando sua designação como BASCs (REYA et al., 2001).

Quando uma mutação em *KRas* é induzida, observa-se células-tronco BASCs proliferando (KIM et al., 2005; ZHANG et al., 2008; VENTURA et al., 2007). Em camundongos com mutação *KRas*, que sofreram lesões com tumor, apresentam expansão de DPCs continuamente, concordando com a progressão tumoral nesses camundongos (VENTURA et al., 2007). Oncogenes humanos como o gene KRAS sofre mutação em  $\pm 30\%$  dos casos, enquanto que os genes MYC e EGFR são superexpressos em 2,5–10% e 6% dos casos, respectivamente (SEKIDO; FONG; MINNA, 1998). Oncogenes novos estão sendo identificados como em BRAF, presente em cerca de 2% dos pacientes com adenocarcinomas, em certas particularidades de tumores que não têm mutações KRAS (NAOKI et al., 2002). Alterações oncogênicas de fatores de transcrição também estão sendo alvos no desenvolvimento do câncer pulmonar (SODA et al., 2007).

O fator de transcrição NKX2-1 é um marcador clássico para adenocarcinomas (KIM et al., 2015). Além disso, o gene MYC está entre os mais frequentemente superexpressos em

---

<sup>7</sup>As DPCs têm capacidade de expressar o marcador de células de Clara (CCA) e também marcadores pró-surfactante apoproteína-C (SP-C), renomeadas como células-tronco bronquioalveolares (BASCs)

câncer de pulmão (TAI et al. 2015). Esse gene codifica um fator de transcrição que regula uma grande variedade de genes envolvidos no controle do crescimento celular, proliferação, e morte celular apoptótica.

Estudos realizados mostram a importância da utilização da técnica de *microarray* para identificar, o perfil de expressão (KABBOUT et al., 2014; LIST et al., 2014) juntamente com a análise das vias funcionais, no qual está ligada a várias vias canônicas associadas ao câncer de pulmão. Então a abordagem que está sendo muito utilizada é uma forma de avaliar a atividade da via de sinalização e para isso utiliza o padrão de expressão gênica resultante da ativação *in vitro* da via para prever sua atividade em amostras de pacientes (PAULING et al., 2014). Esta abordagem tem sido bem sucedida na previsão do estado do caminho nas linhagens celulares, bem como nos tumores com ativação de via conhecida e lesões moleculares. Um dos pontos fortes do uso de assinaturas de expressão gênica para indicar a atividade de sinalização em amostras clínicas é que eles permitem a medição de vias múltiplas em um único experimento de *microarray*. A abordagem computacional para identificar as vias de sinalização que conduzem a oncogênese do câncer de pulmão pode identificar terapias que podem prevenir o desenvolvimento de câncer em populações de alto risco (GUSTAFSON et al., 2010). O tratamento do câncer de pulmão está, portanto, se movendo rapidamente para uma era de medicina personalizada, onde as características moleculares do tumor de um paciente indicarão as modalidades de tratamento adequadas (KIM et al., 2015).

Atualmente, entendemos a função de apenas um número limitado de genes humanos. Para enfrentar esse desafio, novas ferramentas de alto rendimento foram desenvolvidas (RUSSO; ZEGAR; GIORDANO, 2003). O ensaio de *microarray* é uma poderosa tecnologia molecular que permite o estudo simultâneo da expressão de milhares de genes ou seus produtos de RNA (RUSSO; ZEGAR; GIORDANO, 2003). As tecnologias de *microarray* estão agora disponíveis para medir a expressão gênica, o número de cópias de DNA, a metilação, o estado da cromatina, a ligação às proteínas, os SNPs e outros aspectos da fisiologia dos genes (REIMERS, 2010; KOSTINA; SINYAKOV; RYABININ, 2018). É um método que utiliza *microchips* contendo elementos curtos de DNA (conhecidos como sondas) ordenados em *microspots*, cada *spot* contendo uma única espécie de um ácido nucléico e representando os genes de interesse (GIBSON, 2003). Essa tecnologia é baseada na hibridação entre os alvos livres derivados de uma amostra biológica e uma matriz de muitas sondas de DNA que são imobilizadas no *microchip* (GIBSON, 2003; RUSSO; ZEGAR; GIORDANO, 2003). Os alvos são produzidos por transcrição reversa e o sinal de hibridização produzido em

cada sonda é a expressão do mRNA<sup>8</sup>, do gene correspondente na amostra no momento do estudo (STEARNS; MARTINSKY; SCHENA, 2003). Os sinais são detectados, quantificados, normalizados com programa e refletem o perfil de expressão gênica para cada amostra biológica (STEARNS; MARTINSKY; SCHENA, 2003). Muitos milhares ou dezenas de milhares de pontos distintos podem ser impressos em um silício ou lâmina de vidro ou uma base de nylon no estado sólido. Existem basicamente duas variantes de *microarrays*: *microarrays* de cDNA e *microarrays* de oligonucleotídeos (STEARNS; MARTINSKY; SCHENA, 2003; SLONIM; YANAI, 2009).

Em *microarrays* de cDNA, moléculas de DNA relativamente longas são imobilizadas em uma superfície sólida. Este tipo de *microarray* é usado principalmente para estudos de triagem e expressão em larga escala. O *microarray* de oligonucleotídeos é fabricado *in situ* por síntese química dirigida por luz ou por síntese convencional seguida de imobilização numa matriz de vidro, é usado para detecção de mutações, mapeamento de genes e estudos de expressão (RUSSO; ZEGAR; GIORDANO, 2003). Duas plataformas principais são conhecidas no meio científico Illumina (San Diego, CA) e da Affymetrix (Santa Clara, CA) (DISTEFANO; TAVERNA, 2012). Essas duas tecnologias concorrentes, oferecem diferentes abordagens para medir a variação do SNP. Por exemplo, a plataforma Affymetrix imprime sequências de DNA curtas como um ponto no chip que reconhece um alelo de SNP específico. Os alelos (nucleotídeos) são detectados por hibridação diferencial do DNA da amostra. A Illumina, por outro lado, usa uma tecnologia baseada em sequências de DNA ligeiramente mais longas para detectar alelos. Os chips da Illumina são mais caros, mas fornecem melhor especificidade (BUSH; MOORE, 2012).

Os *microarrays* tornaram-se importantes porque são mais fáceis de usar, não requerem sequenciamento de DNA em larga escala e permitem a quantificação paralela de milhares de genes de múltiplas amostras (RUSSO; ZEGAR; GIORDANO, 2003). As assinaturas de expressão gênica podem ser usadas para prever a resposta clínica tanto para terapias convencionais quanto terapias direcionadas (RUSSO; ZEGAR; GIORDANO, 2003). O perfil da expressão gênica de cânceres é representado e abrangente devido a existência da tecnologia de *microarray* (QIU; GENTLES; PLEVRITIS, 2011; WIMMER et al., 2018). Atualmente, a tecnologia de *microarrays* está sendo usada para derivar perfis de expressão gênica que

---

<sup>8</sup> Molécula que faz parte do dogma central da biologia molecular, sintetizada a partir de um molde de DNA pelo mecanismo de transcrição, e necessária para a formação de uma proteína pelo mecanismo da tradução.

predizem a origem precisa do tecido de um tumor em particular (SLONIM; YANAI, 2009). É possível fazer o diagnóstico de um determinado tipo de câncer e subconjuntos, sem examinar a histologia, agora é possível prever o resultado clínico com base nos padrões de expressão gênica (SLONIM; YANAI, 2009; RUSSO; ZEGAR; GIORDANO, 2003).

Embora a tecnologia de *microarray* represente uma oportunidade maravilhosa para pesquisa clínica, questões importantes devem ser consideradas (RUSSO; ZEGAR; GIORDANO, 2003). Estudos relacionados a *microarray* são limitados pelo acesso e pelo custo. A qualidade e quantidade de RNA permanece um grande desafio nos experimentos de *microarray*. A quantidade de tecido obtido e a complexidade da própria amostra de tecido limitam a qualidade e a quantidade de RNA que pode ser isolado. Devido às inúmeras etapas propensas a erros nos experimentos de *microarray*, os experimentos precisam ser replicados para eliminar fontes de erro. A maioria das amostras de tecidos humanos é uma mistura de diferentes tipos de células. Portanto, mudanças nos padrões de expressão gênica, ao comparar duas amostras diferentes de biópsia tecidual, são uma manifestação de todos os tipos de células presentes nessa amostra. Esse problema pode tornar a análise imprecisa (RUSSO; ZEGAR; GIORDANO, 2003). A análise das vias tornou-se a primeira escolha para obter *insights* sobre a biologia de genes e proteínas diferencialmente expressas, pois reduz a complexidade e aumenta o poder explicativo (KHATRI; SIROTA; BUTTE, 2012). A análise de enriquecimento é o processo de usar uma coleção de anotações genéticas para determinar quais termos são "enriquecidos" ou sobre-representados em um conjunto de genes, que são frequentemente produzidos a partir de experimentos em escala genômica (CLARKE, et al., 2013). As redes de genes são comumente interpretadas como codificação de informações funcionais em suas conexões. Os genes que estão associados ou interagindo têm maior probabilidade de compartilhar funções (GILLIS; PAVLIDIS, 2012). A informação de tais redes está agora incorporada em combinação com codificações da função de genes exemplificada pela Gene Ontology (GO)<sup>9</sup>, (GILLIS; PAVLIDIS, 2012). O GO é usado predominantemente para analisar dados de alto rendimento, como resultados de *microarrays* de expressão gênica. Uma análise típica começa identificando uma lista de genes diferencialmente expressos. Então, para obter *insights* sobre o significado biológico das alterações nos níveis de expressão gênica, os pesquisadores usam métodos de análise de enriquecimento GO para determinar se os termos

---

<sup>9</sup> A Ontologia Genética (GO) é a ontologia mais amplamente utilizada para especificar a localização celular, a função molecular e a participação de processos biológicos de genes humanos e de organismos modelo.

GO sobre processos biológicos específicos, funções moleculares ou componentes celulares estão superexpressos ou subexpressos dentro do conjunto de genes de interesse (KHATRI; SIROTA; BUTTE, 2012). Os genes com função relacionada tendem a ser parceiros de interação de proteínas ou compartilham características como padrões de expressão (GILLIS; PAVLIDIS, 2012).



## JUSTIFICATIVA

Recentes avanços nas tecnologias de caracterização genômica têm possibilitado esforços cada vez mais sistemáticos para caracterizar alterações em amostras de câncer humano. A identificação dessas alterações no genoma pode fornecer informações importantes sobre os defeitos celulares que causam o câncer e sugerir possíveis estratégias terapêuticas. Nesse raciocínio, a análise dos proto-oncogenes fornecem informações importantes sobre a patogênese molecular do CPNPC e abrem as portas para novas opções terapêuticas. Atualmente, todos os novos casos de AC são testados quanto evidências de mutação direcionável, além disso, as abordagens imunoterapêuticas para CPNPC também começaram a mostrar-se promissoras. Infelizmente, a mutação proto-oncogênica mais comum, a KRAS, que ocorre em 20% dos ACs, ainda não foi bem-sucedida em alvos farmacológicos. Dessa forma as várias técnicas que variam de imunohistoquímica, hibridização *in situ* e sequenciamento de genes são atualmente utilizadas nos laboratórios de diagnóstico para estabelecer a presença dessas mutações e posteriormente identificar os principais caminhos moleculares no câncer de pulmão para permitir o direcionamento precoce dos processos oncogênicos.

Diversos dados de *microarray* estão disponíveis publicamente, fornecendo dados de expressão de genes através da análise dos tecidos de pacientes com CPNPC. Embora outras plataformas para estudo de transcriptomas estejam mais frequentes, os dados coletados por *microarrays* e a identificação de novos biomarcadores ainda são úteis para a melhoria do diagnóstico e direcionamento do tratamento dos pacientes. Atualmente novas tecnologias estão substituindo o *microarray*, principalmente a tecnologia do sequenciamento genético, no entanto, esta tecnologia continua a gerar grandes quantidades de dados, que também demanda profissionais capacitados para as posteriores análises. O ramo da análise de dados de expressão de genes, precisa urgentemente de um avanço na maneira como se analisa tais dados, ou a coleção de dados depositados em bancos, vão ficar cada vez mais abandonados. Esses dados podem ajudar a explicar diversos processos dentro da célula, mas incapaz de prever qualquer mecanismo, já que estão sendo desvalorizados. Desse modo, a realização dessa meta-análise, mostra como esses dados podem ser valorizados. Uma vez que dados de *microarray* de livre acesso da expressão genica de tecidos tumorais e tecido normal do pulmão de pacientes com adenocarcinoma podem ser coletados e aplicados em ferramentas da bioinformática, com o objetivo de explorar genes diferencialmente expressos como bons candidatos à biomarcadores para diagnóstico e potencial alvo de terapias.

## **OBJETIVOS**

### **Objetivo Geral**

Analisar a expressão de genes em amostras de adenocarcinoma pulmonar através de meta-análise de *microarrays* para identificar conjunto de genes, vias e processos biológicos diferencialmente expressos.

### **Objetivos Específicos**

- Identificar genes diferencialmente expressos (DEG) em amostras de adenocarcinoma pulmonar e identificar genes compartilhados entre diferentes estudos.
- Construir uma rede de interação proteína-proteína codificadas pelos DEG.
- Identificar vias KEGG enriquecidas com DEG.
- Identificar os processos biológicos do Gene Ontology que os DEG estão envolvidos.
- Validar a expressão dos DEG compartilhados usando dados de imunohistoquímica.

## MATERIAL E MÉTODOS

### Amostras de estudo

Este estudo inclui amostras com câncer de pulmão, do subtipo adenocarcinoma e amostras de tecido pulmonar normal (figura 7). A análise em larga escala foi realizada através de dados do transcriptoma (*microarray*) do adenocarcinoma pulmonar. Os estudos primários usados para identificar a expressão dos genes foram obtidos no banco de dados Gene Expression Omnibus database (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) usando a pesquisa “*lung adenocarcinoma*”. A identificação dos DEG foi realizada nas amostras de adenocarcinoma pulmonar e amostras normais como referência (amostra controle). Os conjuntos de dados utilizados foram GSE10072 (Landi et al. 2008), GSE31547, GSE32665 (Kim et al. 2013), GSE32863 (Selamat et al. 2012), GSE75037 (Girard et al. 2016), GSE101929 (Mitchell et al. 2017).

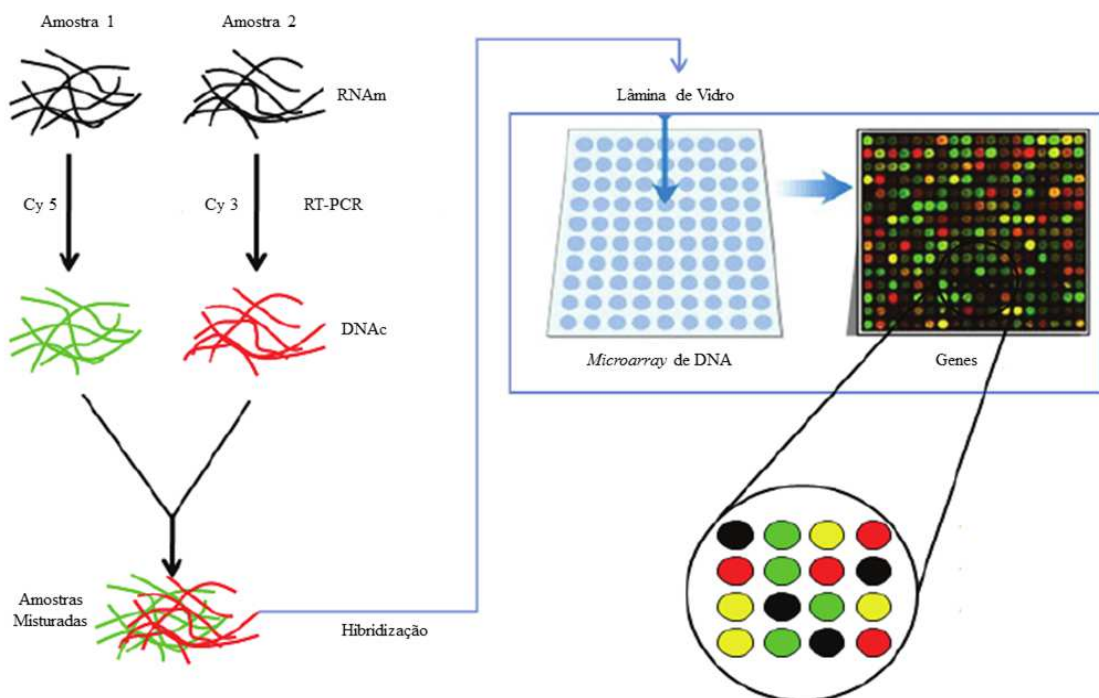


Figura 7. Comparação da expressão de um *microarray*. Amostras em verde representam amostras controle (normal), amostras em vermelho representam amostras tumorais. A intensidade de luz representa a expressão dos genes (ORTUÑO-SAHAGÚN et al. 2011).

Tabela 1. Lista de dados do câncer de pulmão selecionados pelo GEO/NCBI.

Conjunto de Dados	Plataforma	Nº de amostras	Nº de amostras controle/câncer	Referências
<b>GSE10072</b>	GPL96	107	58/49	(Landi et al. 2008)
<b>GSE31547</b>	GPL96	50	20/30	-
<b>GSE32665</b>	GPL6102	179	92/87	(Kim et al. 2013)
<b>GSE32863</b>	GPL6884	116	58/58	(Selamat et al. 2012)
<b>GSE75037</b>	GPL6884	166	83/83	(Girard et al. 2016)
<b>GSE101929</b>	GPL570	66	34/32	(Mitchell et al. 2017)
<b>Total</b>		684	345/339	

### Análise estatística

Todas as análises estatísticas foram realizadas usando a linguagem do programa R. Os dados de expressão dos conjuntos de estudos foram processados usando o pacote GEOquery (DAVIS & MELTZER, 2007) e a normalização e identificação dos DEG usando o pacote de regressão linear múltipla, limma (RITCHIE et al., 2015), disponíveis no *Bioconductor/R*. Os principais genes foram identificados usando critérios para a seleção de DEGs com valor-p ajustado por *false discovery rate* ( $FDR \leq 0,05$  e  $|\log(FC)| > 1$ ). Os DEG compartilhados entre as amostras foram chamados de assinatura gênica (AG) do adenocarcinoma de pulmão.

### KEGG e Anotação Funcional

A análise das vias super-representadas foi realizada para identificar as principais funções KEGG<sup>10</sup>, alteradas no câncer de pulmão. As vias super-representadas foram identificadas independentemente em cada conjunto de DEG proveniente dos conjuntos de dados. O pacote *clusterProfiler* do *Bioconductor/R* (Yu et al. 2012) detectou as vias KEGG super-representadas, e consideramos aquelas com um valor-p ajustado  $\leq 0,05$  e vias com cinco ou mais DEG.

<sup>10</sup> Enciclopédia de Genes e Genoma de Kioto é uma coleção de bancos de dados on-line que lidam com genoma, vias metabólicas e substâncias químicas biológicas.

## Rede de Interação Proteína-Proteína e Rede de Processos Biológicos

O banco de dados APID (Alonso-López et al. 2016) retornou a rede de interação proteína-proteína (PPI) dos DEG e, posteriormente, o *Cytoscape* (Shannon et al. 2003) foi usado para a edição de análise da rede. O *Cytoscape* calculou medidas de centralidade de grau e *betweenness centrality* (BC) da rede. O *plug-in Golorize* para o programa *Cytoscape* (Garcia et al. 2007) investigou os processos biológicos (BP) do *Gene Ontology* (GO) super-representado e classificou as proteínas da rede nesses processos. O *Golorize* identificou o BP usando teste hipergeométrico com um valor-p ajustado  $\leq 0,05$  corrigido por FDR. O *plug-in clueGO* para o programa *Cytoscape* (Bindea et al. 2009) recuperou os GO BP geral e também realizou uma pesquisa direcionada para o sistema imunológico enriquecidos com os DEG. As buscas foram realizadas e consideradas significativas com um valor-p  $\leq 0,05$  ajustado pelo FDR.

## Clusterização de amostras

As amostras de diferentes conjuntos de dados foram agrupadas para verificar como a AG foi capaz de dividir as amostras normais ou tumor. Usamos uma transformação do valor de expressão (normalização min/máx) para permitir a comparação entre os conjuntos de dados de diferentes plataformas de *microarray*. Os processos de padronização foram feitos para cada gene e conjunto de dados individualmente usando o método  $x' = \frac{x-\underline{x}}{sd}$  onde  $x$  representa os valores de expressão gênica,  $\underline{x}$  a expressão média e  $sd$  o desvio padrão (Shabalín et al. 2008). A padronização foi seguida por uma transformação do valor de expressão gênica  $x'' = \frac{x'-min}{max}$ , fazendo com que os valores de expressão estivessem no intervalo [0-1] (Sîrbu, Ruskin e Crane 2010). Os valores de expressão normalizados foram utilizados para construir um *heatmap* com agrupamento hierárquico e análise de componentes principais (PCA).

## Imunohistoquímica

O nível de expressão da proteína foi investigado no banco de dados *The Human Protein Atlas* (UHLEN et al. 2015) ([www.proteinatlas.org](http://www.proteinatlas.org)) para os genes compartilhados entre os conjuntos de *microarrays*. A expressão protéica foi coletada a partir de ensaios imunohistoquímicos de tecidos de câncer de pulmão. Os dados da expressão imunohistoquímica de diferentes anticorpos foram somados como altos, médios, baixos ou não detectados, de acordo com a classificação do *The Human Protein Atlas*. A expressão foi considerada validada quando os dados imunohistoquímicos foram iguais ou maiores que 50% dos ensaios alto +

médio para genes superexpressos e iguais ou menores que 50% dos ensaios baixo + não detectado para genes subexpressos.

## REFERÊNCIAS

ARAUJO, L. H.; BALDOTTO, C.; CASTRO, G. DE; KATZ, A.; FERREIRA, C. G.; MATHIAS, C.; MASCARENHAS, E.; LOPES, G. DE L.; CARVALHO, H.; TABACOF, J.; MARTÍNEZ-MESA, J.; VIANA, L. DE S.; CRUZ, M. DE S.; ZUKIN, M.; MARCHI, P. DE; TERRA, R. M.; RIBEIRO, R. A.; WERUTSKY, V. C. C. DE L. G.; BARRIOS, C. H. Lung cancer in Brazil. **J Bras Pneumol.**, p. 55–64, 2018.

AVRAHAM, R.; YARDEN, Y. Feedback regulation of EGFR signalling: Decision making by early and delayed loops. **Nature Reviews Molecular Cell Biology**, v. 12, n. 2, p. 104–117, 2011.

BEROUKHIM, R.; MERMEL, C. H.; PORTER, D.; WEI, G.; DONOVAN, J.; BARRETINA, J.; BOEHM, J. S.; URASHIMA, M.; HENRY, K.; T MC.; PINCHBACK, R. M.; AZRA, H.; CHO, Y.; HAERY, L.; GREULICH, H.; REICH, M.; LAWRENCE, M. S.; WEIR, B.; TANAKA, K. E.; DEREK, Y.; BASS, A. J.; LOO, A.; HOFFMAN, C.; PRENSNER, J.; GAO, Q.; YECIES, D.; SIGNORETTI, S.; MAHER, E.; FREDERIC, J.; SASAKI, H.; TEPPER, J. E.; FLETCHER, J.; TABERNEIRO, J.; BASELGA, J.; TSAO, MING-SOUND.; DEMICHELI, F.; RUBIN, M.; PASI, A.; DALY, M. J.; NUCERA, C.; LEVINE, R. L.; EBERT, B. L.; GABRIEL, S.; RUSTGI, A. K.; ANTONESCU, C. R.; LADANYI, M.; GARRAWAY, L.; LODA, M.; BEER, D. G.; TRUE, L. D.; POMEROY, S. L.; SINGER, S.; GOLUB, T.R.; ERIC, S. Cancers. **Genetics**, v. 463, n. 7283, p. 899–905, 2010.

BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-Wide Association Studies. **PLoS Computational Biology**, v. 8, n. 12, 2012.

CHIRIEAC, L. R.; ATTANOOS, R. L. 26 – Usual Lung Cancers. **Pulmonary Pathology**, p. 534–551, 2018.

DISTEFANO JK, T. D. **Technological issues and experimental design of gene association studies**. [s.l.] Methods Mol Biol, 2012.

GEISINGER, K. R.; SPENCER RAAB, S. **Metastatic Tumors in the Lung**. Third Edit ed. [s.l.] Elsevier Inc., 2018.

GESTHALTER, Y. B.; BILLATOS, E.; KATHURIA, H. Chapter 9 - Lung Cancer. In: DAVID, S. P. B. T.-G. AND P. M. (THIRD E. (Ed.). . Boston: Academic Press, 2017. p. 165–180.

GIANGRECO, A.; GROOT, K. R.; JANES, S. M. Lung cancer and lung stem cells: Strange bedfellows? **American Journal of Respiratory and Critical Care Medicine**, v. 175, n. 6, p. 547–553, 2007.

GIANGRECO, A.; REYNOLDS, S. D.; STRIPP, B. R. Terminal bronchioles harbor a unique airway stem cell population that localizes to the bronchoalveolar duct junction. **American Journal of Pathology**, v. 161, n. 1, p. 173–182, 2002.

GIBSON, G. Microarray analysis: Genome-scale hypothesis scanning. **PLoS Biology**, v. 1, n. 1, p. 28–29, 2003.

GILLIS, J.; PAVLIDIS, P. “Guilt by association” is the exception rather than the rule in gene

networks. **PLoS Computational Biology**, v. 8, n. 3, 2012.

GORDON, M.; BAKSH, S. *Rassf1a*. **Small GTPases**, v. 2, n. 3, p. 148–157, 2011.

GUSTAFSON, A. M.; Soldi, R.; Anderlind, C.; Scholand, M. B.; Zhang, X.; Cooper, K.; Walker, D.; McWilliams, A.; Liu, G.; Szabo, E.; Brody, J.; Massion, P. P.; Lenburg, M. E.; Lam, S.; Bild, A. H.;

Spira, A. NIH Public Access. v. 2, n. 26, 2013.

HANNA, J.; ONAITIS, M. Cell of origin of lung cancer. **Journal of Carcinogenesis**, v. 12, n. 1, p. 6, 2013.

HOGAN, B. L. M.; BARKAUSKAS, CHRISTINA E.; CHAPMAN, H. A.; JONATHAN, A.; JAIN, R.; HSIA, C. C. W.; NIKLASON, L.; CALLE, E.; LE, A.; RANDELL, S. H.; ROCK, J.; SNITOW, M.; KRUMMEL, M.; STRIPP, B. R.; VU, T.; WHITE, E. S.; WHITSETT, J. A.; MORRISEY, E. E. NIH Public Access. **Cell Stem Cell**, v. 15, n. 2, p. 123–138, 2015.

HYNES, N. E.; MACDONALD, G. ErbB receptors and signaling pathways in cancer. **Current Opinion in Cell Biology**, v. 21, n. 2, p. 177–184, 2009.

KABBOU, M.; GARCIA, M. M.; FUJIMOTO, J.; LIU, D. D.; WOODS, D.; CHOW, C. W.; MENDOZA, G.; MOMIN, A. A.; JAMES, B. P.; SOLIS, L.; BEHRENS, C.; LEE, J. J.; WISTUBA, I. I.; KADARA, H. ETS2 mediated tumor suppressive function and MET oncogene inhibition in human non-small cell lung cancer. **Clinical Cancer Research**, v. 19, n. 13, p. 3383–3395, 2013.

KHATRI, P.; SIROTA, M.; BUTTE, A. J. Ten years of pathway analysis: Current approaches and outstanding challenges. **PLoS Computational Biology**, v. 8, n. 2, 2012.

KIM, M. J.; Lee, J. Y.; STANLEY, E. G.; ELEFANTY, A. G.; PETRATOS, S. Efficient derivation of oligodendrocyte progenitors from nkx2.1-gfp human embryonic stem cell reporter line. **International Journal of Developmental Neuroscience**, v. 47, p. 69, 2015.

KIM, C. F. B.; JACKSON, E. L.; WOOLFENDEN, A. E.; LAWRENCE, S.; BABAR, I.; VOGEL, S.; CROWLEY, D.; BRONSON, R. T.; JACKS, T. Identification of bronchioalveolar stem cells in normal lung and lung cancer. **Cell**, v. 121, n. 6, p. 823–835, 2005.

KOSTINA, E. V; SINYAKOV, A. N.; RYABININ, V. A. RESEARCH PAPER A many probes-one spot hybridization oligonucleotide microarray. 2018.

KOTTON, D. N.; MORRISEY, E. E. Lung regeneration: Mechanisms, applications and emerging stem cell populations. **Nature Medicine**, v. 20, n. 8, p. 822–832, 2014.

KRADIN, R. L. Lung Cancer. **Understanding Pulmonary Pathology: Applying Pathological Findings in Therapeutic Decision Making**, p. 243–269, 2017.

KRIZHANOVSKY, V.; LOWE, S. W. NIH Public Access. v. 460, n. 7259, p. 1085–1086, 2011.

LANDI, D.; GEMIGNANI, F.; BARALE, R.; LANDI, S. A. Catalog of Polymorphisms Falling in MicroRNA-Binding Regions of Cancer Genes. **DNA and Cell Biology**, v. 27, n. 1, p. 35–43, 17 out. 2007.

LINDEMAN, N. I.; CAGLE, P. T.; BEASLEY, M. B.; CHITALE, D. A.; DACIC, S.; GIACCONE, G.; JENKINS, R. B.; KWIATKOWSKI, D. J.; SALDIVAR, J. S.; SQUIRE, J.; THUNNISSEN, E.; LADANYI, M. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: Guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. **Journal of Thoracic Oncology**, v. 8, n. 7, p. 823–859, 2013.

LIST, M.; HAUSCHILD, A. C.; TAN, Q.; KRUSE, T.; MOLLENHAUER, J.; BAUMBACH, J.; BATRA, R. Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data. **Journal of integrative bioinformatics**, v. 11, n. 2, p. 236, 2014.

MAITY, T. K.; Venugopalan, A.; Linnoila, I.; Cultraro, C. M.; Giannakou, A.; Nemati, R.; Zhang, X.; Webster, J. D.; Ritt, D.; Ghosal, S.; Hoschuetzky, H.; Mark Simpson, R.; Biswas, R.; Politi, K.; Morrison, D. K.; Varmus, H. E.; Guha, U. Loss of MIG6 accelerates initiation and progression of mutant epidermal growth factor receptor–driven lung adenocarcinoma. **Cancer Discovery**, v. 5, n. 5, p. 534–549, 2015.

MEERT, A.-P. et al. The role of HER-2/neu expression on the survival of patients with lung cancer: a systematic review of the literature. **British Journal of Cancer**, v. 89, n. 6, p. 959–965, 2003.

MENDELSON, J.; Howley, P. M.; Israel, M. A.; Gray, J. W. B.; Thompson, C. **The molecular basis of cancer**. [s.l.: s.n.].

MOLINA, J. R.; ADJEI, A. A. The Ras/Raf/MAPK pathway. **Journal of Thoracic Oncology**, v. 1, n. 1, p. 7–9, 2006.

MUNNIA, A.; Giese, R. W., Polvani, S.; Galli, A.; Cellai, F.; Peluso, M. E.M. Bulky DNA Adducts, Tobacco Smoking, Genetic Susceptibility, and Lung Cancer Risk. **Advances in Clinical Chemistry**, v. 81, p. 231–277, 1 jan. 2017.

NANAVATI, K. B.; REED, E. **Lung Cancer**. Fourth Edi ed. [s.l.] Elsevier Inc., 2018.

NAOKI, K.; Chen, Tzu-Hsiu; Richards, W. G; Sugarbaker, D. J.; Meyerson, M. Missense mutations of the BRAF gene in human lung adenocarcinoma. **Cancer research**, v. 62, n. 23, p. 7001–3, 2002.

PAULING, J. K.; Hristensen, A. G.; Batra, R.; Alcaraz, N.; Barbosa, E.; Larsen, M. R.; Beck, H. C.; Leth-Larsen, R.; Azevedo, V.; Ditzel, H. J.; Baumbach, J. Elucidation of epithelial-mesenchymal transition-related pathways in a triple-negative breast cancer cell line model by multi-omics interactome analysis. **Integrative Biology (United Kingdom)**, v. 6, n. 11, p. 1058–1068, 2014.

PÉREZ-CARO, M.; Cobaleda, C.; González-Herrero, I.; Vicente-Dueñas, C.; Bermejo-Rodríguez, C.; Sánchez-Beato, M.; Orfao, A.; Pintado, B.; Flores, T.; Sánchez-Martín, M.; Jiménez, R.; Piris, M. A.;

Sánchez-García, I. Cancer induction by restriction of oncogene expression to the stem cell compartment. **The EMBO Journal**, v. 28, n. 1, p. 8–20, 2009.

PRIOR, I. A.; LEWIS, P. D.; MATTOS, C. A comprehensive survey of ras mutations in cancer.



**Cancer Research**, v. 72, n. 10, p. 2457–2467, 2012.

PURI, S.; FOLIAS, A. E.; HEBROK, M. Plasticity and dedifferentiation within the pancreas: Development, homeostasis, and disease. **Cell Stem Cell**, v. 16, n. 1, p. 18–31, 2015.

QIU, P.; GENTLES, A. J.; PLEVITIS, S. K. Discovering biological progression underlying microarray samples. **PLoS Computational Biology**, v. 7, n. 4, 2011.

RAWLINS, E. L.; Okubo, T.; Xue, Y.; Brass, D. M.; Auten, R. L.; Hasegawa, H.; Wang, F.; Hogan, B. L. M. The Role of Scgb1a1+ Clara Cells in the Long-Term Maintenance and Repair of Lung Airway, but Not Alveolar, Epithelium. **Cell Stem Cell**, v. 4, n. 6, p. 525–534, 2009.

REIMERS, M. Making informed choices about microarray data analysis. **PLoS Computational Biology**, v. 6, n. 5, p. 1–7, 2010.

REYA, T.; Morrison, S.J.; Clarke, M.F.; Weissman, I.L. Stem Cells, Cancer, and Cancer Stem Cells. **Nature**, v. 414, n. November, p. 105–111, 2001.

ROY, N.; HEBROK, M. Regulation of Cellular Identity in Cancer. **Developmental Cell**, v. 35, n. 6, p. 674–684, 2015.

RUSSO, G.; ZEGAR, C.; GIORDANO, A. Advantages and limitations of microarray technology in human cancer. **Oncogene**, v. 22, n. 42, p. 6497–6507, 2003.

SEKIDO, Y.; FONG, K. M.; MINNA, J. D. Progress in understanding the molecular pathogenesis of human lung cancer. **Biochimica et Biophysica Acta (BBA) - Reviews on Cancer**, v. 1378, n. 1, p. F21–F59, 1998.

SELAMAT, S. A.; CHUNG, B. S.; GIRARD, L.; LEVY, M.; LOVLY, C. M.; PAO, W.; SELAMAT, S.; CHUNG, B. S.; GIRARD, L.; ZHANG, W.; ZHANG, Y.; CAMPAN, M.; SIEGMUND, K. D.; KOSS, M. N.; HAGEN, J.; LAM, W. L.; LAM, S.; GAZDAR, A. F.; LAIRD-OFFRINGA, I. Adenocarcinoma and integration with mRNA expression paradigm Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. **Genome Research**, v. 22, p. 1197–1211, 2012.

SHIAU, C. J.; TSAO, M.-S. Molecular Testing in Lung Cancer. In: **Diagnostic Molecular Pathology**. [s.l.] Elsevier, 2017. p. 287–303.

SIEGEL, R.; MILLER, K. D.; AHMEDIN, J. Cáncer Statistics. **Ca Cáncer Journal**, v. 67, n. 1, p. 7–30, 2017.

SILVA, S. T. DA.; MARTINS, M. C.; FARIA, F. R. DE; COTTA, R. M. M. Combate ao Tabagismo no Brasil: a importância estratégica das ações governamentais. **Ciência & Saúde Coletiva**, v. 19, n. 2, p. 539–552, 2014.

SLONIM, D. K.; YANAI, I. Getting started in gene expression microarray analysis. **PLoS Computational Biology**, v. 5, n. 10, p. 3–6, 2009.

SMALLEY, M.; ASHWORTH, A. Stem cells and breast cancer: A field in transit. **Nature Reviews Cancer**, v. 3, n. 11, p. 832–844, 2003.

SODA, M.; Choi, YOUNG L.; ENOMOTO, M.; TAKADA, S.; YAMASHITA, Y.; ISHIKAWA, S.; FUJIWARA, S. I.; WATANABE, H.; KURASHINA, K.; HATANAKA, H.; BANDO, M.; OHNO, S.; ISHIKAWA, Y.; ABURATANI, H.; NIKI, T.; SOHARA, Y.;

SUGIYAMA, Y.; MANO, H. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. **Nature**, v. 448, n. 7153, p. 561–566, 2007.

STEARNS, L.; MARTINSKT, T.; SCHEANA, M. Trends in Microarray Analysis. *Nature Medicine*. v. 9, n. 1, p. 140–145, 2003.

SUTHERLAND, K. D.; PROOST, N.; BROUNS, I.; ADRIAENSEN, D.; SONG, J. Y.; BERNIS, A. Cell of origin of small cell lung cancer: Inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. **Cancer Cell**, v. 19, n. 6, p. 754–764, 2011.

TAI, M. C.; KAJINO, T.; NAKATOCHI, M.; ARIMA, C.; SHIMADA, Y.; SUZUKI, M.; MIYOSHI, H.; YATABE, Y.; YANAGISAWA, K.; TAKAHASHI, T. miR-342-3p regulates MYC transcriptional activity via direct repression of E2F1 in human lung cancer. **Carcinogenesis**, v. 36, n. 12, p. 1464–1473, 2015.

TAKAHASHI, K.; YAMANAKA, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. **Cell**, v. 126, n. 4, p. 663–676, 2006.

TEBBUTT, N.; PEDERSEN, M. W.; JOHNS, T. G. Targeting the ERBB family in cancer: Couples therapy. **Nature Reviews Cancer**, v. 13, n. 9, p. 663–673, 2013.

TRAN, T.; KWON, Y. Hierarchical closeness-based properties reveal cancer survivability and biomarker genes in molecular signaling networks. p. 1–14, 2018.

VENTURA, J. J.; TENBAUM, S.; PERDIGUERO, E.; HUTH, M.; GUERRA, C.; BARBACID, M.; PASPARAKIS, M.; NEBREDA, A. R. p38 $\alpha$  MAP kinase is essential in lung stem and progenitor cell proliferation and differentiation. **Nature Genetics**, v. 39, n. 6, p. 750–758, 2007.

VICENTE-DUEÑAS, C.; HAUER, J.; COBALEDA, C.; BORKHARDT, A.; SÁNCHEZ-GARCÍA, I. Epigenetic Priming in Cancer Initiation. **Trends in Cancer**, v. 4, n. 6, 2018.

WANG, M.; VIKIS, H. G.; WANG, Y.; JIA, D.; WANG, D.; BIERUT, L. J.; BAILEY-WILSON, J. E.; AMOS, C. I.; PINNEY, S. M.; PETERSEN, G. M.; DE ANDRADE, M.; YANG, P.; WIEST, J. S.; FAIN, P. R.; SCHWARTZ, A. G.; GAZDAR, A.; MINNA, J.; GABA, C.; ROTHSCILD, H.; MANDAL, D.; KUPERT, E.; SEMINARA, D.; LIU, Y.; VISWANATHAN, A.; GOVINDAN, R.; ANDERSON, M. W.; YOU, M. Identification of a novel tumor suppressor gene p34 on human chromosome 6q25.1. **Cancer Research**, v. 67, n. 1, p. 93–99, 2007.

WILSON, F. H.; POLITI, K. ERBB Signaling Interrupted: Targeting Ligand-Induced Pathway Activation. **Cancer discovery**, v. 8, n. 6, p. 676–678, 2018.

WIMMER, I. et al. Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. **Scientific Reports**, v. 8, n. 1, p. 1–17, 2018.

WONG, M. C. S.; LAO, X. Q.; HO, KIN-F.; GOGGINS, W. B.; TSE, S. L. A. Incidence and mortality of lung cancer: global trends and association with socioeconomic status. **Scientific Reports**, v. 7, n. 1, p. 14300, 2017.

XINARIANOS, G.; MCRONALD, F. E.; RISK, J. M.; BOWERS, N. L.; NIKOLAIDIS, G.; FIELD, J. K.; LILOGLOU, T. Frequent genetic and epigenetic abnormalities contribute to the deregulation of cytoglobin in non-small cell lung cancer. **Human Molecular Genetics**, v. 15, n. 13, p. 2038–2044, 2006.

YANG, P. Epidemiology of Lung Cancer Prognosis: Quantity and Quality of Life. **Methods Mol Biol**, v. 471, p. 469–486, 2009.

YARDEN, Y.; PINES, G. The ERBB network: At last, cancer therapy meets systems biology. **Nature Reviews Cancer**, v. 12, n. 8, p. 553–563, 2012.

ZHANG, Y.; GOSS, A. M.; COHEN, E. D.; KADZIK, R.; LEPORE, J. J.; MUTHUKUMARASWAMY, K.; YANG, J.; DEMAYO, F. J.; WHITSETT, J. A.; PARMACEK, M. S.; MORRISEY, E. E. A Gata6-Wnt pathway required for epithelial stem cell development and airway regeneration. **Nature Genetics**, v. 40, n. 7, p. 862–870, 2008.

ZHANG, Y.; GOSS, A. M.; COHEN, E. D.; KADZIK, R.; LEPORE, J. J.; MUTHUKUMARASWAMY, K.; YANG, J.; DEMAYO, F. J.; WHITSETT, J. A.; PARMACEK, M. S.; MORRISEY, E. E. Development and Airway Regeneration. v. 40, n. 7, p. 862–870, 2009.

ZHOU, C.; WU, Y. L.; CHEN, G.; FENG, J.; LIU, X. Q.; WANG, C.; ZHANG, S.; WANG, J.; ZHOU, S.; REN, S.; LU, S.; ZHANG, L.; HU, C.; HU, C.; LUO, Y.; CHEN, L.; YE, M.; HUANG, J.; ZHI, X.; ZHANG, Y.; XIU, Q.; MA, J.; ZHANG, L.; You, C. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. **The Lancet Oncology**, v. 12, n. 8, p. 735–742, 2011.

## CAPÍTULO 1

Trabalho para ser enviado à revista **CANCER GENOMICS & PROTEOMICS / ISSN: 1109-6535**

### **Title: Differentially gene expression in pulmonary adenocarcinoma meta-analysis**

Full names and full postal addresses of all Authors,

**Authors:** Fernanda Castro Pires dos Santos<sup>1</sup>; Cristina Padre Cardoso<sup>1,3</sup>; Leandro Martins de Freitas<sup>1\*</sup>, Sávio Torres de Farias<sup>2</sup>;

#### **Filiation:**

1 Laboratório de Genética e Bioinformática, Núcleo de Biointegração, Universidade Federal da Bahia, Rua Rio de Contas, 58, Vitória da Conquista, BA, 45029-094, Brazil

2 Departamento de Biologia Molecular, Universidade Federal da Paraíba, João Pessoa, PB, 58059-900, Brazil

3 Faculdade Santo Agostinho, Vitória da Conquista, Brazil

\*Correspondence: [leandromartins@ufba.br](mailto:leandromartins@ufba.br)

#### **Abstract**

Over the past decade, comprehensive efforts have been made to understand the biomarkers of non-small cell lung cancer (NSCLC). This is of clinical relevance, the new biomarkers can be screened in samples of cancer patients, and the therapeutic regimens can be improved. The objective was to apply the computational tools in NSCLC samples to investigate the differentially expressed genes (DEG), and how they interact, and activate biological process.

We used six microarray dataset of the lung adenocarcinoma subtype. The tumor and normal samples were retrieved, and DEG were identified with the Bioconductor/R liming package using  $|\log_{2}FC| \geq 1$  and adjusted p-value  $\leq 0.05$ . The DEG was used to find the KEGG enriched pathways using Cluster Profiler Bioconductor/R package. A protein-protein interaction network was build

in the APID database and analyzed with Cytoscape. The sample clustering were performed using mRNA expression values and displayed as heatmap and principal component analysis (PCA). The Human Protein Atlas was used to validate the expression microarray results using immunohistochemical information.

We have identified that there are more downregulated genes than upregulated in NSCLC samples. We obtained 49 genes upregulated, and 164 genes downregulated shared by all the six studies. The KEGG pathways enriched by upregulated genes are involved in cellular replication and by the downregulated genes are more diverse. The PPI network constructed reveal CDH1 and XPO1 proteins as hub nodes in the upregulated network/ while the downregulated indicates JUN and MYC. The heatmap and PCA showed clusterization of the samples, although there are some few samples distant from their group. The analysis in Human Protein Atlas showed 13 upregulated genes present with high or medium expression in more than 50% of the immunohistochemical samples, and 133 downregulated genes confirmed as validated by Protein Atlas as low or not detected expression.

This study demonstrated that our computational methodologies for the separation of DEG with defined patterns could be used to identify patients with NSCLC, since it reveals some candidate genes for the genetic signature (GS) in adenocarcinoma. Therefore, microarray data helps to identify diagnostic and predictive factors that can guide a personalized therapy approach. Also, understanding biological processes is one of the most pertinent needs in cancer research, so doctors have significant predictive power. Future studies should validate these results in ways of diagnosing low cost and accessible to public hospitals.

**Keywords:** lung adenocarcinoma, gene expression, protein network

## Introduction

The lung cancer is a respiratory disease with an estimative of 1.8 million new cases in 2012 (82). It is the second most common type of cancer in both sexes (about 13%), behind prostate in males and breast in females, in the USA (82). Besides the second most common, it is the first leading to death, around 25% of deaths among all cancers. After diagnosis, the five-year

survival is about 15-18%, one with the lowest survival rate (71, 82). The risk factors associated with lung cancer are tobacco and pollution, causing to somatic mutations (33). Countries with tobacco control have achieved a reduction in lung cancer, showing the success of this program (82). Despite studies in understanding the molecular pathogenesis, therapeutic advances, and decrease in rates in some countries, lung cancer is still the most common cause of deaths regarding all cancer, about 1.4 million deaths annually (81).

Despite campaigns against tobacco and the reduction of smoking, the high number of deaths demands new investigations to understand the progression of the disease and new strategies for diagnosis and treatment. The genes associated with lung cancer can put a light on this issue. It is essential the identification of genes expressed and responsible for cancer development and this information can be helpful in cancer treatment choice because these genes can be the target of drugs improving the treatment of patients. Genome screening identified genes associated with lung cancer as LKB1 (50, 32), BRAF (33), SLC22A18 (40), KRAS (3, 33), ERBB2 (66, 6), PPP2R1B (86), IRF1 (41), and CASP8 (76, 70). Some researchers pointed a set of genes associated with NSCLC as DUSP6, MMD, STAT1, ERBB3, and UBE2C, TPX2, MCM2, MCM6, FEN1, and SFN (REXO2) (35, 36). Some genes are involved in DNA replication (REXO2, MCM2, MCM2), and ErbB signaling pathway (BRAF, KRAS, EGFR, ERBB2, PIK3CA) which in KEGG pathway would result in activation of the process involved in cancer progression as survival/proliferation, angiogenesis, migration/invasion.

The KRAS gene plays a crucial role in the development of cancer, also associate with lung cancer, and after decades of research investigating a possible inhibitor it appears to be undruggable (13, 51). However, some types of NSCLC present mutations in the EGFR gene that can be deactivated by tyrosine kinase inhibitor (TKI) (49). Genetic investigation allows the descriptions of mutations that are responsive to treatments like the common mutations in EGFR present a higher survival rate contrasted with rare mutations when treated with gefitinib (87). New researchers also showed that microRNA-128b regulates EGFR, and demonstrated the correlation of this miRNA and survival of patients treated with gefitinib (88).

The microarray has been used in cancer research to discover changes in gene expression, biomarkers, classification of subtypes of cancer, and identification of biological functions (72, 18). The immune profile of the lung cancer pointed differences among the types of lung cancer. While the secretory subtype has an elevated immune cell expression, the adenocarcinoma has low

expression of immune cells (18). The gene expression helps to understand and improve the cancer treatment. The genes associated with lung cancer has been used in the diagnosis of suspicious cases through investigation of autoantibodies and concentration in the patients' serum. The use of proteins, peptides, mi-RNA has been proposed and tested as indicative of lung cancer.

To understand the molecular changes involved in lung adenocarcinoma, we analyzed six microarray datasets and checked the common genes differentially expressed among these data. The gene expression data showed 49 genes upregulated, and 164 downregulate shared by the six datasets (Gene Signature, GS). The immunohistochemical images from Protein Atlas confirmed 13 and 56 genes as upregulated and downregulated respectively. These results indicate a set of genes shared by lung cancer cells.

## Material and methods

### 1. Study samples

The datasets from lung cancer adenocarcinoma were retrieved from Gene Expression Omnibus database/NCBI. The search was performed using the term “lung adenocarcinoma” and selecting datasets with mRNA data expression and studies with control and cancer samples. The procedure was, for each study retrieve, create two groups (adenocarcinoma and control) to compare and identify DEG (table 1).

Table 1. List of lung cancer datasets selected in GEO/NCBI.

<b>Dataset</b>	<b>Platform</b>	<b>#samples (control/cancer)</b>	<b>Reference</b>
<b>GSE10072</b>	GPL96	107 (58/49)	(39)
<b>GSE31547</b>	GPL96	50 (20/30)	-
<b>GSE32665</b>	GPL6102	179 (92/87)	(38)
<b>GSE32863</b>	GPL6884	116 (58/58)	(65)
<b>GSE75037</b>	GPL6884	166 (83/83)	(23)
<b>GSE101929</b>	GPL570	66 (34/32)	(53)
<b>Total</b>		684 (345/339)	

- The data in the Gene Expression Omnibus without publication associated
- The number of samples is presented as control and cancer in each study

### 2. Data processing

In each dataset, the microarray data were retrieved with GEOquery Bioconductor/R package (64), and mRNA expression values were used aslog<sub>2</sub> scale. A nonspecific filter removed gene expression without variation using the genefilter Bioconductor/R package (22). The WCNA Bioconductor/R package selected the spot from the same gene with maximum variance without the control/tumor group information. After grouping samples as control or tumor, the limma



Bioconductor/R package (61) identified the DEG as those with  $|\logFC| \geq 1$  and adjusted p-value  $\leq 0.05$ . The list of DEGs was investigated in the Malacards database Score  $> 1$  was analyzed for gene association with NSCLC or human diseases (59).

### 3. KEGG pathway overrepresentation and functional annotation

The overrepresented pathways analysis was performed to identify the high-level functions changed in the lung cancer. Overrepresented pathways were identified independently in each set of DEG coming from the different studies. The clusterProfiler Bioconductor/R package (93) detected the overrepresented KEGG pathway, and we considered those with an adjusted p-value  $\leq 0.05$  and pathways with at least five DEG.

### 4. Protein-Protein interaction and biological process network

The APID database (5) retrieve the protein-protein interaction (PPI) information of the DEG, and later we used the Cytoscape (68) to build, edit, and analyze the networks. The Cytoscape calculated the centrality measures degree and betweenness centrality (BC) in the network (minimum centrality score  $\geq 0.05$ ). The Golorize Cytoscape plug-in (21) investigated the overrepresented Gene Ontology (GO), biological process (BP), and classified the proteins in these processes. The Golorize identified the BP using a hypergeometric test with an adjusted p-value of 0.05 and corrected by false discovery rate (FDR).

### 5. Samples clusterization

The samples from different datasets were clustered to check how the GS was able to divide the healthy from tumor samples. We used a transformation of expression value to allow the comparison among the datasets as follow: the standardization processes was done for each gene and dataset individually using the method  $x' = \frac{x - \bar{x}}{sd}$ , where  $x$  represents the gene expression values,  $\bar{x}$  the average expression, and  $sd$  standard deviation (67). The standardization was followed by a transformation of gene expression value  $x'' = \frac{x' - min}{max}$ , making the

expression values be in a range [0-1] (73). The expression values standardized were used to construct a heatmap with hierarchical clusterization and principal component analysis (PCA).

## **6. Immunohistochemistry**

The protein expression level was investigated in The Human Protein Atlas (85) ([www.proteinatlas.org](http://www.proteinatlas.org)) for those genes shared among datasets. The protein expression was collected from immunohistochemistry assays from lung cancer tissue. The immunohistochemistry expression data was summarized as high, medium, low, or not detected, accordingly The Protein Atlas classification. The expression was considered validated when immunohistochemical data was equal or greater than 50% of the assays, high + medium regarding for upregulated genes, and low + not detected for downregulated genes.

## Results

### Study samples

The molecular characterization of NSCLC helps to understand the changes in gene and protein expression that leads to uncontrolled cell division. Besides that, the identification of genes that are essential to the development of NSCLC could be useful in new targets drug design. We used six microarrays NSCLC studies, and we identified that the number of upregulated genes is lower than downregulated in all samples (Table 2), this pattern indicates the NSCLC downregulate more genes than upregulate.

Table 2. Number of DEG up and downregulated in the six microarray studies used.

<b>Dataset</b>	<b>#Genes up</b>	<b>#Gene down</b>
<b>GSE10072</b>	228	427
<b>GSE31547</b>	129	448
<b>GSE32665</b>	505	647
<b>GSE32863</b>	513	752
<b>GSE75037</b>	1536	1828
<b>GSE101929</b>	1347	1866

We considered the DEG shared by all studies as critical genes for the NSCLC and from now on we refer to this set as gene signature (GS). The DEG shared by all samples are 49 genes upregulated and 164 genes downregulated. The pattern of more genes downregulated than upregulated was also observed in the GS. The list of DEG was investigated to check if they are associated with NSCLC using information from Malacards database. The Malacards returned the score indicating the association of the gene with human diseases. We found the upregulated gene CEACAM1 already associated with cancer with a score of 7.36 (table 3). The other genes

presented association and high score, but they did not present an association with cancer or pulmonary diseases.

Table 3. Frequency and Malacards score of the lung cancer DEG upregulated

	<b>GeneID</b>	<b>Freq</b>	<b>Gene Symbol</b>	<b>uniprot</b>	<b>Malacards (disease)</b>	<b>Malacard score</b>
<b>1</b>	1311	6	COMP	P02461	Epiphyseal Dysplasia, Multiple, 1	13.326
<b>2</b>	1281	6	COL3A1	P00751	Ehlers-Danlos Syndrome, Vascular Type	9.383
<b>3</b>	56521	6	DNAJC12	P39900	Hyperphenylalaninemia, Mild, Non-Bh4-Deficient	7.906
<b>4</b>	6690	6	SPINK1	Q12834	Pancreatitis, Hereditary	7.589
<b>5</b>	634	6	CEACAM1	Q06210	Colorectal Cancer	7.360
<b>7</b>	3149	6	HMGB3	P21741	Microphthalmia, Syndromic 13	7.018
<b>8</b>	6768	6	ST14	Q3KP66	Ichthyosis, Congenital, Autosomal Recessive 11	6.356
<b>9</b>	1001	6	CDH3	NA	Ectodermal Dysplasia, Ectrodactyly, and Macular Dystrophy Syndrome	6.335
<b>10</b>	1300	6	COL10A1	P00995	Metaphyseal Chondrodysplasia, Schmid Type	6.842

The downregulated genes presented a higher score in the Malacard database compared with the upregulated (table 4). The downregulated genes did not present an association with cancer. However, the gene ABCA3 is associated with Surfactant Metabolism Dysfunction, Pulmonary 3. The ABCA3 protein is associated with the product of the lung surfactant and the absence of this protein cause breathing problems (74, 28). There is a variation in phenotype among the individuals carrying the mutation in ABCA3 that could be due to the environment such as smoke and infection (7). The long smoking history could cause somatic mutations or epigenetic changes that cause the reduced expression of ABCA3 and the surfactant lung problem.

Table 4. Frequency and Malacards score of the lung cancer DEG downregulated

	<b>GeneID</b>	<b>Freq</b>	<b>Gene Symbol</b>	<b>uniprot</b>	<b>Malacards(disease)</b>	<b>Malacard score</b>
<b>1</b>	3949	6	LDLR	P53805	Hypercholesterolemia, Familial	23.689
<b>2</b>	7466	6	WFS1	Q99518	Wolfram Syndrome 1	17.512
<b>3</b>	948	6	CD36	Q9UPQ0	Platelet Glycoprotein Iv Deficiency	16.894
<b>4</b>	8547	6	FCN3	Q9NRD9	Ficolin 3 Deficiency	16.876
<b>5</b>	2624	6	GATA2	P24310	Immunodeficiency 21	16.763
<b>6</b>	4023	6	LPL	P55083	Familial Lipoprotein Lipase Deficiency	16.401
<b>7</b>	21	6	ABCA3	P07550	Surfactant Metabolism Dysfunction, Pulmonary 3	16.018
<b>8</b>	1410	6	CRYAB	O76024	Myopathy, Myofibrillar, 2	14.735
<b>9</b>	33	6	ACADL	P41134	Acyl-Coa Dehydrogenase, Very Long-Chain, Deficiency of	12.924
<b>10</b>	2690	6	GHR	P26447	Growth Hormone Insensitivity, Partial	12.399

### **KEGG pathway overrepresentation**

We used the total number of DEG in each studied to identify the KEGG pathway enriched and find the most frequent as the high-level function changed in NSCLC (Table 5). The enriched KEGG pathways can give an overview of the modifications that are happening in the NSCLC cells. The upregulated genes are present in 10 KEGG enriched pathways. The most common KEGG pathway identified was Cell cycle in four studies. The Cell cycle pathway has 124 genes, and it was identified 20 upregulated (16%). The following pathways are Biosynthesis of amino acids, and p53 signaling pathway, both present in three studies with 18 and 19% of genes upregulated.

Table 5. Pathway KEGG of Genes UP.

<b>Pathway</b>	<b>Freq</b>	<b>Number of genes</b>	<b>Number of genes in the pathway</b>	<b>Percentage mean number of pathway</b>
Cell cycle	4	20	124	16,0%
Biosynthesis of amino acids	3	133	74	18,0%
p53 signaling pathway	3	13	68	19,0%
ECM-receptor interaction	2	7	82	9,0%
Protein digestion and absorption	2	7	90	7,0%

The higher number of downregulated genes resulted in a number 89 of enriched KEGG pathways. The pathways enriched with downregulated genes are most shared compared with the upregulated (table 6). The pathways downregulated had lower expression in 15-26% of the genes. The most common pathways are involved in Endocrine and metabolic diseases (AGE-RAGE signaling pathway in diabetic complications), Cardiovascular diseases (Fluid shear stress and atherosclerosis), Immune system (Complement and coagulation cascades), and Infectious diseases (Malaria).

This result showed that few pathways are upregulated and shared by the NSCLC. That could indicate that these studies have gene expression very different. However, the total number of activated pathways are reduced and is more like the NSCLC induce few pathways. Taking a look at the downregulated pathways, we can realize that the studies shared enriched pathways, four pathways are present in five studies. Thus these studies have similar gene expression.

Table 6. Pathway KEGG of Genes DOWN.

<b>Pathway</b>	<b>Freq</b>	<b>Number of genes</b>	<b>Number of genes in the pathway</b>	<b>Percentage mean number of pathway</b>
----------------	-------------	------------------------	---------------------------------------	--

AGE-RAGE signaling pathway in diabetic complications	5	17.6	99	17%
Complement and coagulation cascades	5	20.4	79	25%
Fluid shear stress and atherosclerosis	5	21.2	139	15%
Malaria	5	13	49	26%

### Protein-Protein interaction network

The investigation of critical genes and functions in lung cancer was also conducted through network analyses. The DEGs identified in the microarray analysis were used to build a PPI network. The APID database was used to screen protein interactions, which were then visualized as a PPI network using Cytoscape (figure 1). A total of 49 different DEG upregulated, and 164 downregulated genes were used to construct the two PPI networks. The proteins with the highest centrality in the upregulated PPI network were the Epidermal growth factor receptor (EGFR), DNA Repair Associated (BRCA1), Cadherin 1 (CDH1), Exportin 1 (XPO1), minimum centrality score 0.05.

The proteins with highest centrality in the downregulated PPI network (figure 2) were the Jun Proto-Oncogene, AP-1 Transcription Factor Subunit (JUN) e Transcription Factor (MYC), minimum centrality score 0.05.

We used the Golorize to classified the genes in the GO categories. The classification of the upregulated genes in the distinct GO biological process is related to the cell division and positive regulation of apoptotic process. It is noteworthy that EGFR and XPO1 genes, two of the most central in PPI network, are not involved in these biological processes. The downregulated genes are classified as distinct GO biological process categories with the highest number of genes related to the process highlighted are signal transduction, cell adhesion, immune response, innate immune response, apoptotic, inflammatory response, negative regulation of cell proliferation, positive regulation of cell proliferation. We believe that, together these downregulated biological processes could produce an aggressive tumor because of the reduced activity of the components of the immune system, cell adhesion, and apoptosis, that would result in the absence of immune reaction against the tumor cells, rapidly spreading of the cells creating metastasis, and unresponsive cell to apoptosis signals respectively.

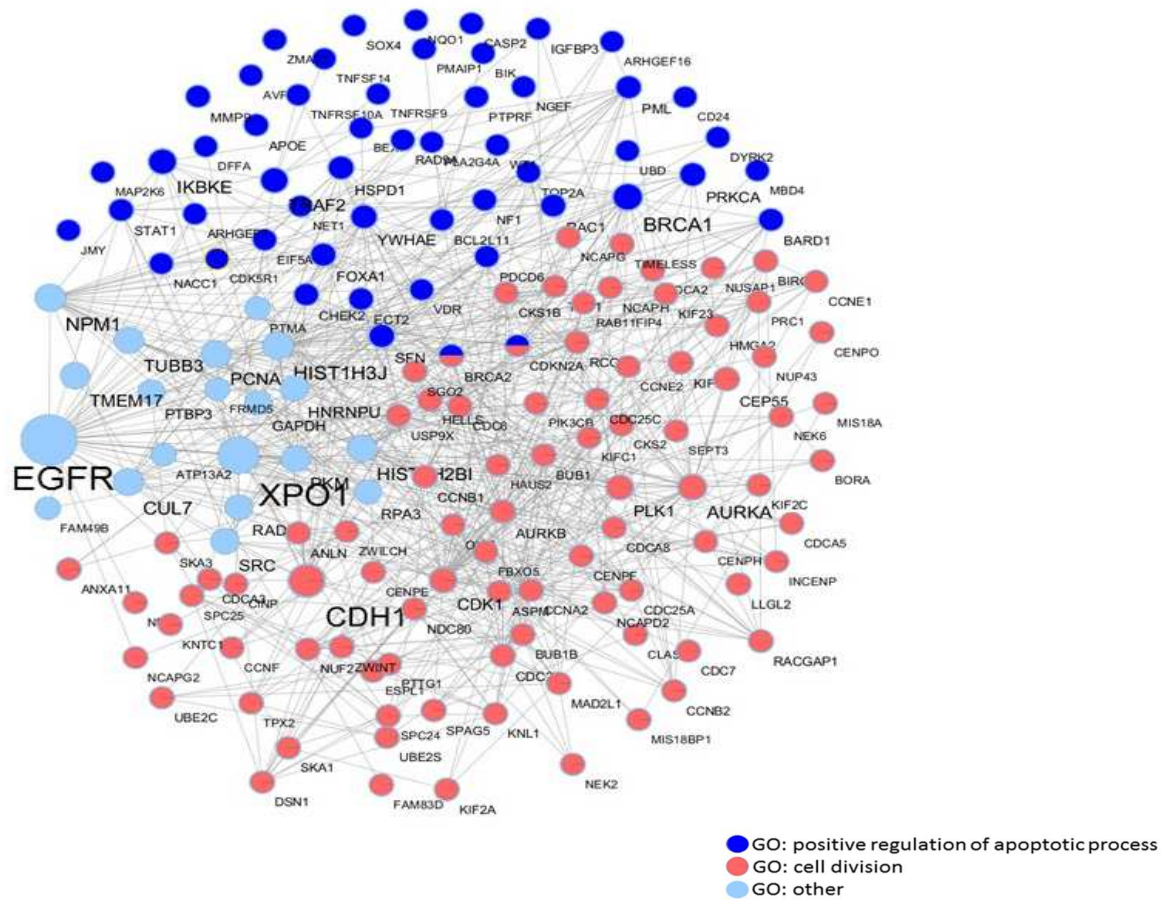


Figure 1. Protein-protein interaction of upregulated differentially expressed genes. The node size is proportional to the centrality of the protein. The genes with the highest centrality are EGFR, XPO1, and CDH1 (minimum centrality score  $\geq 0.05$ ). The GO process highlighted by Golorize are cell division (GO:0051301) (red nodes), and positive regulation of apoptotic process (GO:0043065) (blue nodes).



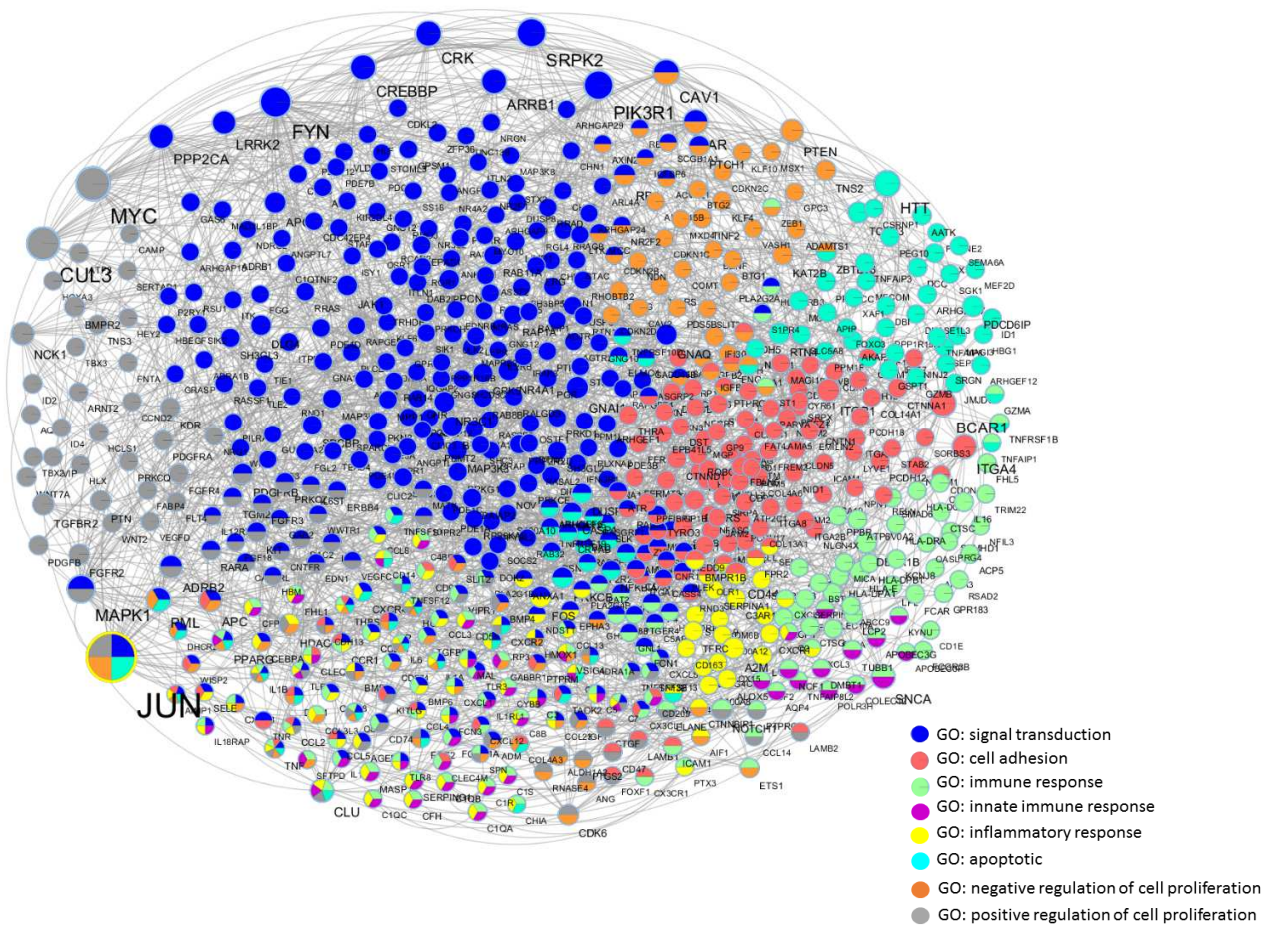


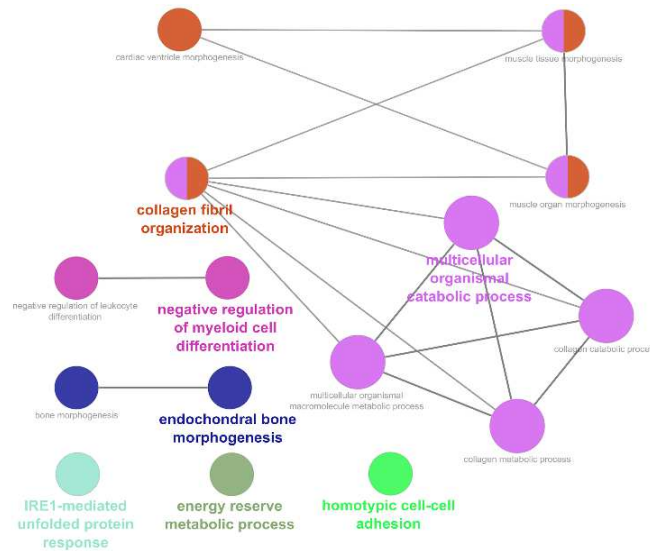
Figure 2. Protein-protein interaction of downregulated differentially expressed genes. The node size is proportional to the centrality of the protein. The genes with the highest centrality are JUN, MYC. (minimum centrality score  $\geq 0.05$ ). The GO process highlighted by Golorize are signal transduction (GO:0007165) (royal blue nodes), cell adhesion (GO:0007155) (red nodes), immune response (GO:0006955) (green nodes), innate immune response (GO:0045087) (pink nodes), apoptotic (GO:0006915) (baby blue nodes), inflammatory response (GO:0006954) (yellow), negative regulation of cell proliferation (GO:0008285) (orange), positive regulation of cell proliferation (GO:0008284) (gray).

### Gene ontology classification biological process

The investigation of functions activated in lung cancer is also complemented by ClueGO to retrieve the connection among these biological processes. The GS was used to access and compare the biological process in tissue lung cancer. In the upregulated genes, the ClueGO pointed the process ‘multicellular organismal catabolic process’, and ‘collagen fibril

organization' as the GO with more categories activated representing 38.89 and 22.22% respectively (figure 3).

(a)



(b)

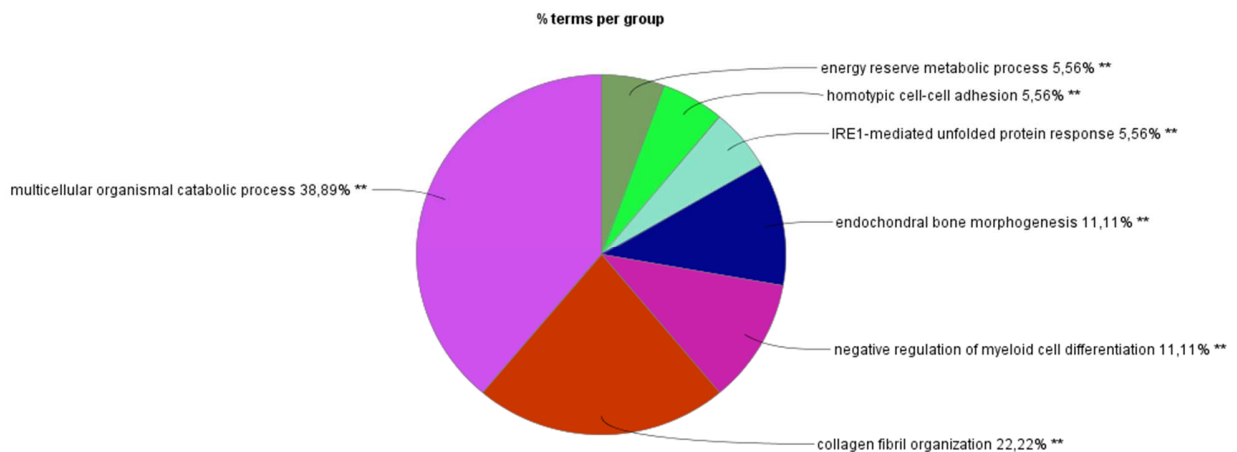
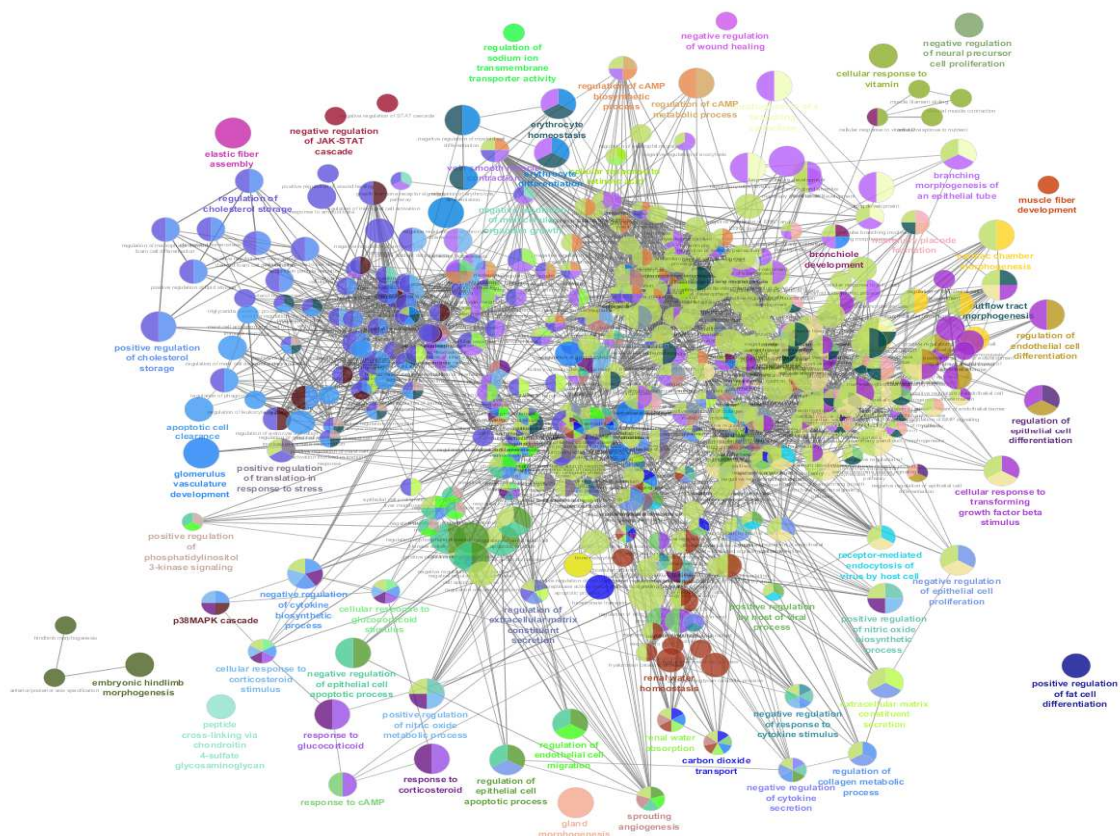


Figure 3. The GO biological processes network upregulated. The network was retrieved by ClueGO using the upregulated DEG from the six lung cancer. The panel (a) is the ClueGO network, (b) is the percentage of biological processes in each category. Overview chart with functional groups including specific terms for upregulated genes. The node size represents the

term enrichment significance. The color gradient shows the gene proportion of each cluster associated with the GO biological process.

The GO network for downregulated genes highlighted several terms in lung cancer (Figure 4a). These terms indicate processes turned off during lung cancer. In addition, one term was most represented in the analysis (Figure 4b) 'transforming growth factor receptor beta signaling pathway' with 25.6% of GO terms. This process presented downregulated genes, that indicates no pathway functioning or activity decrease. This can lead to defective signaling inside the cell.

(a)





(b)

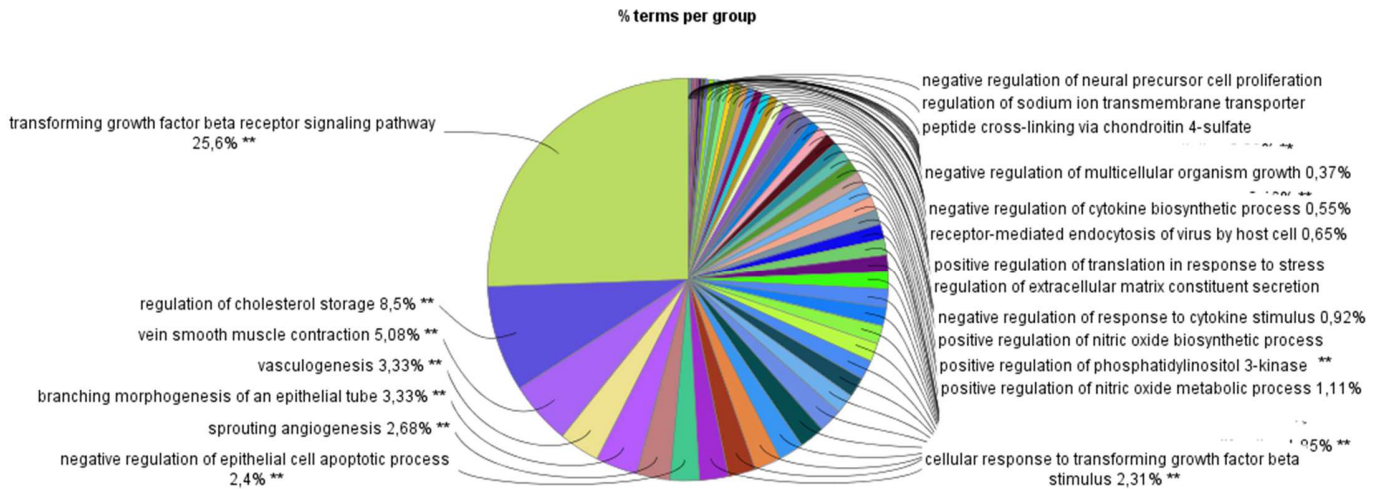


Figure 4. The GO biological processes network downregulated. The network was retrieved by ClueGO using the downregulated DEG from the six lung cancer. The panel (a) is the ClueGO network, (b) the percentage of biological processes in each category. Overview chart with functional groups including specific terms for downregulated genes. The node size represents the term enrichment significance. The color gradient shows the gene proportion of each cluster associated with the GO biological process.

## Clusterization heatmap and PCA

To evaluate the cluster of the samples using the gene expression signature we grouped the samples using hierarchical clustering and PCA. The hierarchical split the samples in two main groups, clustering the normal samples in one cluster together with eight tumor samples, and another cluster with tumor plus two healthy samples (figure 5/ table 7). These results give a true positive rate (TPR) of 0,97, and a true negative rate (TNR) of 0,99 (table 8).

The PCA analysis also separated well the healthy and tumor samples in two groups along the first principle component (figure 6). Although the samples with the same label are closer, there are some samples belonging to the healthy group mixed the tumor samples, and vice-versa.

Taken together these results showed that the expression values from the signature genes could group the healthy and tumor samples, indicating the presence of a distinct biological function in lung tumor cells.)

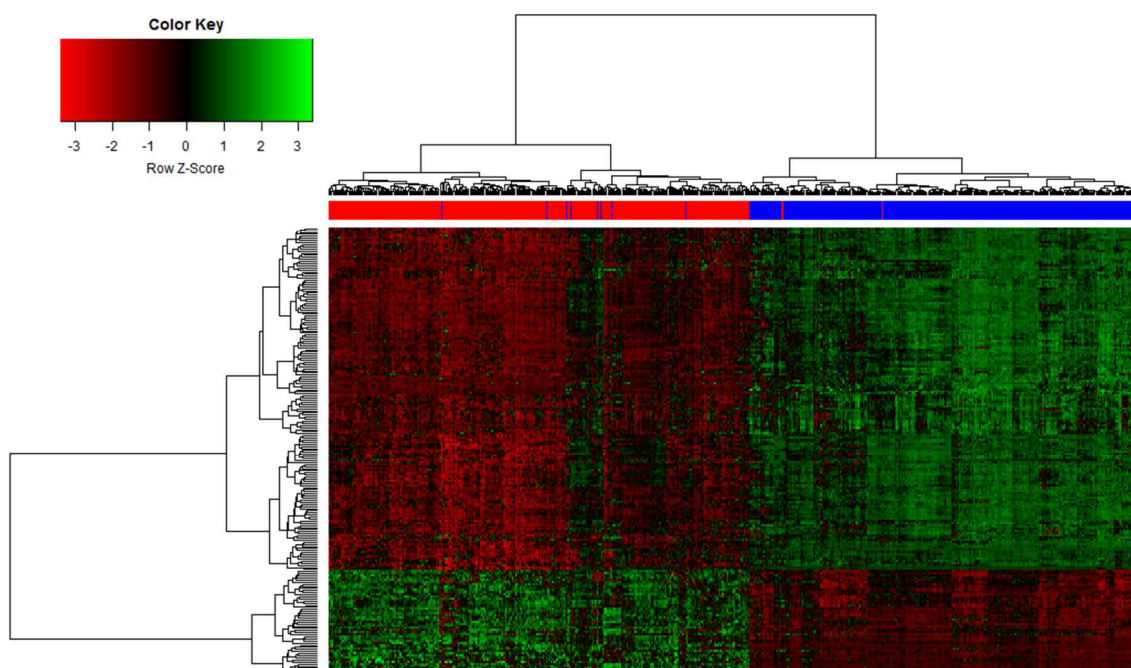


Figure 5. Heatmap and hierarchical clustering of samples using gene expression values. The heatmap was constructed using expression values of 49 genes upregulated and 164 downregulated. The red color represents low expression and green color high expression. The label bar represents the sample group (healthy or tumor), blue label represent healthy samples, while red labels represent lung tumor samples.

Table 7. Hierarchical classification

	Healthy cluster	Tumor cluster	Total
Healthy samples	328	8	336
Tumor samples	2	346	348

Table 8. Binary classification

	Predictive values

<b>Sensitivity</b>	99,4 (346/(346+2))
<b>Specificity</b>	97,6 (328/(328+8))

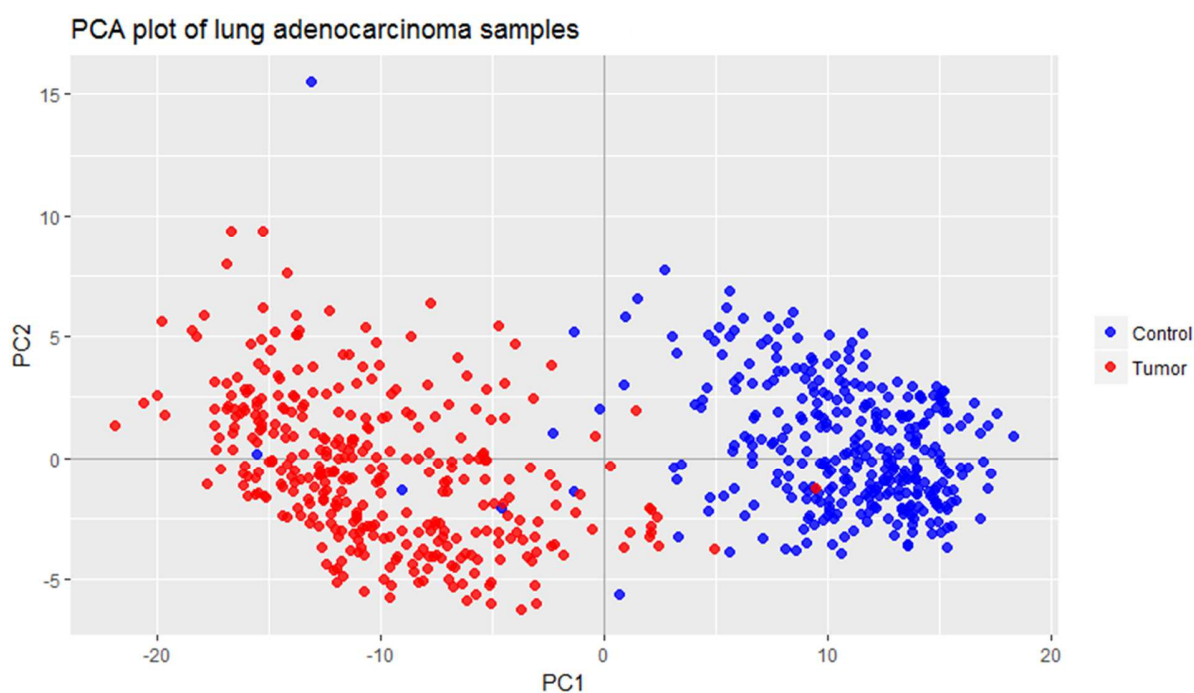


Figure 6. Principal component analysis of healthy and tumor samples. The samples grouped well in healthy and tumor samples in the first component. Nevertheless, there are some samples distant from their group, indicating some samples with different expression values in these signatures list. The expression values of the 213 GS were used in the PCA to investigate the sample grouping. The blue dots represent healthy samples, while red dots represent lung tumor samples.

## Immunohistochemistry

The expression of the GS was investigated in the protein level using information from The Protein Atlas to validate the results. We investigated protein expression of the DEG using

immunohistochemical images from lung cancer and the protein expression information classified as high, medium, low, or not detected (figura 8). We considered the genes with protein expression high+medium more than 50% as validated. The investigation showed that 13 upregulated genes present expression classified as high or medium in more than 50% of the immunohistochemical samples (figure 7).

These genes were investigated in the Malacards to check association with lung or cancer diseases. Among the upregulated, the genes TOP2A, NME1, CEACAM5 presented an association with lung cancer and the highest scores in Malacards, 5.88, 4.93, 3.15 respectively. The genes AGR2, CDC20, CDH3, HMGB3, PAICS, and SOX4 are associated with lung cancer, but the Malacards presented a low score to these genes, lower than 1. Although the gene CLDN3 is not associated with lung cancer, is associated with other types of cancer with a high score (> 4.8), such as ovarian, gastric, and endometrial cancer. The genes DNAJC12, GOLM1, and PDIA4, are not associated with lung cancer nor present a high score to any cancer. The genes AGR2, CDC20, CEACAM5, NME1, and TOP2A are involved in mutagenesis and cell/DNA replication.

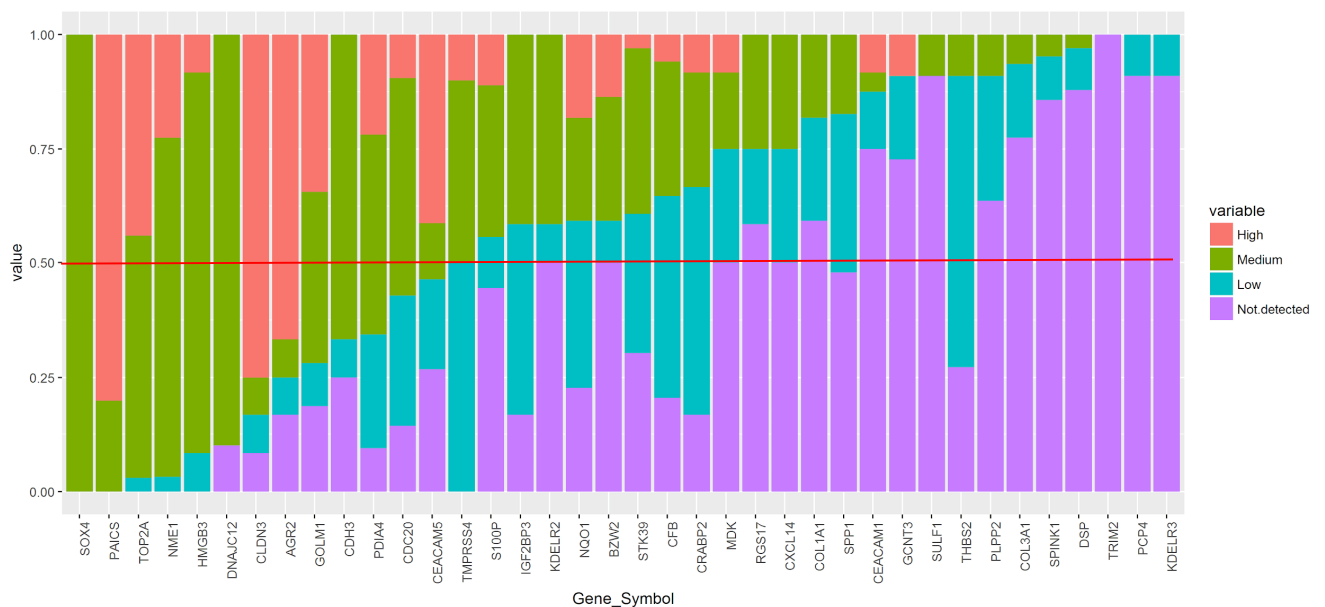


Figure 7. Level of protein expression of 38 upregulated genes, of total 49 genes. Samples with high gene expression (pink), samples with medium gene expression (green), samples with low

expression (blue) and samples with not detected expression (purple). Data credit: Human Protein Atlas. (High + medium > 50%).

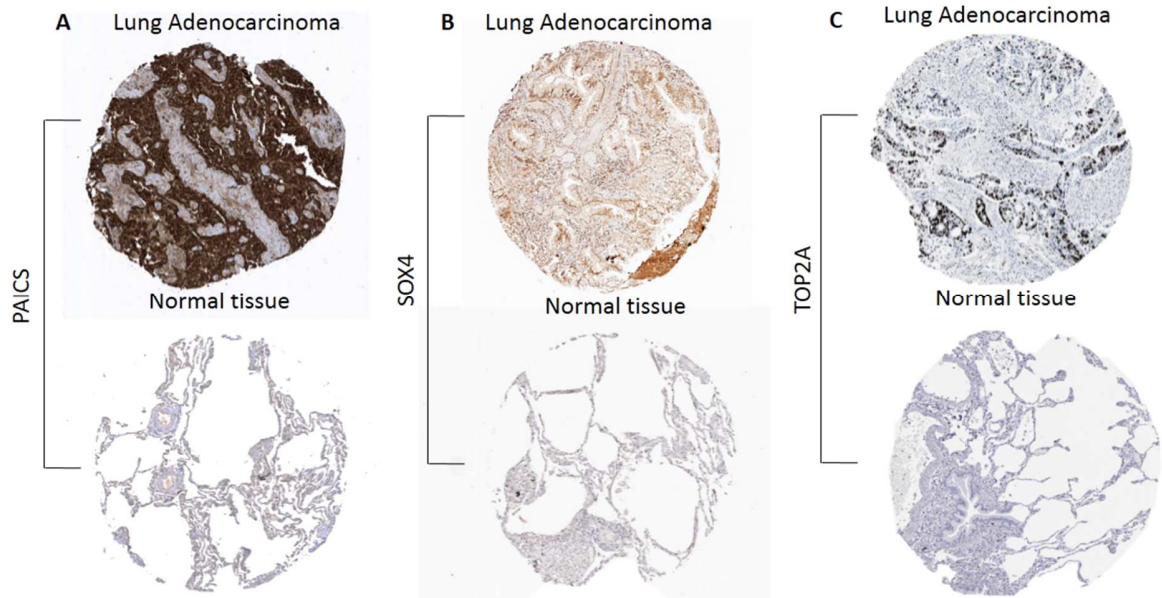


Figure 8. Immunohistochemical staining of PAICS (A), SOX4 (B) and TOP2A (C), shows the differential expression pattern, both high and normal tissue in samples from lung adenocarcinoma patients female. Image credit: Human Protein Atlas.

The list of 164 downregulated genes with protein expression 133 genes downregulated confirmed as validated by Protein Atlas (figure 9). Investigated in the Malacard has the SCGB1A1 associated to lung cancer with a score of 5.1, the highest among the downregulated. The downregulated gene list also has genes associated with lung cancer, but with a score lower than 1.0, these genes are FAM107A, ADAMTS8, LIMCH1, TMEM100, and ABCA8. Although the gene SFTPC and ALOX5 are not associated with lung cancer, they are associated with lung disease and asthma with a high score of 4.5 and 7.0, respectively. The genes CYP4B1, LYVE1, ZBTB16, HYAL2 are not associated with lung cancer but are associated with other cancer with high scores of 4.7 and 7.0, such as bladder cancer, prostate cancer, leukemia, and bronchiolo-alveolar adenocarcinoma. Taken together, these data indicate that the validated genes are involved in cell cycle and cellular replication in lung cancer.



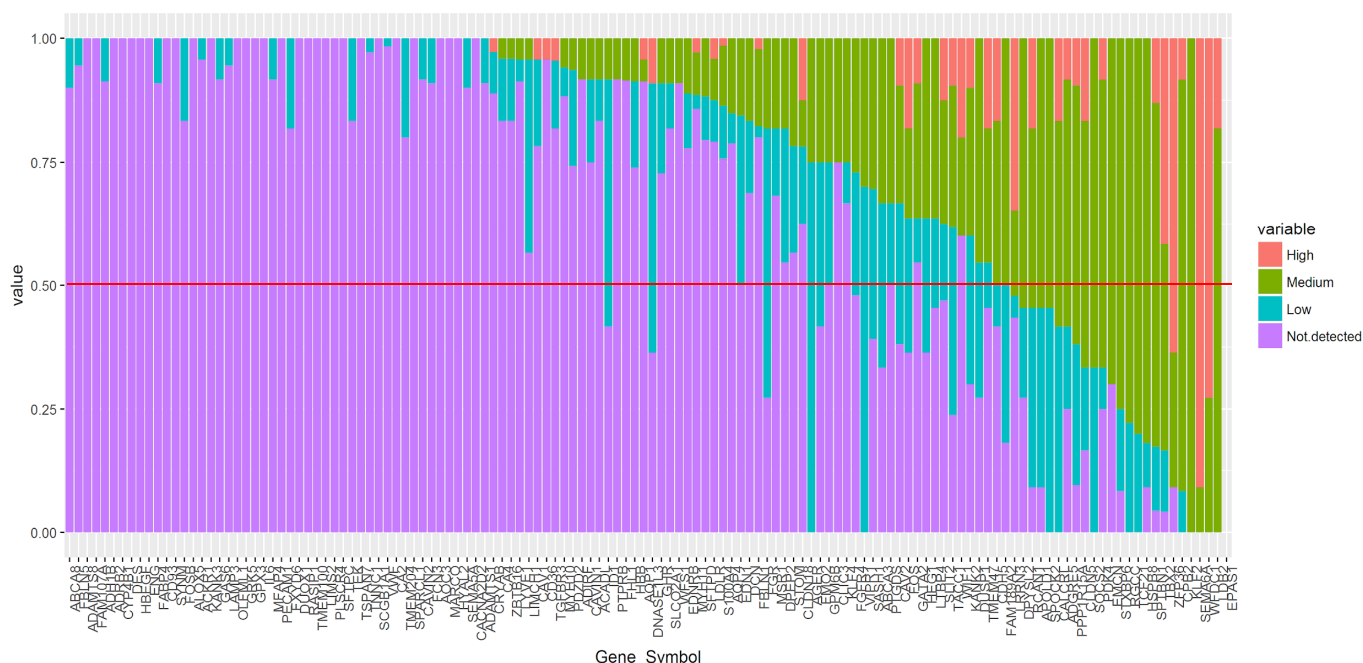


Figure 9. Level of protein expression of 133 downexpressed genes, of total 164 genes. Samples with high gene expression (pink), samples with medium gene expression (green), samples with low expression (blue) and samples with not detected expression (purple). Data credit: Human Protein Atlas. (Low + Not detected > 50%).

## Discussion

The complexity of NSCLC allows the disease to be of a heterogeneous nature, affecting patients differently. Patients in similar clinical and the pathological state can trigger the re-establishment of health or the onset of the disease. A versatile investigation of lung tumor tissue through molecular methods adds value to characterization because only clinical and pathological investigation does not allow a reliable diagnostic. Including the analysis of gene expression of tumor tissues is useful for greater precision of the patient's prognosis. In this study, we analyzed the microarray gene expression data from pulmonary adenocarcinoma data from six studies and combined the expression of available proteins using immunohistochemical data. Thanks to this

meta-analysis approach, we identified 213 differentially expressed genes at the mRNA level shared by the six studies, 49 up and 164 downregulated, respectively.

The use of molecular data from expression and mutations associated with lung cancer allowed the identification of the set of genes and biological processes altered in the pathology. The investigation mutations present in lung cancer samples indicated the involvement of TP53, KRAS, EGFR, RB1, and PIK3CA (12). The gene EGFR is a central point in this gene list because it is the receptor of the ligand stimuli of replication. The genes KRAS, PI3K, and RB1, are genes downstream in the EGFR signal transduction. These set of genes have an essential role in the origin and development of cancer, altering cell survival, proliferation, and differentiation. The PI3K stimulates the cell survival while KRAS stimulated cell proliferation (34). The gene expression studies indicate a set of genes altered during lung cancer development. The biological processes in which these genes are involved include mainly mitosis and cell division, (38, 39, 53, 65), cytoskeletal organization (39, 65), invasion (38, 53). These altered function in lung cancer contribute to rapid growth and metastasis.

The current study points important genes with altered expression in NSCLC using 684 samples (345 healthy, 339 tumor) from six studies. Among the differentially expressed genes identified, SOX4 is considered a transcription factor, in which its expression is involved in embryonic development in the heart, central nervous system, lung and thymus (9). As found in the analysis of the current study, the SOX4 gene is considered upregulated in cancer and forms part of a generic gene expression signature of cancer (60). The lung cancer type directives are related to different patterns of gene expression concerning the SOX gene. In the case of NSCLC, and more specifically in adenocarcinoma, increased levels of SOX4 are identified (9). The pathological degree of patients with the SOX4 gene with high expression is associated with a higher pathological grade, metastasis, tumor proliferation in progression and poor prognosis, than patients with low expression of SOX4 (96). This gene in lung cancer has characteristics of apoptotic changes, cell proliferation, invasion and migration (96, 97). Moreover, the silencing of SOX4 was investigated in some studies, and it was possible to verify the decrease of tumor growth in mice (96).

The PAICS gene encodes a bifunctional enzyme a carboxylase and synthetase in purine biosynthesis. Besides that, PAICS gene is upregulated in lung adenocarcinomas (25). In cancer, rapid cell division favors the biosynthesis pathway of new purine nucleotides for the synthesis of

adenine and guanine (42). This gene is related to the progression of cancer, and this was verified in prostate cancer as well as lung cancer (11). The alteration of PAICS expression modulates proliferation activity as well as the activity of tissue invasion (25). The PAICS gene is currently considered an example of a prognostic biomarker for aggressive lung adenocarcinoma (25).

Our data highlighted the gene TOP2A as upregulated; this information is consistent in several types of human tumors, in which topoisomerases are overexpressed (29). In the event of tumor cell proliferation, the TOP2A gene has its contribution (57). Specifically, this gene makes changes in DNA, and these events are essential for transcription, replication, recombination (29). In several types of cancer, the unfavorable prognosis is related to changes in the TOP2A gene, as the amount of translated proteins, the number of copies, and the mRNA expression (48).

The NME1 gene is a member of the nucleoside diphosphate kinase family (69) and plays a role in metastasis inhibition (92). This gene encodes NDK A protein and was the first tumor metastasis suppressor gene discovered from the 13 genes identified (58, 92). To replicate the cells need nucleoside triphosphates and the NME1 gene encodes the protein responsible for producing these molecules (58). NME1 is related to the regulation of gene expression and can impact the metastatic potential in different types of cancers (51). The expression of tumor and metastatic suppressor genes has a central position in tumor development as experimental studies indicate by down-regulating expression of this kind of genes, including NME1 that promotes cancer progression (24). At the function level, the NME1 gene has a specific function in the NSCLC (69). Loss of NDK A protein expression in some tumors was observed, and after the restoration of NDK A production, metastasis to the lungs and other cancers was significantly decreased (90). However, the reduction of the NME1 gene seems to be associated with a poor prognosis in lung cancer. Thus, the downregulation of the gene in the process of tumor progression is indicated as an unfavorable outcome in NSCLCs, mainly in lung adenocarcinoma (24).

The HMGB3 gene has been reported as a DNA protein binding through one or more motifs (75). The expression of HMGB3 has been reported as a poor prognosis in patients with a variety of human cancers such as esophageal squamous cell carcinoma (ESCC), Human urinary bladder cancer (UBC) and NSCLC (20, 43). Tissues present both, RNA and HMGB3 protein expression significantly higher in tumor than in adjacent normal tissues (20). At function level, it is believed that HMGB3 proteins play a crucial role in DNA replication, nucleosome assembly, and transcription. (79). Specifically, these proteins have the role of distorting, doubling or

structurally modifying linear DNA in a curved conformation (2). This indicates the role of HMGB3 gene proteins in DNA replication, transcription, recombination, and repair (20). In the case of gene transcriptional regulation, studies report the role of HMGB3 protein interacting with transcription regulators (transcription factors) and chromatin (histones) (84). Expression of HMGB3 in NSCLC samples is high compared to normal tissues, demonstrating important role as a biomarker. Besides, the expression of HMGB3 correlates with tumor size, stage, and metastasis (75).

The protein expressed from the DNAJC12 gene is a member of the HSP40/DnaJ heat shock protein family and play a role in assembly, protein folding, and export (14, 27). The use of chaperones in different tissues was studied in order to act on the enrichment of other proteins in the transport and expression of genes (14). Evidence from gastric and colorectal cancers indicate that HSP40 and HSP70 expression in cancer tissues were overexpressed compared to normal tissues (27). Moreover, breast cancer stem cells (CSCs) has higher expression of the DNAJC12 gene in comparison to adherent breast cancer cells (63). Although these data indicate overexpression of DNAJC12 and HSP40/DnaJ heat shock protein family genes, little is known about the role of these proteins in cancers (27).

The CLDN3 gene encodes the claudin protein, which is required for cell-to-cell adhesion, regulating the entry and exit of solutes and water (44, 4). It has been observed that the expression of this protein plays a different role in tumor tissues (4). In many types of cancers, such as epithelial, colon, liver, and thyroid cancer, the modulation of this protein has already been reported (1). In NSCLC, has been reported to be upregulated (44, 4). The overexpression of claudin-3 and claudin-4 proteins were also shown in ovarian cells (44) and in colon cancer (15). In ovarian cancer, the overexpression of claudin-3 and claudin-4 led to cell invasion (1). Claudin-3 when downregulated increased metastasis in esophageal cancer (78). The expression of CLDN3 in lung adenocarcinoma predicts poor prognosis (94). Interestingly, the claudin-3 protein was induced to overexpress EGF, which is the EGFR ligand, an event mediated by the MEK/ERK and PI3K/Akt pathways (94). This divergence still has much to discuss; the most plausible explanation is the specific expression of claudin in different tissues which may lead to different functions (1). The expression of CLDN3 is reported to be upregulated in lung adenocarcinoma and play a role in migration (30). Consequentially, the expression of CLDN3 in lung cancer might contribute to the metastasis and poor prognostic.

The AGR2 gene expresses an endoplasmic reticulum chaperone that is classified as a proto-oncogene in breast adenocarcinoma, colon, and esophagus (52). Poor prognosis is reported in patients with NSCLC with high levels of AGR2 (91). The AGR2 in lung tumors is included in the positive regulation of the EGFR signaling pathway induced by the ligand-transforming growth factor (TGFA) (80). The lung tumorigenesis by the AGR2 gene is related to ligands (TGFA, BTC, and EGF) in mutant EGFR adenocarcinoma (80). In EGFR mutations in lung cancer, TGFA expression was associated with poor patient survival, as it induces the EGFR signaling pathway, indicating proliferation and survival of tumor cells (80, 52). In human lung cancer, the AGR2 gene can be used as a biomarker (52, 91).

The networks identified the interaction of central genes and biological processes expressed in NSCLC, and our results demonstrated the involvement of 52 genes related to the regulation of the apoptotic process and 89 genes related to cell division. The analysis of new signaling pathways allows us to learn more about pathogenesis, biological processes and the significant genes that are pathogenic for cancers. The biological process of apoptosis plays a vital role during the progression of lung cancer (46). In tumorigenesis, the inhibition of apoptosis is considered an essential step to cancer development, since it allows the survival of cells (56). In NSCLC, apoptosis and cell division are part of EGFR signaling, which can mediate signals for cancer cell proliferation, inhibition of apoptosis, invasion, metastasis and neovascularization (55). EGFR is characterized by several cellular processes such as signal transduction at the level of cell division, angiogenesis, and apoptosis (16). Apoptosis can be mediated by autophagy in NSCLC, by persistent activation, causing cell death, thus decreasing cancerous cells (46). The autophagy process is characterized by being a self-feeding process (46). Mechanism to degrade and recycle cellular components in autophagosomes, and later in the late endosome or lysosome may occur (46). When activated by oncogenes, it can prevent the degradation of proteins in tumor cells, favoring the progression of cancerous cells (46).

It is important to be aware of the role of exogenous cells in the production of cell-mediated proteins in the production of cell-mediated proteins (63). In patients with cancers, overexpressing exportin-1 performs some erroneous mechanisms such as suppressor protein delivery, cell signaling, inactivation of apoptosis and tumor growth (63). Studies have suggested the use of the XPO1 gene as a prognostic biomarker in patients with adenocarcinoma (62). In our study, EGFR and XPO1 has been shown to be an important gene between the centrality of the

network of protein interaction, but to demonstrate contradiction to the biological processes of apoptosis and cell division.

Another gene with important betweenness centrality in the network of protein interaction and that in our analysis has shown to participate in the biological process of apoptosis is the gene BRCA1. This gene is an important tumor suppressor because it has the characteristic of ADN repair (76). If mutations in this gene are diagnosed, the possibility of developing cancerous cells in the breast can be inferred (54). This mutation was also diagnosed in patients with NSCLC (10). Findings indicate a high expression of BRCA1 mRNA in patients with NSCLC (76, 83). The occurrence of the best response to the disease is related to the expression of this gene in the primary tumor (83). In addition to being a target gene for treatment in NSCLC (76).

The biological process related to cell division also appeared in our analysis. This may be significant, since uncontrolled cell division is considered characteristic of oncogenesis (47), and the genes pointed here could be investigated as a therapy target to improve treatment. The result is in the gene cadherin1 (CDH1), which is among the genes with expression and with important centrality in the network of protein interaction. The cadherin protein has the characteristic of connecting cells, being able to regulate the mobility and proliferation of epithelial cells. (95). In patients with NSCLC, the decreased expression of the CDH1 gene is related to tumor invasion and consequently poor prognostic (46). The CDH1 gene was analyzed on NSCLC and has been shown to be a low expression gene, which goes in contrast with our findings (95). Besides that, CDH1 gene expression is associated with EGFR signaling (47). The gene has the character of a tumor suppressor gene (89, 45). Our data support the hypothesis that increased gene expression of CDH1 may play a role in the pathogenesis of NSCLC in the attempt to control cell proliferation, leading to its accumulation in the cytoplasm and inability to exert its effects.

Studies have shown the classification of samples from healthy lung tissue and tumor tissue. Some of the approaches are the analysis of the behavior of these cells and the identification of signaling (38). Other studies have been based on the identification of the regulation of DNA methylation and deregulation of biological processes (65). Relevant use for precision medicine is to distinguish different NSCLC subtypes between themselves and the healthy tissue (23). Our data showed the separation of the samples in tumor and healthy samples based in the DEG, suggesting these DEG can be used in molecular diagnosis and help to improve inconclusive results.

Immunohistochemistry of the analyzed PAICS gene has been shown to be upregulated in lung cancer, and this result predicts poor prognostic (25). As with the CDC20 gene, immunohistochemistry also revealed poor prognosis of patients with overexpression of PAICS gene in NSCLC (37). In immunohistochemical analyzes performed for the TOP2A gene, the increased expression of this gene in NSCLC and its correlation with metastasis for the brain was also revealed (). The immunohistochemical results here validate the mRNA expression analysis and indicate a list of genes also expressed as the protein level in NSCLC.

In conclusion, our results provide clues about investigating genetic biomarkers and their importance in predicting lung cancer. If confirmed in assays, the results of differentially expressed genes found in this study may become important targets for the treatment of NSCLC, especially lung adenocarcinoma. Future validation of this research would be an essential step.

### **Funding**

This work was supported by PNPD/CAPES fellowship from UFBA/Biociências postgraduate program and FASPESB fellowship. The authors declare that the funding agency had no influence in data analyses, or journal selection.

### **Acknowledgements**

We thank all past and present members of the lab for their contributions to the scientific work leading to knowledge advancement on lung cancer and DEGs.

### **References**

1. Agarwal R, D'Souza T, Morin PJ. Claudin-3 and claudin-4 expression in ovarian epithelial cells enhances invasion and is associated with increased matrix metalloproteinase-2 activity. *Cancer Res.* 2005;65(16):7378–85.
2. Agresti A, Bianchi ME. HMGB proteins and gene expression. *Curr Opin Genet Dev.* 2003;13(2):170–8.
3. Ahrendt SA, Anthony Decker P, Alawi EA, Zhu YR, Sanchez-Cespedes M, Yang SC, et al. Cigarette smoking is strongly associated with mutation of the K-ras gene in patients with primary adenocarcinoma of the lung. *Cancer.* 2001;92(6):1525–30.
4. Akizuki R, Maruhashi R, Eguchi H, Kitabatake K, Tsukimoto M, Furuta T, et al. Decrease

- in paracellular permeability and chemosensitivity to doxorubicin by claudin-1 in spheroid culture models of human lung adenocarcinoma A549 cells. *Biochim Biophys Acta - Mol Cell Res* [Internet]. Elsevier; 2018;1865(5):769–80. Available from: <https://doi.org/10.1016/j.bbamcr.2018.03.001>
5. Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. APID interactomes: Providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res*. 2016;44(W1):W529–35.
  6. Bakör K, UcĖak R, TuncĖözgür B, Elbeyli L. Prognostic Factors and c-erbB-2 Expression in Non-Small-Cell Lung Carcinoma (c-erbB-2 in Non-Small Cell Lung Carcinoma). 2002;90(342).
  7. Beers MF, Mulugeta S. The biology of the ABCA3 lipid transporter in lung health and disease. *Cell Tissue Res*. 2017;367(3):481–93.
  8. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25(8):1091–3.
  9. Castillo SD, Matheu A, Mariani N, Carretero J, Lopez-Rios F, Lovell-Badge R, et al. Novel transcriptional targets of the SRY-HMG box transcription factor SOX4 link its expression to the development of small cell lung cancer. *Cancer Res*. 2012;72(1):176–86.
  10. Cedrés S, Felip E, Cruz C, Martinez de Castro A, Pardo N, Navarro A, et al. Activity of HSP90 Inhibitor in a Metastatic Lung Cancer Patient With a Germline BRCA1 Mutation. *JNCI J Natl Cancer Inst* [Internet]. 2018;110:914–7. Available from: <https://academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djy012/4911464>
  11. Chakravarthi BVSK, Goswami MT, Pathi SS, Dodson M, Chandrashekar DS, Agarwal S, et al. Expression and Role of PAICS, a De Novo Purine Biosynthetic Gene in Prostate Cancer. *Prostate*. 2017;77(1):10–21.
  12. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature* [Internet]. Nature Publishing Group; 2014;511(7511):543–50. Available from: <http://dx.doi.org/10.1038/nature13385>
  13. Cox AD, Fesik SW, Kimmelman AC, Luo J, Der CJ. Drugging the undruggable RAS: Mission Possible? *Nat Rev Drug Discov*. 2014;13(11):828–51.
  14. De Bessa SA, Salaorni S, Patr??o DFC, Neto MM, Brentani MM, Nagai MA. JDP1 (DNAJC12/Hsp40) expression in breast cancer and its association with estrogen receptor status. *Int J Mol Med*. 2006;17(2):363–7.
  15. de Souza WF, Fortunato-Miranda N, Robbs BK, de Araujo WM, de-Freitas-Junior JC, Bastos LG, et al. Claudin-3 Overexpression Increases the Malignant Potential of Colorectal Cancer Cells: Roles of ERK1/2 and PI3K-Akt as Modulators of EGFR signaling. *PLoS One*. 2013;8(9):1–13.
  16. Denisenko T V., Budkevich IN, Zhivotovsky B. Cell death-based treatment of lung adenocarcinoma article. *Cell Death Dis* [Internet]. Springer US; 2018;9(2). Available from: <http://dx.doi.org/10.1038/s41419-017-0063-y>
  17. Dennis, Glynn. J, Sherman B, Hosack D, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* [Internet]. 2003;4(5):P3. Available from: <http://genomebiology.com/2003/4/5/P3>
  18. Faruki H, Mayhew GM, Fan C, Wilkerson MD, Parker S, Kam-Morgan L, et al. Validation of the lung subtyping panel in multiple fresh-frozen and formalin-fixed,



- paraffin-embedded lung tumor gene expression data sets. *Arch Pathol Lab Med*. 2016;140(6):536–42.
19. Faruki H, Mayhew GM, Serody JS, Hayes DN, Perou CM, Lai-Goldman M. Lung Adenocarcinoma and Squamous Cell Carcinoma Gene Expression Subtypes Demonstrate Significant Differences in Tumor Immune Landscape. *J Thorac Oncol* [Internet]. Elsevier Inc; 2017;12(6):943–53. Available from: <http://dx.doi.org/10.1016/j.jtho.2017.03.010>
  20. Gao J, Zou Z, Zhang H, Lin Z, Zhang Y, Luo X, et al. Increased expression of HMGB3: a novel independent prognostic marker of worse outcome in patients with esophageal squamous cell carcinoma. *Int J Clin Exp Pathol* [Internet]. 2015;8(1):345–52. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed13&NEWS=N&AN=25755721>
  21. Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B, et al. GOLORize: A Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*. 2007;23(3):394–6.
  22. Gentleman AR, Functions D, Rgraphviz S. Package ‘Rbioinf.’ 2018;
  23. Girard L, Rodriguez-Canales J, Behrens C, Thompson DM, Botros IW, Tang H, et al. An expression signature as an aid to the histologic classification of non-small cell lung cancer. *Clin Cancer Res*. 2016;22(19):4880–9.
  24. Goncharuk VN, Del-Rosario A, Kren L, Anwar S, Sheehan CE, Carlson JA, et al. Co-Downregulation of PTEN, KAI-1, and nm23-H1 Tumor/Metastasis Suppressor Proteins in Non-Small Cell Lung Cancer. *Ann Diagn Pathol*. 2004;8(1):6–16.
  25. Goswami MT, Chen G, Chakravarthi BVSK, Pathi SS, Anand SK, Carskadon SL, et al. Role and regulation of coordinately expressed de novo purine biosynthetic enzymes PPAT and PAICS in lung cancer. *Oncotarget* [Internet]. 2015;6(27):23445–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26140362><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4695129>
  26. Gupta A, Saltarski JM, White MA, Scaglioni PP, Gerber DE. Therapeutic Targeting of Nuclear Export Inhibition in Lung Cancer. *J Thorac Oncol* [Internet]. Elsevier Inc; 2017;12(9):1446–50. Available from: <http://dx.doi.org/10.1016/j.jtho.2017.06.013>
  27. He HL, Lee YE, Chen HP, Hsing CH, Chang IW, Shiue YL, et al. Overexpression of DNAJC12 predicts poor response to neoadjuvant concurrent chemoradiotherapy in patients with rectal cancer. *Exp Mol Pathol* [Internet]. Elsevier Inc.; 2015;98(3):338–45. Available from: <http://dx.doi.org/10.1016/j.yexmp.2015.03.029>
  28. Henderson LB, Melton K, Wert S, Couriel J, Bush A, Ashworth M, et al. Large ABCA3 and SFTPC deletions resulting in lung disease. *Ann Am Thorac Soc*. 2013;10(6):602–7.
  29. Hou GX, Liu P, Yang J, Wen S. Mining expression and prognosis of topoisomerase isoforms in non-small-cell lung cancer by using Oncomine and Kaplan-Meier plotter. *PLoS One*. 2017;12(3):1–16.
  30. Hsu Y-L, Hung J-Y, Lee Y-L, Chen F-W, Chang K-F, Chang W-A, et al. Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. *Oncotarget* [Internet]. 2017;8(62):104831–54. Available from: <http://www.oncotarget.com/fulltext/21022>
  31. Huang H, Liu J, Meng Q, Niu G. Multidrug resistance protein and topoisomerase 2 alpha expression in non-small cell lung cancer are related with brain metastasis postoperatively. *Int J Clin Exp Pathol*. 2015;8(9):11537–42.

32. Ji H, Ramsey MR, Hayes DN, Fan C, McNamara K, Kozlowski P, et al. LKB1 modulates lung cancer differentiation and metastasis. *Nature*. 2007;448(7155):807–10.
33. Ji H, Wang Z, Perera SA, Li D, Liang MC, Zaghlul S, et al. Mutations in BRAF and KRAS converge on activation of the mitogen-activated protein kinase pathway in lung cancer mouse models. *Cancer Res*. 2007;67(10):4933–9.
34. Jiang W, Cai G, Hu PC, Wang Y. Personalized medicine in non-small cell lung cancer: a review from a pharmacogenomics perspective. *Acta Pharm Sin B* [Internet]. Elsevier B.V.; 2018;8(4):530–8. Available from: <http://dx.doi.org/10.1016/j.apsb.2018.04.005>
35. Kadara H, Lacroix L, Behrens C, Solis L, Gu X, Lee JJ, et al. Identification of gene signatures and molecular markers for human lung cancer prognosis using an in vitro lung carcinogenesis system. *Cancer Prev Res*. 2009;2(8):702–11.
36. Kadara H, Behrens C, Yuan P, Solis L, Liu D, Gu X, et al. A five-gene and corresponding protein signature for stage-I lung adenocarcinoma prognosis. *Clin Cancer Res*. 2011;17(6):1490–501.
37. Kato T, Daigo Y, Aragaki M, Ishikawa K, Sato M, Kaji M. Overexpression of CDC20 predicts poor prognosis in primary non-small cell lung cancer patients. *J Surg Oncol*. 2012;106(4):423–30.
38. Kim IJ, Quigley D, To MD, Pham P, Lin K, Jo B, et al. Rewiring of human lung cell lineage and mitotic networks in lung adenocarcinomas. *Nat Commun* [Internet]. Nature Publishing Group; 2013;4:1701–11. Available from: <http://dx.doi.org/10.1038/ncomms2660>
39. Landi D, Gemignani F, Barale R, Landi S. A Catalog of Polymorphisms Falling in MicroRNA-Binding Regions of Cancer Genes. *DNA Cell Biol* [Internet]. Mary Ann Liebert, Inc., publishers; 2007 Oct 17;27(1):35–43. Available from: <https://doi.org/10.1089/dna.2007.0650>
40. Lee MP, Reeves C, Schmitt A, Su K, Connors TD, Hu RJ, et al. Somatic mutation of TSSC5, a novel imprinted gene from human chromosome 11p15.5. *Cancer Res*. 1998;58(18):4155–9.
41. Lee SJ, Jang BC, Lee SW, Yang Y II, Suh S II, Park YM, et al. Interferon regulatory factor-1 is prerequisite to the constitutive expression and IFN- $\gamma$ -induced upregulation of B7-H1 (CD274). *FEBS Lett*. 2006;580(3):755–62.
42. Li SX, Tong YP, Xie XC, Wang QH, Zhou HN, Han Y, et al. Octameric Structure of the Human Bifunctional Enzyme PAICS in Purine Biosynthesis. *J Mol Biol*. 2007;366(5):1603–14.
43. Li M, Cai Y, Zhao H, Xu Z, Sun Q, Luo M, et al. Overexpression of HMGB3 protein promotes cell proliferation, migration and is associated with poor prognosis in urinary bladder cancer patients. *Tumor Biol*. 2015;36(6):4785–92.
44. Li W, Zhang X, Li Z, Jiang F, Zhao H, Wei BO. Identification of genes associated with matrix metalloproteinases in invasive lung adenocarcinoma. *Oncol Lett*. 2018;123–30.
45. Ling ZQ, Li P, Ge MH, Zhao X, Hu FJ, Fang XH, et al. Hypermethylation-modulated down-regulation of CDH1 expression contributes to the progression of esophageal cancer. *Int J Mol Med*. 2011;27(5):625–35.
46. Liu G, Pei F, Yang F, Li L, Amin AD, Liu S, et al. Role of autophagy and apoptosis in non-small-cell lung cancer. *Int J Mol Sci*. 2017;18(2).
47. Liu X, Su L, Liu X. Loss of CDH1 up-regulates epidermal growth factor receptor via phosphorylation of YBX1 in non-small cell lung cancer cells. *FEBS Lett* [Internet].

- Federation of European Biochemical Societies; 2013;587(24):3995–4000. Available from: <http://dx.doi.org/10.1016/j.febslet.2013.10.036>
48. Lu B, Zhang H, Zhang T, Cai Y, Hu Y, Zheng H, et al. Topoisomerase I expression is associated with prognosis in postoperative non-small cell lung cancer patients. *Thorac Cancer*. 2016;7(4):486–94.
  49. Masago K, Fujita S, Irisa K, Kim YH, Ichikawa M, Mio T, et al. Good clinical response to gefitinib in a non-small cell lung cancer patient harboring a rare somatic epidermal growth factor gene point mutation; codon 768 AGC > ATC in exon 20 (S768I). *Jpn J Clin Oncol*. 2010;40(11):1105–9.
  50. Matsumoto S, Iwakawa R, Takahashi K, Kohno T, Nakanishi Y, Matsuno Y, et al. Prevalence and specificity of LKB1 genetic alterations in lung cancers. *Oncogene*. 2007;26(40):5911–8.
  51. McCorkle JR, Leonard MK, Kraner SD, Blalock EM, Ma D, Zimmer SG, et al. The metastasis suppressor NME1 regulates expression of genes linked to metastasis and patient outcome in melanoma and breast carcinoma. *Cancer Genomics and Proteomics*. 2014;11(4):175–94.
  52. Milewski D, Balli D, Ustiyev V, Le T, Dienemann H, Warth A, et al. FOXM1 activates AGR2 and causes progression of lung adenomas into invasive mucinous adenocarcinomas. *PLoS Genet*. 2017;13(12):1–21.
  53. Mitchell KA, Zingone A, Toulabi L, Boeckelman J, Ryan BM. Comparative transcriptome profiling reveals coding and noncoding RNA differences in NSCLC from African Americans and European Americans. *Clin Cancer Res*. 2017;23(23):7412–25.
  54. Nielsen FC, Van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: New genes in confined pathways. *Nat Rev Cancer* [Internet]. Nature Publishing Group; 2016;16(9):599–612. Available from: <http://dx.doi.org/10.1038/nrc.2016.72>
  55. Othman N, Nagoor NH. The role of microRNAs in the regulation of apoptosis in lung cancer and its application in cancer treatment. *Biomed Res Int*. Hindawi Publishing Corporation; 2014;2014.
  56. Pore MM, Hiltermann TJN, Kruyt FAE. Targeting apoptosis pathways in lung cancer. *Cancer Lett* [Internet]. Elsevier Ireland Ltd; 2013;332(2):359–68. Available from: <http://dx.doi.org/10.1016/j.canlet.2010.09.012>
  57. Proszek J, Roy A, Jakobsen AK, Frøhlich R, Knudsen BR, Stougaard M. Topoisomerase I as a biomarker: Detection of activity at the single molecule level. *Sensors (Switzerland)*. 2014;14(1):1195–207.
  58. Qu S, Long J, Cai Q, Shu X-O, Cai H, Gao Y-T, et al. Genetic polymorphisms of metastasis suppressor gene NME1 and breast cancer survival. *Clin Cancer Res* [Internet]. 2008;14(15):4787–93. Available from: <http://clincancerres.aacrjournals.org/content/14/15/4787.full>
  59. Rappaport N, Twik M, Plaschkes I, Nudel R, Stein TI, Levitt J, et al. MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res*. 2017;45(D1):D877–87.
  60. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci* [Internet]. 2004;101(25):9309–14. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.0401994101>

61. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
62. Robles AI, Arai E, Mathé EA, Okayama H, Schetter AJ, Brown D, et al. An integrated prognostic classifier for Stage I lung adenocarcinoma based on mRNA, microRNA, and DNA methylation biomarkers. *J Thorac Oncol* [Internet]. International Association for the Study of Lung Cancer; 2015;10(7):1037–48. Available from: <http://dx.doi.org/10.1097/JTO.0000000000000560>
63. Ross CKC, Martinez SR, Woods- L, Durán AM, Roy S, Basu A, et al. RNA sequencing reveals upregulation of a transcriptomic program associated with stemness in metastatic prostate cancer cells selected for taxane resistance. 2018;9(54):30363–84.
64. Sean D, Meltzer PS. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007;23(14):1846–7.
65. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res.* 2012;22(7):1197–211.
66. Selvaggi G, Scagliotti G V., Torri V, Novello S, Leonardo E, Cappia S, et al. HER-2/neu overexpression in patients with radically resected nonsmall cell lung carcinoma: Impact on long-term survival. *Cancer.* 2002;94(10):2669–74.
67. Shabalina AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics.* 2008;24(9):1154–60.
68. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res* [Internet]. 2003;(22):426. Available from: <http://ci.nii.ac.jp/naid/110001910481/>
69. Shi X, Jin H, Peng M, Li B, She M, Zhu T, et al. Association between NME1 polymorphisms and cancer susceptibility: A meta-analysis based on 1644 cases and 2038 controls. *Pathol Res Pract* [Internet]. Elsevier; 2018;214(4):467–74. Available from: <https://doi.org/10.1016/j.prp.2018.02.020>
70. Shivapurkar N, Reddy J, Matta H, Sathyanarayana UG, Huang CX, Toyooka S, et al. Loss of expression of death-inducing signaling complex (DISC) components in lung cancer cell lines and the influence of MYC amplification. *Oncogene.* 2002;21(55):8510–4.
71. Siegel R, Miller K, Jemal A. Cancer statistics , 2015 . *CA Cancer J Clin* [Internet]. 2015;65(1):29. Available from: <http://onlinelibrary.wiley.com/doi/10.3322/caac.21254/pdf>
72. Singhal S, Amin KM, Kruklytis R, DeLong P, Friscia ME, Litzky L a, et al. Alterations in cell cycle genes in early stage lung adenocarcinoma identified by expression profiling. *Cancer Biol Ther* [Internet]. 2003;2(3):291–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12878869>
73. Sîrbu A, Ruskin HJ, Crane M. Cross-platform microarray data normalisation for regulatory network inference. *PLoS One.* 2010;5(11).
74. Somaschini M, Presi S, Ferrari M, Vergani B, Carrera P. Genetic surfactant dysfunction in newborn infants and children with acute and chronic lung disease. *J Pediatr Neonatal Individ Med* [Internet]. 2017;6(1):1–8. Available from: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019008235&doi=10.7363%2F060134&partnerID=40&md5=d08ce118f4e883507f08120c3e911cf3>

75. Song N, Liu B, Wu J-L, Zhang R-F, Duan L, He W-S, et al. Prognostic value of HMGB3 expression in patients with non-small cell lung cancer. *Tumor Biol* [Internet]. 2013;34(5):2599–603. Available from: <http://link.springer.com/10.1007/s13277-013-0807-y>
76. Sun SY, Yue P, Zhou JY, Wang Y, Choi Kim HR, Lotan R, et al. Overexpression of Bcl2 blocks TNF-related apoptosis-inducing ligand (TRAIL)-induced apoptosis in human lung cancer cells. *Biochem Biophys Res Commun*. 2001;280(3):788–97.
77. Sun X, Cui F, Yin H, Wu D, Wang N, Yuan M, et al. Association between EGFR mutation and expression of BRCA1 and RAP80 in non-small cell lung cancer. *Oncol Lett* [Internet]. 2018;2201–6. Available from: <http://www.spandidos-publications.com/10.3892/ol.2018.8938>
78. Takala H, Saarnio J, Wiik H, Soini Y. Claudins 1, 3, 4, 5 and 7 in esophageal cancer: loss of claudin 3 and 4 expression is associated with metastatic behavior. *APMIS* [Internet]. 2007;115(7):838–47. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17614851>
79. Tang HR, Luo XQ, Xu G, Wang Y, Feng ZJ, Xu H, et al. High mobility group-box 3 overexpression is associated with poor prognosis of resected gastric adenocarcinoma. *World J Gastroenterol*. 2012;18(48):7319–26.
80. Tomoshige K, Guo M, Tsuchiya T, Fukazawa T, Fink-Baldauf IM, Stuart WD, et al. An EGFR ligand promotes EGFR-mutant but not KRAS-mutant lung cancer in vivo. *Oncogene* [Internet]. Springer US; 2018;37(28):3894–908. Available from: <http://dx.doi.org/10.1038/s41388-018-0240-1>
81. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-tieulent J, Jemal A. Global Cancer Statistics, 2012. *CA a cancer J Clin* [Internet]. 2015;65(2):87–108. Available from: <http://onlinelibrary.wiley.com/doi/10.3322/caac.21262/abstract>
82. Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends - An update. *Cancer Epidemiol Biomarkers Prev*. 2016;25(1):16–27.
83. Tryfonidis K, Papadaki C, Assele S, Lagoudaki E, Menis J, Koutsopoulos A, et al. Association of BRCA1, ERCC1, RAP80, PKM2, RRM1, RRM2, TS, TSP1, and TXR1 mRNA expression levels between primary tumors and infiltrated regional lymph nodes in patients with resectable non-small cell lung cancer. *Pharmacogenomics J* [Internet]. Springer US; 2018;1–10. Available from: <http://dx.doi.org/10.1038/s41397-018-0013-9>
84. Ueda T, Yoshida M. HMGB proteins and transcriptional regulation. *Biochim Biophys Acta - Gene Regul Mech* [Internet]. Elsevier B.V.; 2010;1799(1–2):114–8. Available from: <http://dx.doi.org/10.1016/j.bbagr.2009.11.005>
85. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science* (80- ). 2015;347(6220).
86. Wang SS, Esplin ED, Li JL, Huang L, Gazdar A, Minna J, et al. Alterations of the PPP2R1B gene in human lung and colon cancer. *Science* (80- ). 1998;282(5387):284–7.
87. Watanabe S, Minegishi Y, Yoshizawa H, Maemondo M, Inoue A, Sugawara S, et al. Effectiveness of gefitinib against Non-Small-Cell lung cancer with the uncommon EGFR mutations G719X and L861Q. *J Thorac Oncol* [Internet]. International Association for the Study of Lung Cancer; 2014;9(2):189–94. Available from: <http://dx.doi.org/10.1097/JTO.0000000000000048>
88. Weiss GJ, Bemis LT, Nakajima E, Sugita M, Birks KD, Robinson AW, et al. EGFR regulation by microRNA in lung cancer: correlation with clinical response and survival to gefitinib and EGFR expression in cell lines. *Ann Oncol*. 2008;1053–1059.

89. Wong AST, Gumbiner BM. Adhesion-independent mechanism for suppression of tumor cell invasion by E-cadherin. *J Cell Biol.* 2003;161(6):1191–203.
90. Wu M-F, Hsiao Y-M, Huang C-F, Huang Y-H, Yang W-J, Chan H-W, et al. Genetic Determinants of Pemetrexed Responsiveness and Nonresponsiveness in Non-small Cell Lung Cancer Cells. *J Thorac Oncol* [Internet]. International Association for the Study of Lung Cancer; 2010;5(8):1143–51. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1556086415305724>
91. Xue X, Fei X, Hou W, Zhang Y, Liu L, Hu R. miR-342-3p suppresses cell proliferation and migration by targeting AGR2 in non-small cell lung cancer. *Cancer Lett* [Internet]. Elsevier Ltd; 2018;412:170–8. Available from: <https://doi.org/10.1016/j.canlet.2017.10.024>
92. You DJ, Park CR, Lee HB, Moon MJ, Kang JH, Lee C, et al. A splicing variant of NME1 negatively regulates NF- $\kappa$ B signaling and inhibits cancer metastasis by interacting with IKK $\beta$ . *J Biol Chem.* 2014;289(25):17709–20.
93. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi A J Integr Biol* [Internet]. 2012;16(5):284–7. Available from: <http://online.liebertpub.com/doi/abs/10.1089/omi.2011.0118>
94. Zhang L, Wang Y, Zhang B, Zhang H, Zhou M, Wei M, et al. Claudin-3 expression increases the malignant potential of lung adenocarcinoma cells : role of epidermal growth factor receptor activation. 2017;8(14):23033–47.
95. Zhao W, Zhang L-N, Wang X-L, Zhang J, Yu H-X. Long noncoding RNA NSCLCAT1 increases non-small cell lung cancer cell invasion and migration through the Hippo signaling pathway by interacting with *CDH1*. *FASEB J* [Internet]. 2018;fj.201800408R. Available from: <https://www.fasebj.org/doi/10.1096/fj.201800408R>
96. Zhou Y, Wang X, Huang Y, Chen Y, Zhao G, Yao Q, et al. Down-regulated SOX4 expression suppresses cell proliferation, metastasis and induces apoptosis in xuanwei female lung cancer patients. *J Cell Biochem.* 2015;116(6):1007–18.
97. Zhu Y, Li Y, Wei J, Liu X. The role of sox genes in lung morphogenesis and cancer. *Int J Mol Sci.* 2012;13(12):15767–83.

## CONSIDERAÇÕES FINAIS

Este estudo abrangente aborda o panorama atual do câncer de pulmão. Dados específicos são frequentemente escassos ou ausentes em vários casos, o que deve motivar esforços concentrados para abordar esses tópicos. Vários estudos de perfis de câncer por análise de *microarray* são usadas com estratégias diferentes, tais como tumor contra controle. Nesse estudo o perfil de expressão gênica do tumor é comparado com sua amostra controle correspondente para medir as diferenças e semelhanças entre ambos os fenótipos, em que os perfis de expressão gênica de diferentes amostras do mesmo tipo de câncer são comparados para revelar vias enriquecidas para melhor definir a classificação molecular de um tipo histológico comum de câncer. Embora muitos estudos de análise de *microarray* em doenças humanas tenham sido publicados, apresentamos aqui alguns dos que têm interesse clínico

pela oncologia. No entanto, limitações técnicas não devem ser ignoradas. Um dos principais fatores que determinarão o sucesso futuro do *microarray* é a padronização, não apenas no sentido experimental, mas também no sentido computacional.

Prof. Dr. Lucas Miranda Marques  
*Coordenador do Programa de Pós-graduação em  
Biociências IMS/UFBA*

Vitória da Conquista, 13 de Novembro de 2018.

# Submission #12680

[Home \(author/home.html\)](author/home.html)

## Submissions

[Active \(1\) \(author/submissions/inreview/\)](author/submissions/inreview/)

[Archive \(0\) \(author/submissions/archive/\)](author/submissions/archive/)

## Orders

[Saved orders \(0\) \(author/orders/saved/\)](author/orders/saved/)

[Placed Orders \(0\) \(author/orders/placed/\)](author/orders/placed/)

## Summary

<b>ID</b>	12680
<b>Date Submitted</b>	04/02/2019
<b>Journal</b>	Cancer Genomics & Proteomics
<b>Submitted by</b>	Fernanda Castro Pires dos Santos
<b>Authors</b>	Fernanda Castro Pires dos Santos, Cristina Padre Cardoso, Leandro Martins de Freitas, Sávio Torres de Farias
<b>Title</b>	Differentially gene expression in pulmonary adenocarcinoma meta-analysis
<b>Original File</b>	1549303687.docx (/data/upl/files/1549303687.docx)
<b>Comments</b>	<p>We are pleased to submit our manuscript entitled "Differentially gene expression in pulmonary adenocarcinoma meta-analysis" to be considered for publication in <i>Cancer Genomics &amp; Proteomics (CGP)</i>. In this paper, we report the analyses of gene expression in lung adenocarcinoma through comparison of microarray mRNA expression profile analyses and immunohistochemical detection. Lung cancer is a pathology with high mortality, poor prognosis, and low survival rate. We think that our paper contributes to the effort toward a better comprehension and also report new possible targets to drug/therapy to lung cancer, and it is therefore of interest to readers in areas of cancer research, and bioinformatics. Our bioinformatic analyses integrate information from high-throughput gene and protein expression, and interaction of pathways activated by the differentially expressed genes (DEG), thus giving a broader view of the PDAC molecular aspects. We analyzed the public microarray gene expression of lung cancer looking for shared DEG among samples. The data have information on gene expression from six studies. This shared DEG were able to distinguish between PDAC and normal samples, indicating a robust gene expression pattern among the samples from different studies and moreover different biological process in lung cancer. For these reasons, we think that our study provides insights into lung cancer research field, which is potentially useful to both the molecular understanding of the disease and in the development of new therapeutic targets and, therefore, it should be appropriately published in <i>CGP</i>. We confirm that this work is original and has not been published elsewhere nor it is currently under consideration for publication elsewhere.</p>

## Status

**Submission Status:** Unassigned Submission  
**Initiated:** 04/02/2019  
**Last Modified:** 04/02/2019



[Home \(/home.html\)](/home.html)[Author Home \(author/home.html\)](author/home.html)[Edit My Profile \(author/updateprofile\)](author/updateprofile)**Principal contact for editorial correspondence.**[Reset Password \(/reset-author-password.html\)](/reset-author-password.html)

**Name** Leandro Martins de Freitas  
**Affiliation** Universidade Federal da Bahia  
**Country** Brazil  
**Bio statement** Núcleo de Biointegração  
**E-mail** leandromartinsdefreitas@gmail.com  
**Url**

**Authors**

**Name** Fernanda Castro Pires dos Santos  
**Affiliation** UFBA  
**Country** Brazil  
**Bio statement** -  
**E-mail** fe.castro\_@hotmail.com  
**Url**

**Name** Cristina Padre Cardoso  
**Affiliation** Faculdade Santo Agostinho, Vitória da Conquista, Brazil  
**Country** Brazil  
**Bio statement** -  
**E-mail** crispadre@gmail.com  
**Url**

**Name** Sávio Torres de Farias  
**Affiliation** Universidade Federal da Paraíba  
**Country** Brazil  
**Bio statement** -  
**E-mail** stfarias@yahoo.com.br  
**Url**

**Title and Abstract****Title:** Differentially gene expression in pulmonary adenocarcinoma meta-analysis

**Abstract:** **Background:** Over the past decade, comprehensive efforts have been made to understand the biomarkers of non-small cell lung cancer (NSCLC). This is of clinical relevance, the new biomarkers can be screened in samples of cancer patients, and the therapeutic regimens can be improved. The objective was to apply the computational tools in NSCLC samples to investigate the differentially expressed genes (DEG), and how they interact, and activate biological process. **Materials and Methods:** We used six microarray dataset of the lung adenocarcinoma subtype. The tumor and normal samples were retrieved, and DEG were identified with the Bioconductor/R liming package using  $|\log_{2}FC| \geq 1$  and adjusted p-value  $\leq 0.05$ . The DEG was used to find the KEGG enriched pathways using Cluster Profiler Bioconductor/R package. A protein-protein interaction network was build in the APID database and analyzed with Cytoscape. The sample clustering were performed using mRNA expression values and displayed as heatmap and principal component analysis (PCA). The Human Protein Atlas was used to validate the expression microarray results using immunohistochemical information. **Results:** We have identified that there are more downregulated genes than upregulated in NSCLC samples. We obtained 49 genes upregulated, and 164 genes downregulated shared by all the six studies. The KEGG pathways enriched by upregulated genes are involved in cellular replication and by the downregulated genes are more diverse. The PPI network constructed reveal CDH1 and XPO1 proteins as hub nodes in the upregulated network/ while the downregulated indicates JUN and MYC. The heatmap and PCA showed clusterization of the samples, although there are

[Home \(/home.html\)](#)

[Author Home \(author/home.html\)](#)

[Reset Password \(/reset-author-password.html\)](#)

some few samples distant from their group. The analysis in Human Protein Atlas showed 13 upregulated genes present with high or medium expression in more than 50% of the immunohistochemical samples, and 133 downregulated genes confirmed as validated by Protein Atlas as low or not detected expression. **Conclusion:** This study demonstrated that our computational methodologies for the separation of DEG with defined patterns could be used to identify patients with NSCLC, since it reveals some candidate genes for the genetic signature (GS) in adenocarcinoma. Therefore, microarray data helps to identify diagnostic and predictive factors that can guide a personalized therapy approach. Also, understanding biological processes is one of the most pertinent needs in cancer research, so doctors have significant predictive power. Future studies should validate these results in ways of diagnosing low cost and accessible to public hospitals.

**Supporting Agencies**

**Agencies:** Fundação de Amparo à Pesquisa do Estado da Bahia

[Terms and Conditions \(http://iiar-anticancer.org/terms-and-conditions/\)](http://iiar-anticancer.org/terms-and-conditions/)

[Privacy Policy \(http://iiar-anticancer.org/privacy-policy/\)](http://iiar-anticancer.org/privacy-policy/)

Designed and built by [Slides Center \(http://www.slidescenter.com\)](http://www.slidescenter.com)