

**UNIVERSIDADE DE SÃO PAULO**  
**INSTITUTO DE QUÍMICA**  
**Programa de Pós Graduação em Ciências Biológicas (Bioquímica)**

**DIOGO DE OLIVEIRA PESSOA**

**IDENTIFICAÇÃO E ANOTAÇÃO FUNCIONAL DE RNAS LONGOS  
NÃO CODIFICADORES ASSOCIADOS À SUBTIPOS  
MOLECULARES EM TUMORES PANCREÁTICOS**

Versão Original da Dissertação

São Paulo  
Data do Depósito na SPG:  
03/12/2018

DIOGO DE OLIVEIRA PESSOA

**IDENTIFICAÇÃO E ANOTAÇÃO FUNCIONAL  
DE RNAS LONGOS NÃO CODIFICADORES  
ASSOCIADOS À SUBTIPOS MOLECULARES  
EM TUMORES PANCREÁTICOS**

*Dissertação apresentada ao Instituto de  
Química da Universidade de São Paulo para  
obtenção do Título de Mestre em  
Ciências (Bioquímica)*

**Orientador:** Dr. Eduardo M. R. Reis

**Co-orientador:** Dr. João Carlos Setubal

São Paulo  
2018

## Agradecimentos

*À minha avó por ter me criado e por me ensinar o valor da disciplina e dos estudos.*

*À Amable, por todo o amor, carinho e apoio constante.*

*Aos meus amigos Eric, Leonel, Thiago e Rodrigo, pelo companheirismo nessa jornada.*

*Aos meus amigos de laboratório Ester, Bruno, Vinícius, Diogo (Pellegrina), Bianca, Julio, Felipe, Lutero, Luiza, Thalita, Vanessa, Allan, Dimitrius e Beatriz pelo aprendizado e colaborações durante esse período de pós graduação.*

*Aos meus orientadores Eduardo Reis e João Carlos Setubal por terem propiciado uma experiência de aprendizado tão frutífera.*

*À agência de fomento CAPES pelo apoio financeiro que possibilitou o desenvolvimento deste trabalho.*

## Epígrafe

*"Success flows from perspiration, and inspiration from diligence and effort."*

*"If you have no more tears left to weep, then don't weep. Laugh."*

*- Amos Oz*

## Resumo

PESSOA, D. O. **Identificação e Anotação Funcional de RNAs Longos Não Codificadores Associados à Subtipos Moleculares em Tumores Pancreáticos**. 2018. 64p. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciências (Bioquímica). Instituto de Química, Universidade de São Paulo, São Paulo.

O Adenocarcinoma Pancreático Ductal (*Pancreatic Ductal Adenocarcinoma* - PDAC) é a sétima causa de mortes por câncer no mundo, com uma taxa de sobrevivência de apenas 6%. Embora alguns genes estejam recorrentemente mutados em grande parte dos tumores e sejam críticos para a oncogênese, a heterogeneidade das alterações moleculares tanto no tumor quanto em componentes do microambiente tumoral se reflete em diferentes características fenotípicas com comportamentos clínicos distintos e que têm sido associados a diferentes subtipos moleculares através da análise computacional de dados de alterações somáticas e transcricionais no PDAC. RNAs não codificadores longos (lncRNAs) têm sido reconhecidos como importantes reguladores da expressão gênica em doenças proliferativas mas sua associação com subtipos em PDAC e sua contribuição para o estabelecimento de diferentes fenótipos moleculares e clínicos da doença não foi explorada até o momento. Neste trabalho, foi implementada uma abordagem computacional com o objetivo de identificar e anotar funcionalmente lncRNAs associados a subtipos moleculares de PDAC. Inicialmente, a classificação não supervisionada por Fatoração Matricial Não Negativa (*Non-Negative Matrix Factorization* - NMF) de dados de expressão gênica global de amostras clínicas disponíveis publicamente (*The Cancer Genome Atlas* - TCGA) resultou na identificação de quatro subgrupos distintos de PDAC, que recapitulam os fenótipos Exócrino/Endócrino, Imunogênico, Escamoso e Progenitor descritos na literatura. Uma análise de expressão diferencial permitiu a identificação de assinaturas de expressão gênica características que incluem lncRNAs associados a cada subgrupo. Através da construção de redes de coexpressão de mRNAs e lncRNAs e a identificação de módulos da rede significativamente enriquecidos em genes que participam em vias moleculares conhecidas foi possível inferir possíveis funções biológicas à lncRNAs associados aos diferentes subtipos moleculares, tais como funções exócrinas/neuroendócrinas, imunogênicas, reparo de DNA/progressão do ciclo celular e progenitoras/morfogênicas. Entre eles, o subgrupo 3, enriquecido para fenótipo Escamoso e associado a hiper-expressão do supressor tumoral *TP63*, possui dois lncRNAs hiper-expressos neste subgrupo em relação aos outros subgrupos, sendo que o lncRNA antissenso *FAM83A-*

*ASI* tem a predição de interagir com as proteínas FGFR2, AXIN1, PTEN, BRAF, SMAD4, TGFBR2, TP53 e CDKN2A, que exercem funções importantes na transdução de sinal e supressão tumoral no câncer incluindo o de pâncreas. Entre os lncRNAs hipo-regulados no subgrupo 3 em relação aos outros subgrupos, alguns, como *FLJ42875*, *LOC338651*, *C20orf56* e *LOC38838* tem predição de interação com alta afinidade à proteína BRCA2, que está envolvida no reparo de DNA e participa de processos de resistência à quimioterápicos. As informações trazidas por este estudo permitem gerar hipóteses sobre a contribuição de lncRNAs para a definição de subtipos moleculares de PDAC e priorizar candidatos e experimentos para estudos funcionais de modo a contribuir para um melhor entendimento sobre os mecanismos de ação de lncRNAs na tumorigênese e agressividade do câncer de pâncreas.

**Palavras-chave:** *lncRNAs, PDAC, NMF, Heterogeneidade Tumoral.*

## Abstract

PESSOA, D. O. **Identification of Long Non Coding RNAs Associated to Molecular Subtypes in Pancreatic Tumors**. 2018. 64p. Masters Thesis - Graduate Program in Biochemistry.

Pancreatic Ductal Adenocarcinoma (PDAC) is the seventh cause of worldwide cancer related deaths, with an overall survival rate of only 6%. Some genes might be recurrently mutated in a large number of tumors, and be critical for oncogenesis, molecular alteration heterogeneity both in the tumor as well as in the tumor microenvironment is reflected in diverse phenotypic features with distinct clinical outcomes, and this distinction in multiple molecular subtypes has been drawn through transcriptional and somatic alteration computational analysis within PDAC. Long Non Coding RNAs (lncRNAs) have been recognized as important gene expression regulators in proliferative diseases, but its association to molecular subtypes in PDAC and its contribution in the establishment of diverse molecular and clinical phenotypes hasn't been explored at length until the present. This work focused on the implementation of a computational approach with the objective of lncRNA identification and functional annotation associated to distinct molecular subtypes in PDAC. Initially, Non-negative Matrix Factorization (NMF), an unsupervised classification method, applied to global gene expression data from publicly available clinical samples (The Cancer Genome Atlas - TCGA) resulted in the identification of four distinct PDAC molecular subgroups reminiscent of Exocrine/Endocrine, Immunogenic, Squamous and Progenitor phenotypes. Differential expression analysis allowed a characteristic gene expression signature identification, including distinct molecular subtype associated lncRNAs. mRNA and lncRNA containing gene co-expression modules significantly enriched annotated pathways containing the molecular subtype associated lncRNAs allowed to designate possible molecular functions of the distinct molecular subtype associated lncRNAs, such as exocrine/neuroendocrine, immunogenic, DNA repair/cell cycle progression and progenitor/morphogenic functions. Subgroup 3, enriched with a Squamous phenotype and associated to *TP63* over-expression contains two lncRNAs over-expressed compared to other subgroups; furthermore, the antisense lncRNA *FAM83A-AS1* yielded a predicted lncRNA-protein interaction to FGFR2, AXIN1, PTEN, BRAF, SMAD4, TGFBR2, TP53 and CDKN2A, proteins that play important signal transduction and tumor suppressor roles in several cancer types, including pancreas. Among under-expressed lncRNAs in subgroup 3 compared to the other subgroups, some, such as *FLJ42875*, *LOC338651*, *C20orf56* and *LOC38838* yielded a high protein interac-

tion prediction score with BRCA2, a protein involved in DNA repair and processes resulting in chemotherapy resistance. The information brought by this study allowed to generate hypothesis on lncRNA contribution to define PDAC molecular subtypes, helping prioritize candidates and experiments for functional studies, thus contributing to a better understanding on lncRNA mechanisms related to tumor progression and aggressiveness in pancreatic cancer.

**Key-words:** *lncRNAs, PDAC, NMF, Tumor Heterogeneity.*

## Lista de ilustrações

- Figura 1 – Modelo de progressão de células normais em Adenocarcinomas Pancreáticos Ductais. O acúmulo de mutações e alterações epigenéticas leva à formação de lesões precursoras conhecidas como Lesões Intraepiteliais Pancreáticas (PanIN), que se apresentam em diferentes graus de acordo o nível de atipia citológica e nuclear. . . . . 18
- Figura 2 – Heterogeneidade tumoral em função de expansão e seleção clonal. . . . . 20
- Figura 3 – Componentes celulares do microambiente tumoral. . . . . 22
- Figura 4 – Representação da metodologia de deconvolução de sinal por NMF. . . . . 23
- Figura 5 – Os 2000 genes com expressão mais variável (por desvio absoluto da mediana) foram utilizados como entrada para a classificação não supervisionada por NMF. **A.** Resultado gráfico da iteração de dois a sete possíveis grupos/classes. O coeficiente de correlação cofenética (gráfico superior esquerdo) foi utilizado como um indicador quantitativo da estabilidade do agrupamento obtido a partir das possíveis soluções, onde valores mais próximos de 1 refletem grupos mais consistentes. Para maior resolução, foi escolhida a solução que considera quatro subgrupos. **B.** Representação gráfica da matriz consenso mostrando o agrupamento das amostras para diferentes soluções de número de grupos. . . . . 37
- Figura 6 – **A.** “Heatmap” com abundância relativa dos 4437 genes (linhas) diferencialmente expressos entre os quatro subgrupos de amostras (colunas) resultantes da análise por NMF. **B.** Dos genes representados em **A.**, 60 lncRNAs. Para cada gene a expressão está representada em valores de desvio-padrão (*Z-score*) de todas as amostras. . . . . 38

- Figura 7 – **A.** Silhueta com genes representativos de cada um dos quatro subgrupos de amostras obtidos por NMF. Amostras com valores de silhueta menores que zero foram excluídas para as análises de expressão diferencial. **B.** A análise de expressão diferencial comparando cada um dos grupos versus todos os outros. As barras laterais à esquerda indicam o número total de genes hiperexpressos em cada grupo, e as barras verticais indicam os subconjuntos de genes que são exclusivos do grupo (círculos na legenda) ou que possuem sobreposição (linhas ligando os círculos) entre os grupos definidos. Como critérios para considerar um gene como diferencialmente expresso em um grupo foi utilizado  $\log_2 FC > |1, 2|$  e  $p - \text{valor} < 0, 05$ . . . . . 42
- Figura 8 – Avaliação do enriquecimento para componentes do microambiente tumoral. Utilizando a ferramenta *ESTIMATE*, foi possível estimar de maneira quantitativa o grau de enriquecimento para células estromais e do sistema imune e, também, a pureza tumoral das amostras estratificadas por subgrupo identificado. . . . . 43
- Figura 9 – **A.** Determinação dos valores ideais dos parâmetros para utilização na função de adjacência e determinação dos módulos. Aqui foi escolhido o valor de  $\beta = 10$ , que representa o menor valor de potência obtido quando a curva atinge um platô. **B.** Mensuração da conectividade média estimada para cada valor de Beta correspondente. A obtenção de módulos significativos depende da conectividade entre os genes dentro de um determinado módulo (conectividade intra-modular). . . . . 44
- Figura 10 – **A.** Matriz de dissimilaridade topológica. A distância medida por agrupamento hierárquico aglomerativo do grau de conectividade entre os genes permite a visualização de módulos de genes coexpressos. **B.** Cada linha do dendograma representa um gene, sendo que os ramos delimitam os módulos identificados, os quais são representados por cores distintas (barra inferior). A identificação de módulos de genes permite maior grau de contextualização dos diversos processos biológicos operantes e sua associação com os subgrupos resultantes da classificação não supervisionada por NMF. . . . . 45

Figura 11 – Significância dos módulos de coexpressão. Cada os genes foi testados quanto ao seu potencial prognóstico por meio de análise de regressão de Cox. Neste contexto, módulos com maior número de genes significativos são mais enriquecidos para possíveis marcadores prognósticos. . . . .	46
Figura 12 – Curva de Kaplan Meier para representar a taxa de sobrevida de pacientes com PDAC em função da expressão. Dois genes exemplares dos módulos enriquecidos ( <i>Brown e Tan</i> ) e listados na Tabela 3 foram selecionados para visualização. Para evitar a inserção de grandes flutuações dos dados clínicos na avaliação foi aplicada uma censura à esquerda de 60 dias e à direita de 3 anos. . . . .	47
Figura 13 – Integração de dados dos subgrupos de amostras com os módulos de genes. .	49
Figura 14 – Nível de abundância da proteína BRCA2 estratificada por subgrupo. A abundância do produto proteico é significativamente maior no subgrupo 3, de fenótipo escamoso, o que está associado a menor expressão de lncRNAs com alta afinidade por esta proteína, representando um possível papel de complexo regulador desempenhado por RNAs não codificadores. . . . .	51
Figura 15 – Curva de Kaplan Meier representando a taxa de sobrevida de pacientes com PDAC em função de cada um dos subgrupos identificados. Apesar do resultado estratificado por subgrupo de pacientes não ser significativo, a tendência prognóstica é similar à apontada na literatura por Bailey e colaboradores (BAILEY et al., 2016). Para evitar a inserção de grandes flutuações dos dados clínicos na avaliação foi aplicada uma censura à esquerda de 60 dias e à direita de 3 anos. . . . .	52
Figura 16 – Padrão de expressão dos lncRNAs identificados como diferencialmente expressos no subgrupo 3 e com anotação em amostras obtidas do hospital AC Camargo. . . . .	53

## Lista de tabelas

Tabela 1 – Lista de lncRNAs diferencialmente expressos nos subgrupos. . . . .	38
Tabela 2 – Vias biológicas enriquecidas por subgrupo identificado. . . . .	40
Tabela 3 – Regressão univariada e multivariada de Cox. . . . .	47
Tabela 4 – Sumarização das alterações identificadas em PDAC. . . . .	50

## Lista de Abreviaturas

PDAC\* - *Pancreatic Ductal Adenocarcinoma*

UICC - União Internacional Contra o Câncer

INCA - Instituto Nacional do Câncer

HNPCC\* - *Hereditary Monopolyposis Coloretcal Cancer*

PanIN\* - *Pancreatic Intraepithelial Neoplasia*

NMF\* - *Non-Negative Matrix Factorization*

lncRNA\* - *long non coding RNA*

XIC\* - *X Inactivation Center*

TCGA\* - *The Cancer Genome Atlas*

miRNA\* - *micro RNA*

snRNA\* - *small nuclear RNA*

piRNA\* - *piwi interacting RNA*

MLL\* - *Mixed Lineage Leukemia*

NSCLC\* - *Non Small Cell Lung Cancer*

ENCODE\* - *Encyclopedia of genes and genes variants*

\* Abreviaturas derivadas do inglês.

## SUMÁRIO

	<b>Lista de ilustrações</b>	<b>8</b>
	<b>Lista de tabelas</b>	<b>11</b>
	<b>SUMÁRIO</b>	<b>13</b>
<b>1</b>	<b>INTRODUÇÃO</b>	<b>16</b>
1.1	Câncer de Pâncreas	16
1.2	Aspectos Moleculares da progressão de tumores de pâncreas	17
1.3	Heterogeneidade e Microambiente Tumoral	19
1.4	Abordagens Computacionais para Investigação da Heterogeneidade Tumoral em tumores de Pâncreas	22
1.5	RNAs longos não codificadores e regulação da expressão gênica	25
1.6	RNAs longos não codificadores e Câncer	27
<b>2</b>	<b>OBJETIVOS</b>	<b>30</b>
2.1	Objetivos Gerais	30
2.2	Objetivos Específicos	30
<b>3</b>	<b>MATERIAIS E MÉTODOS</b>	<b>31</b>
3.1	Casuística	31
3.2	Aquisição e Processamento de Dados de Expressão Gênica	31
3.3	Aquisição de Dados dos Níveis de Proteínas	32
3.4	Aquisição de Dados de Mutação	32
3.5	Classificação de Amostras Tumorais em Subtipos moleculares	32
3.6	Identificação de Assinaturas Moleculares por Análise de Expressão Diferencial	32
3.7	Análise de Enriquecimento para Vias Biológicas	33
3.8	Análise dos Componentes do Microambiente Tumoral	33
3.9	Análise de redes de Co-expressão Gênica	33
3.10	Análise de Sobrevida	34

3.11	Anotação estrutural e funcional de lncRNAs . . . . .	34
4	RESULTADOS . . . . .	36
4.1	Classificação Não Supervisionada de Tumores Pancreáticos em Subgrupos moleculares . . . . .	36
4.2	Identificação de Assinaturas Moleculares Associadas a Subtipos Moleculares de PDAC . . . . .	37
4.3	Anotação Funcional das Assinaturas de Genes Associada aos Subgrupos . . . . .	40
4.4	Análise do Microambiente Tumoral dos Subgrupos de PDAC . . . . .	43
4.5	Análise de rede de Coexpressão Gênica . . . . .	43
4.6	Contexto Biológico e Clínico dos Módulos de Coexpressão . . . . .	45
4.7	Integração de Dados de coexpressão gênica com subtipos moleculares de PDAC . . . . .	48
4.8	Anotação de lncRNAs associados com subtipos moleculares de PDAC . . . . .	50
4.9	Análise dos lncRNAs Identificados em Amostras do AC Camargo . . . . .	52
5	DISCUSSÃO . . . . .	54
5.1	Desafios Impostos pela Heterogeneidade Tumoral . . . . .	54
5.2	Influência do Microambiente no Contexto de Subgrupos Tumoriais . . . . .	55
5.3	O Contexto de lncRNAs em Subtipos Moleculares de PDAC . . . . .	56
5.4	Integração de Dados . . . . .	57
6	CONCLUSÕES . . . . .	60
	REFERÊNCIAS . . . . .	61

# 1 INTRODUÇÃO

## 1.1 Câncer de Pâncreas

O câncer pancreático é a sétima causa de morte associada a câncer no mundo, e estima-se cerca de 418.000 novos casos diagnosticados até o ano de 2020 (KHAN et al., 2017). A sobrevida média de pacientes com tumores pancreáticos é de 2 a 8 meses após o diagnóstico, e apenas 6% dos pacientes permanece vivo após 5 anos (SCHLITTER et al., 2017). Entre os fatores mais influentes para baixa sobrevida pode-se citar o estágio avançado do tumor quando diagnosticado e a resistência a tratamentos quimioterápicos.

O tipo de tumor de pâncreas mais comum é o adenocarcinoma ductal (*Pancreatic Ductal Adenocarcinoma* – PDAC), originado no tecido glandular de função exócrina e correspondente a 90% dos casos diagnosticados. No Brasil, é responsável por cerca de 2% de todos os tipos de câncer diagnosticados e por 4% do total de mortes por essa doença. É mais raro entre os jovens e mais recorrente em indivíduos de idade mais avançada. De acordo com a União Internacional Contra o Câncer (UICC), o número de casos aumenta com o avanço da idade: de 10 em cada 100.000 habitantes entre 40 e 50 anos para 116 em cada 100.000 habitantes entre 80 e 85 anos. A incidência é mais significativa em homens, com número total de mortes de 8.710, sendo 4.373 homens e 4.335 mulheres (INCA, 2016).

A epidemiologia do câncer pancreático aponta vários possíveis fatores de risco para o surgimento do tumor, os quais podem ser divididos em esporádico e familiar. Entre os fatores esporádicos que podem levar a tumorigênese, a pancreatite crônica apresenta alto risco, possivelmente em função da divisão celular epitelial que ocorre durante o processo de reparo tecidual, ou de espécies reativas de oxigênio que causam danos no DNA (GALL; WASAN; JIAO, 2015). Processos inflamatórios possibilitam que populações celulares clonais com mutações somáticas, normalmente eliminadas pelo sistema imune, sobrevivam e se proliferem. Outro fator de risco é o tabagismo, que pode vir a causar dano ao DNA em células pancreáticas, levando a oncogênese, expansão clonal e acúmulo de mutações.

A obesidade pode levar ao aumento de câncer pancreático por induzir um estado pró inflamatório e a hiperinsulinemia. A diabetes tipo II é um conhecido fator associado ao câncer pancreático, possivelmente aumentando o risco de formação tumoral através da hiperinsuline-

mia e hiperglicemia, que leva à desregulação dos níveis de glicose circulantes. A hiperglicemia pode aumentar o risco de câncer por facilitar a sobrevivência e expansão clonal de mutantes para *KRAS*, devido à sua dependência variável de glicose.

Entre algumas síndromes hereditárias que possam ser relacionadas à formação de câncer pancreático, pode-se citar a síndrome de Peutz-Jeghers, uma doença autossômica rara, em que mutações no supressor tumoral *SKT11* estão associadas ao risco de formação tumoral. Outro exemplo é a Monopolipose de Câncer Colorretal Hereditário (*Hereditary Monopolyposis Colorectal Cancer* – HNPCC), na qual mutações germinativas de reparo nos genes *MLH1*, *MSH2*, *MSH6* e *PMS2* levam à instabilidade genômica e consequente formação de tumores colorretais e pancreáticos (AMUNDADOTTIR, 2016).

Além disso, estudos genômicos de associação identificaram variantes que aumentam o risco para o câncer pancreático, em loci gênicos que codificam a transcritase reversa da telomerase (*TERT*), receptor nuclear subfamília 5 grupo A membro 2 (*NR5A2*), *zinc finger 3* (*ZNRF3*) e *TP63* ou eficiência de reparo do DNA (manutenção estrutural do cromossomo 2 (*SMC2*)) (MAKOHON-MOORE; IACOBUZIO-DONAHUE, 2016).

## 1.2 Aspectos Moleculares da progressão de tumores de pâncreas

Acredita-se que o desenvolvimento de PDAC ocorra através de uma série de passos (textbf-Figura 1), tendo como ponto inicial o surgimento de lesões precursoras não cancerosas conhecidas como Neoplasias Intraepiteliais Pancreáticas (*Pancreatic Intraepithelial Neoplasia* - PanIN) que eventualmente progridem para o carcinoma invasivo (GHARIBI; ADAMIAN; KELBER, 2016).

O modelo de progressão tumoral mais aceito é que as PanINs se formem a partir de epitélio pancreático ductal, progressivamente dando origem ao carcinoma *in situ*, seguindo finalmente para carcinoma pancreático invasivo (HRUBAN et al., 2000). Em função do grau de alteração celular e molecular, lesões PanIN podem se subdividir em lesões baixas (PanIN-1A/B), intermediárias (PanIN-2) e altas (PanIN-3).

Mutações no oncogene homólogo de *Kirsten rat sarcoma* (*KRAS*), assim como a hiperexpressão de microRNAs e a ativação de fatores estromais são frequentes em lesões PanIN de baixo grau (PanIN-1). A medida que ocorre a progressão tumoral, outras alterações comuns são a hiper expressão do gene Mucina 1 (*MUC1*) e mutações que levam à inativação de

p16/*CDKN2A* (um importante supressor tumoral), características de lesões de grau intermediário (PanIN-2).

Lesões caracterizadas como PanIN-3 são marcadas pela inativação de genes supressores de tumor como *TP53*, o qual expressa a proteína p53, crítico para o controle do ciclo celular, *BRCA2*, gene de suscetibilidade de câncer de mama 2 e *SMAD4* (*mothers against decapentaplegic homolog 4*) (TINDER; SUBRAMANI; BASU, 2008).

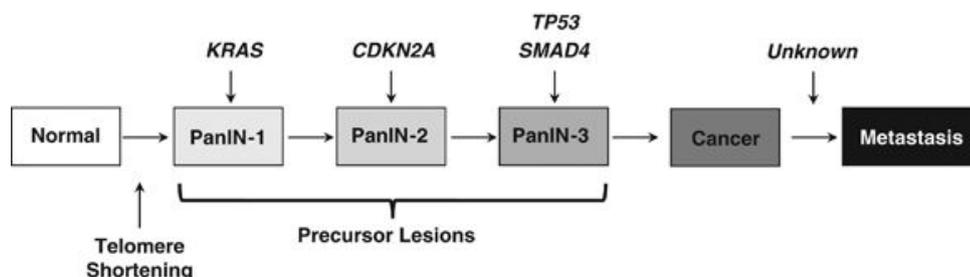


Figura 1 – Modelo de progressão de células normais em Adenocarcinomas Pancreáticos Ductais. O acúmulo de mutações e alterações epigenéticas leva à formação de lesões precursoras conhecidas como Lesões Intraepiteliais Pancreáticas (PanIN), que se apresentam em diferentes graus de acordo o nível de atipia citológica e nuclear.

Os tumores pancreáticos exibem alterações em cascatas de sinalização autócrinas e parácrinas que promovem a proliferação celular, migração, invasão e metástase. A auto suficiência mitogênica alcançada por meio da ativação da expressão de fatores proliferativos como *transforming growth factor- $\alpha$*  (*TGF $\alpha$* ), *insulin-like growth factor 1* (*IGF1*), *fibroblast growth factors* (*FGFs*) e *hepatocyte growth factor* (*HGF*) - e seus respectivos receptores de tirosina quinase - como *epidermal growth factor receptor* (*EGFR*), *receptor tyrosine-protein kinase erbB-2* (*ERBB2*; também conhecido como *HER2*), *HER3*, *receptor IGF1* (*IGF1R*), receptores *FGF* (*FGFRs*) e *receptor HGF* (*HGFR*; também conhecido como *MET*), contribuem para conferir habilidades de migração e proliferação celular.

Essas cascatas de sinalização são ativadas em conjunto com vias como a do *signal transducer and activator of transcription 3* (*STAT3*), *nuclear factor- $\kappa$ B* (*NF- $\kappa$ B*) e *AKT*, levando à ativação de processos celulares de sobrevivência e anti-apoptóticos (PREIS; KORC, 2011). Vias comumente ativadas durante o desenvolvimento tecidual, como de *WNT*, *SHH* e *NOTCH*, também tendem a ser reativadas em alguns casos de tumores pancreáticos (MAGLIANO et al., 2007).

PDAC também é caracterizado por altas taxas de metástase, sendo que a maioria dos tumores, quando diagnosticados, encontra-se já disseminados em outros órgãos. Estudos anteriores

mostraram a relação entre o gene que codifica o inibidor de metaloproteinase (*TIMP1*) e metástases hepáticas do câncer pancreático. Outro importante mecanismo que leva a formação de metástases é a sinalização através de vesículas extracelulares (exossomos) secretadas pelas células tumorais, que modula a função de células do microambiente e auxilia na formação de nichos pré-metastáticos (GHARIBI; ADAMIAN; KELBER, 2016).

O desafio imposto pelo microambiente na formação tumoral, com baixas concentrações de glicose, estresse oxidativo, pouca vascularização, baixa pressão de oxigênio e baixa perfusão intratumoral pode influenciar no grau de resistência por exercer uma pressão seletiva, favorecendo o crescimento de células mais agressivas de PDAC (CHAND et al., 2016).

O estroma já foi previamente associado a facilitação de quimiorresistência por gerar uma barreira física que limita o acesso de agentes quimioterápicos, assim como a interação parácrina e transformação conferida às células tumorais (PROVENZANO et al., 2012). Estudos anteriores observaram que a combinação de gemcitabina com um inibidor de *JAK2* resulta na depleção de células estromais e uma diminuição do crescimento tumoral com melhora na sobrevivência (WORMANN et al., 2016).

Contudo, existe uma contradição sobre o papel do estroma como barreira física para conferir quimioresistência. Um estudo de Aiello e colaboradores (AIELLO et al., 2016) com modelos de camundongo demonstrou uma redução de metástases por meio de administração de quimioterápicos em lesões tanto de baixa como de alta densidade estromal, mostrando a necessidade de estudo e caracterização contexto específico do papel do estroma em PDAC.

### 1.3 Heterogeneidade e Microambiente Tumoral

Os avanços em técnicas de sequenciamento de DNA e RNA tem sido instrumentais para a elucidação de assinaturas moleculares de diferentes tipos de câncer. Os avanços nos estudos de tumores pancreáticos por meio da análises de alterações somáticas no exoma utilizando ferramentas de sequenciamento de última geração levou à descoberta de importantes mutações *driver* recorrentes nos genes *KRAS*, *TP53*, *SMAD4* e *CDKN2A*, além da identificação de vias de sinalização alteradas no tumor (SCHLITTER et al., 2017).

Embora algumas alterações tenham sido associadas a lesões precursoras do PDAC, como discutido acima, o processo de transição que células normais sofrem ao se transformarem em células malignas têm um componente estocástico, ou seja, seu desenvolvimento não segue um

curso pré-determinado, pois deriva da alteração de processos celulares chave que conferem uma vantagem proliferativa às células cancerosas. Essa evolução por caminhos alternativos resulta em uma heterogeneidade genética, transcriptômica, epigenética e fenotípica (DAGOGO-JACK; SHAW, 2018).

A heterogeneidade intratumoral denota a heterogeneidade entre as células tumorais de um único paciente resultantes de seleção e expansão clonal, enquanto que a heterogeneidade intertumoral se refere a heterogeneidade entre pacientes que foram acometidos por tumores do mesmo tipo histológico (**Figura 2**) (DAGOGO-JACK; SHAW, 2018).

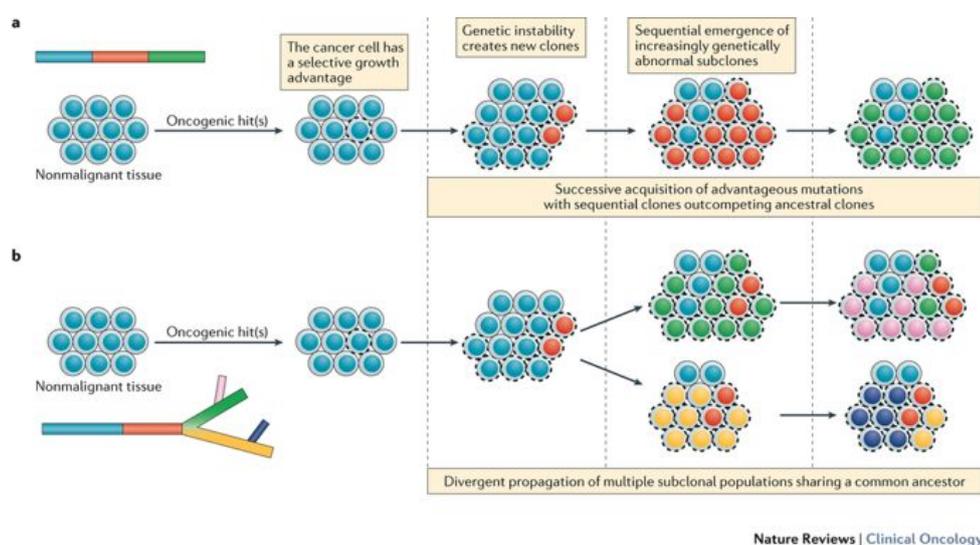


Figura 2 – Heterogeneidade tumoral em função de expansão e seleção clonal.

É importante considerar a possibilidade de que células geneticamente idênticas possam existir em diferentes estados/configurações celulares, devido à diferenças epigenéticas e influência do microambiente tumoral. Alterações epigenéticas são modificações fenotípicas (na função de um gene) que são herdadas sem a ocorrência de uma concomitante alteração na sequência do DNA (JONES; BAYLIN, 2007).

Desta maneira, a epigenética pode ser definida como um conjunto de processos dinâmicos que regulam a expressão gênica, conferindo um maior grau de plasticidade genômica e diversidade na identidade celular. As modificações epigenéticas mais comuns para a modulação da expressão de genes são a metilação de DNA em dinucleotídeos CpG, a modificação química de histonas e a ação regulatória RNAs não codificadores, que são vitais na regulação epigenética e nas funções tecido-específicas (WIDSCHWENDTER et al., 2018).

A descoberta da perda bialélica de função do gene remodelador de cromatina *SMARCB1* (*SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b*,

*member 1*; também conhecido como *SNF5*) em tumores pediátricos rabdóides é um dos primeiros exemplos de mecanismo epigenético de driver de tumorigênese (VERSTEEGE et al., 1998). *SMARCB1* é um dos componentes do complexo remodelador de cromatina dependente de ATP (Adenosina Trifosfato) *SWI/SNF*, que pode atuar dissociando, trocando ou deslocando nucleossomos e assim levando tanto a ativação quanto a repressão de genes (KALIMUTHU; CHETTY, 2016).

Embora as modificações epigenéticas componham apenas uma parte das alterações moleculares importantes para a tumorigênese, a sua natureza dinâmicas e responsiva à mudanças do microambiente tumoral faz com que exerçam um importante papel na definição do destino do estado celular e comportamento em um determinado momento ou em resposta à terapia. Além disso, marcadores epigenéticos representam uma espécie de “histórico” do câncer, visto que algumas dessas marcas epigenéticas se mantêm após mudanças de estado celular (ALIZADEH et al., 2015).

Outro fator que contribui para a heterogeneidade tumoral é a interação que há entre células malignas e células não transformadas, e que compõem o microambiente tumoral (**Figura 3**). A constante comunicação intercelular é mediada por moléculas como citocinas, quimiocinas, fatores de crescimento, além de enzimas inflamatórias e remodeladoras da matriz.

Existem muitos aspectos em comum entre a atividade das células do microambiente tumoral e processos celulares inflamatórios e de reparo celular, que ocorrem devido à ativação de processos inflamatórios por mutações oncogênicas em células malignas (MANTOVANI et al., 2008). O melhor entendimento dos componentes celulares e moleculares do microambiente tumoral pode, então, levar à formulação de terapias complementares alternativas.

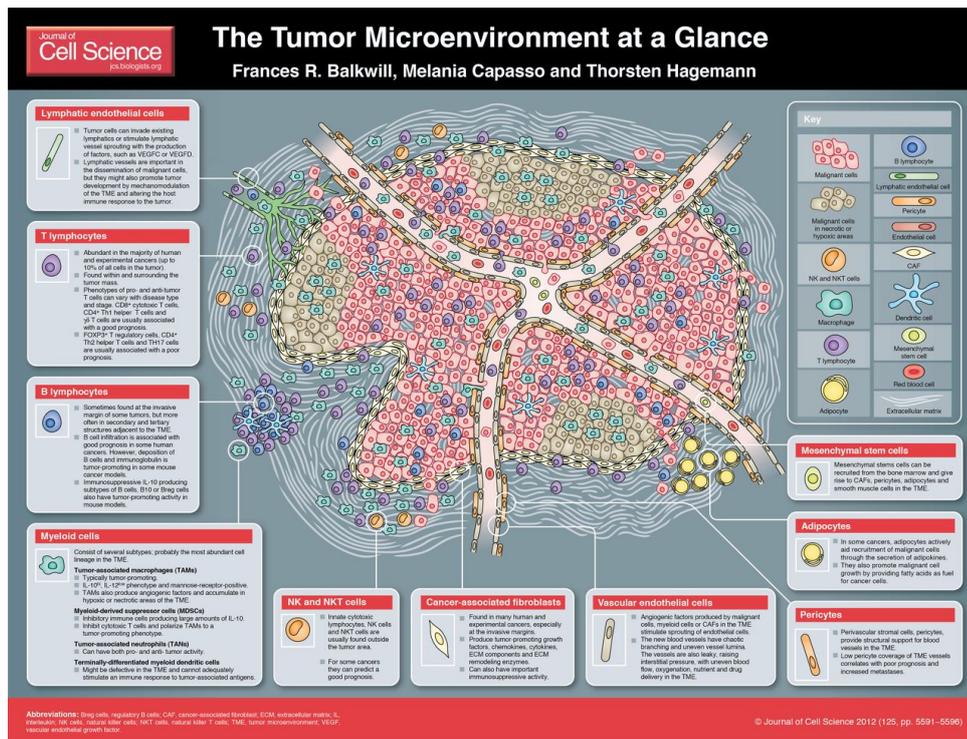


Figura 3 – Componentes celulares do microambiente tumoral.

#### 1.4 Abordagens Computacionais para Investigação da Heterogeneidade Tumoral em tumores de Pâncreas

A heterogeneidade celular e molecular existente, seja em um único tumor, em diferentes sítios de neoplasia de um paciente ou entre tumores de vários pacientes representa uma variável confundidora no entendimento da evolução tumoral e na habilidade de selecionar terapias efetivas que possam contornar a resistência ao tratamento (ALIZADEH et al., 2015).

A heterogeneidade deriva principalmente de alterações genéticas/epigenéticas e da consequente seleção evolutiva de células com fenótipos que venham a conferir uma vantagem proliferativa (ALIZADEH et al., 2015). O aprofundamento do contexto de formação de subclones clinicamente distintos pode auxiliar na interpretação das informações biológicas significativas (MAKOHON-MOORE; IACOBUZIO-DONAHUE, 2016).

Estudos recentes abordaram a questão da heterogeneidade tumoral utilizando ferramentas computacionais capazes de identificar assinaturas moleculares subjacentes a tipos celulares específicos que compõem a massa tumoral (COLLISSON et al., 2011; MOFFITT et al., 2015; BAILEY et al., 2016). Estes estudos utilizaram um método de decomposição de sinal conhecido como Fatoração Matricial Não Negativa (*Non Negative Matrix Factorization* – NMF), que consiste em uma metodologia de classificação não supervisionada que permite agrupar as amos-

tras segundo um padrão característico, como por exemplo de perfis de expressão gênica ou de mutação somática.

A classificação de amostras em função de alterações na expressão gênica ou de mutações somáticas feita por NMF, proposta por Brunet et al. em 2004 (BRUNET et al., 2004), toma um conjunto de dados com  $N$  genes de  $M$  amostras, representado como uma matriz  $A$  de dimensão  $N \times M$ , em que as linhas representam os genes e as colunas as amostras. O objetivo é encontrar um pequeno número de metagenes, definidos como uma combinação linear positiva dos  $N$  genes. É possível então aproximar o padrão de expressão das amostras como combinações lineares desses metagenes.

A fatoração da matriz  $A$  é feita em duas matrizes positivas,  $A \sim WH$  (**Figura 4**). A matriz  $W$  tem dimensão  $N \times k$ , em que cada coluna  $k$  define um fator (ou metagene) e cada entrada  $w_{ij}$  tem um coeficiente  $i$  do metagene  $j$ . A matriz  $H$  tem dimensão  $k \times M$ , em que cada coluna  $M$  corresponde ao padrão de expressão para a amostra correspondente e cada entrada  $h_{ij}$  representa a expressão do metagene  $i$  da amostra  $j$ . Assim, a fatoração  $A \sim WH$  permite agrupar as  $M$  amostras e  $k$  grupos, em que cada amostra  $j$  é inserida em um grupo  $i$  se o valor  $h_{ij}$  for o maior para a coluna  $j$  (BRUNET et al., 2004).

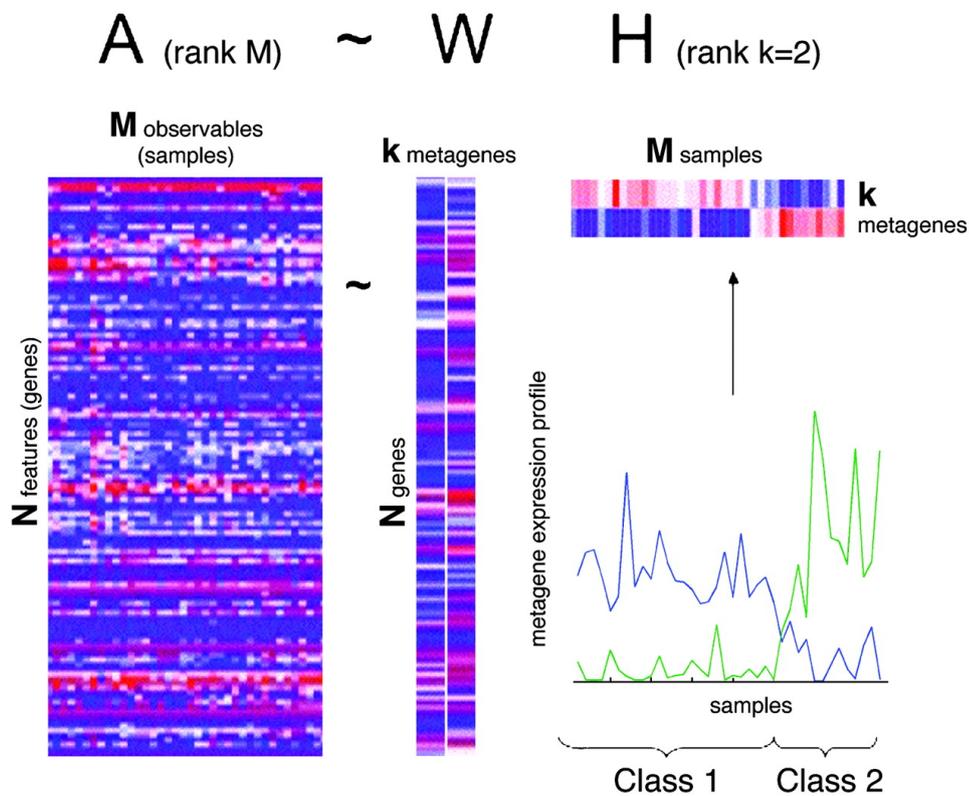


Figura 4 – Representação da metodologia de deconvolução de sinal por NMF.

Em 2011 Collisson e colaboradores (COLLISSON et al., 2011), aplicaram a metodologia de

classificação por NMF em dados de expressão gênica de tumores pancreáticos micro-dissecados obtidos por hibridização com microarranjos de DNA. Esta análise levou à identificação de três subtipos de tumores pancreáticos, que foram definidos como “Clássico”, “Quasi-mesenquimal” e “Exócrino”, em função da interpretação e correlação com a literatura do padrão de expressão de assinaturas de expressão gênica identificadas nas amostras.

O subtipo Clássico possui alta expressão de genes epiteliais e associados à adesão celular, o subtipo Quasi-Mesenquimal possui expressão alta de genes associados ao mesênquima e o subtipo Exócrino expressão aumentada de genes relacionados à digestão enzimática. Interessantemente, foram observadas associações entre os subtipos moleculares das diferentes assinaturas de expressão e características clínicas de PDAC; por exemplo, a associação entre o subtipo molecular e o padrão de resposta a tratamento com as drogas gemcitabina e erlotinibe em linhagens celulares mostraram que o subtipo “Quasi-mesenquimal” tem maior sensibilidade média a gemcitabina em relação ao subtipo “Clássico”, enquanto que com erlotinibe verificou-se o inverso.

Moffitt e colaboradores (MOFFITT et al., 2015) avaliaram o perfil de expressão gênica em PDAC pela análise de amostras de tumores primários, metástases e amostras normais utilizando NMF. A abordagem levou à obtenção de duas assinaturas tumorais, denominadas “Clássica” e “Basal”, sendo o segundo de pior prognóstico e com semelhanças moleculares a tumores basais de bexiga e de mama. Além disso, a avaliação do perfil de expressão gênica em diferentes linhagens celulares permitiu a identificação de dois subtipos de estroma pancreático, denominados “estroma normal” e “estroma ativado”, sendo essa diferença representativa na avaliação de sobrevida para pacientes em cada um desses sub grupos.

Bailey e colaboradores (BAILEY et al., 2016) avaliaram o perfil de expressão gênica em 96 amostras de tumores de pâncreas sequenciadas por RNA-seq para RNA total, o que permite maior sensibilidade e representatividade das possíveis espécies de RNA a serem encontradas. A análise por NMF dos perfis globais de expressão permitiu a identificação de quatro subtipos de amostras, denominadas como “Escamoso”, “Anormalmente Diferenciado Endócrino-Exócrino”, “Pancreático Progenitor” e “Imunogênico”. O enriquecimento para vias biológicas em cada um dos subtipos mostrou alteração em vias como *TP53*, *KDM6A* e *TP63* para o subtipo Escamoso, *FOXA2*, *PDX1* e *NMX1* para o subtipo Pancreático Progenitor, *KRAS* para o subtipo ADEX, *NR5A2* e *RBPJL* para o subtipo Exócrino e *NEUROD1*.

Recentemente, Raphael e colaboradores (RAPHAEL et al., 2017), analisaram dados genô-

micos, transcriptômicos e proteômicos de amostras de PDAC do banco de dados do TCGA, corroborando observações anteriores de mutações frequentes em *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, além de *RNF43*, *ARID1A*, *TGFbR2*, *GNAS*, *RREB1*, e *PBRM1*. Foi documentado que as amostras *KRAS* selvagem possuíam mutações recorrentes em outros genes *driver* oncogênicos como *GNAS*, *BRAF*, *CTNNB1*, além de genes adicionais da via de *RAS*. A comparação das assinaturas moleculares geradas por Collisson et al., Moffitt et al. e Bailey et al. nas amostras do TCGA mostrou sobreposição entre as amostras classificadas como “Progenitoras” por Bailey et al., “Clássicas” por Collisson et al. e “Clássicas” por Moffitt et al. Também foi reportada a sobreposição entre as classes “Escamosa” de Bailey et al. e “Basal” de Moffitt et al., assim como da classe “Aberrantemente Diferenciada Exócrina/Endócrina” de Bailey et al. e “Exócrina” de Collisson et al., reforçando os fenótipos característicos identificados em PDAC.

A comparação das diferentes classificações em subtipos moleculares a partir dos perfis de expressão das diferentes assinaturas obtidas por Collisson, Moffitt e Bailey (descritas acima) leva à interpretação de que apesar dos diferentes resultados obtidos por cada um dos grupos, há uma sobreposição significativa dos fenótipos identificados, corroborando de maneira robusta a abordagem computacional utilizada, assim como a existência de subtipos moleculares definidos.

## 1.5 RNAs longos não codificadores e regulação da expressão gênica

O desenvolvimento de tecnologias de sequenciamento em larga escala de genoma e transcriptoma permitiu a anotação com alta-resolução dos elementos funcionais no genoma humano. Existem diversas anotações do genoma humano, cada uma com suas vantagens e desvantagens. As anotações são baseadas em dois principais métodos: anotação automatizada, na qual o transcriptoma é montado de novo, gerando resultados de maneira rápida e de custo mais baixo, porém com anotações incompletas e de baixa acurácia. A segunda é a anotação manual, que rende catálogos de alta qualidade, mas requer mais tempo e recursos para gerar novas anotações (USZCZYNSKA-RATAJCZAK et al., 2018).

Um catálogo amplamente utilizado é GENCODE, produzido pelo consórcio ENCODE (*Encyclopedia of genes and genes variants*) que identificou os elementos gênicos presentes no genoma humano através de análise computacional, anotação manual e validação experimental (HARROW et al., 2012). Baseado na anotação mais atual do (GENCODE versão 25) aproximadamente 51,8% do genoma é transcrito, mas somente 1,2% do genoma codifica produtos

proteicos (RANSOHOFF; WEI; KHAVARI, 2017), e 23,5% dos genes expressos representam RNAs longos não codificadores (DJEBALI et al., 2012).

O atual estado da arte em estudos moleculares da heterogeneidade tumoral de PDAC teve seu foco direcionado em espécies de RNAs codificadores de proteínas, mas pouca atenção foi dada às outras espécies de RNAs. Uma espécie de RNA de grande interesse são os RNAs longos não codificadores (*long non coding RNA* – lncRNA), transcritos com mais de 200 nucleotídeos que podem ser formados a partir de regiões intrônicas, intergênicas, antissenso ou enhancer do genoma (KAPRANOV et al., 2010).

Essa definição por tamanho serve como um critério ad hoc para separar os lncRNAs de outras espécies de RNAs não codificadores menores, como micro-RNAs (miRNAs), RNAs de transferência (transfer RNAs - tRNAs), pequenos RNAs nucleares (small nuclear RNAs - snRNAs), RNAs ribossomais (ribossomal RNAs - rRNAs) ou *piwi-interacting RNAs* (piRNAs).

Apesar de possuírem potencial codificador limitado ou ausente determinado através da análise computacional, dados gerados em estudos de proteômica revelaram que alguns lncRNAs possuem fases de leitura aberta (*Open Reading Frame* - ORF) e assim codificando produtos ainda pouco caracterizados (QUINN; CHANG, 2016).

A grande versatilidade dos lncRNAs permite sua interação com moléculas de DNA, RNA através de pareamento de trechos da cadeia nucleotídica, ou com proteínas através de domínios estruturais gerados por estruturas secundárias e dobramento espacial do RNA assumindo estruturas tridimensionais (PARALKAR; WEISS, 2013).

Os transcritos provenientes de *loci* de lncRNAs podem regular genes localmente (em *cis*) através de: recrutamento de fatores regulatórios para o locus e/ou modular sua função, se o processo de transcrição e/ou *splicing* do lncRNA vier a conferir habilidade regulatória independente da sequência do transcrito, ou se a regulação em *cis* depender somente de elementos do DNA co-localizados no *locus* do promotor ou do gene do lncRNA, sendo independente de sua expressão. A regulação em *trans* pode ocorrer por meio de lncRNAs que atuam em sítios distantes do seu local de transcrição, através do recrutamento e regulação de complexos proteicos que influenciam na estrutura e organização da cromatina e no padrão de expressão gênica ou interagindo com e regulando o comportamento de proteínas e/ou moléculas de RNA (KOPP; MENDELL, 2018).

Vários exemplos de funcionalidade em diversos processos celulares foram descobertos, desde células tronco embrionárias, pluripotência, regulação do ciclo celular, além de doen-

ças como câncer. Achados recorrentes de lncRNAs atuando como formadores de complexos ribonucleico-proteico mostram seu papel enquanto reguladores da expressão gênica (RINN; CHANG, 2012).

Um exemplo bem estabelecido de funcionalidade de lncRNAs é o da compensação de dosagem gênica por inativação de um dos cromossomos X em fêmeas de mamíferos. O silenciamento depende de um *locus* específico no cromossomo X, chamado de Centro de Inativação de X (*X Inactivation Center* - XIC), o qual contém o gene *Xist*. Esse gene, apesar de ser processado por inserção de 7-metilguanossina na extremidade 5' e poli-adenilação na região 3', é retido no núcleo.

O modelo atual mais aceito de silenciamento do cromossomo X é o de recrutamento do Complexo Repressor *Polycomb 2* (*Polycomb Repressor Complex 2* - PCR2) pelo RNA não codificador *Xist*. O RNA *Xist* atua em *cis*, acumulando-se no cromossomo a ser inativado, assim ativando uma cascata de eventos e o consequente recrutamento do complexo PCR2, que atua por inserção de marcas epigenéticas (trimetilação da histona três na lisina 27 - *H3K27me3*), o que leva ao remodelamento da cromatina e ao silenciamento estável do cromossomo (CERASE et al., 2015; KOPP; MENDELL, 2018).

O RNA antissenso do transcrito *HOX* (*HOX transcript antisense RNA* - *HOTAIR*) funciona de maneira cooperativa com o complexo PCR2, mediando a repressão do *locus* gênico *HOXD* através da inserção de marcas epigenéticas (*H3K27me3*), associadas ao silenciamento de conjuntos de genes. A formação de estruturas secundárias por *HOTAIR* permite sua interação com o complexo remodelador de cromatina PCR2 (RINN et al., 2007).

## 1.6 RNAs longos não codificadores e Câncer

LncRNAs são de grande interesse porque são mais abundantes que genes codificadores de proteína, sendo uma modulação nos níveis de expressão de uma gama de lncRNAs observada em um determinado subtipo de câncer, fornecendo uma janela maior para a detecção de biomarcadores baseados em lncRNAs subtipo-específicos.

Níveis de expressão subtipo/tecido específico são cruciais no desenvolvimento de novos biomarcadores e terapias personalizadas. Por serem de maior tamanho, lncRNAs podem dobrar-se e formar estruturas secundárias/terciárias complexas capaz de interagir com várias proteínas, fatores de transcrição, mRNAs complementares e sequências de DNA, assim auxiliando na tu-

morigênese e progressão tumoral. Sua presença em grande variedade também serve como uma plataforma no desenvolvimento de novas drogas baseadas em estrutura. A participação de lncRNAs em diversos processos de sinalização celular tecido-específica possibilita sua utilização na formulação de novas estratégias de diagnóstico e tratamento subtipo-específico em câncer (BHAN; SOLEIMANI; MANDAL, 2017).

O papel funcional de lncRNAs em processos tumorigênicos ainda não é totalmente compreendido. Um exemplo significativo do envolvimento de um RNA não codificador em câncer é o da promoção de metástase em tumores de mama por *HOTAIR* (GUPTA et al., 2010). Neste trabalho Gupta e colaboradores identificaram o lincRNA *HOTAIR* com expressão aumentada em tumores primários e metastáticos, além do potencial preditivo de eventual metástase em função dos níveis de expressão em tumores primários.

A super expressão de *HOTAIR* induzida em células epiteliais cancerosas levou a um maior recrutamento de *PRC2*, com a consequente reprogramação gênica por meio de remodelamento da cromatina, alterando o fenótipo celular para características mesenquimais, assim como conferindo maior habilidade de invasão e metástase dependente de *PCR2*.

Outro reconhecido exemplo de lncRNA envolvido em câncer é o transcrito *HOXA* da ponta distal (*transcript at the distal tip - HOTTIP*). Já foi demonstrado anteriormente que *HOTTIP* regula interações de cromatina no *locus HOX* do qual é produzido, levando a ativação dos genes *HOX* pela ligação da proteína adaptadora WDR5. A interação de WDR5 com complexos de linhagem mista de leucemia (*mixed lineage leukemia - MLL*) leva à metilação da histona três na lisina quatro (*H3K4*), assim ativando a transcrição de vários genes (WANG et al., 2011).

Transcritos Naturalmente Antissenso (*Naturally Antisense Transcripts - NAT*) são comuns no genoma de mamíferos, sendo que cerca de 20% dos transcritos em humanos podem vir a formar pares senso-antissenso. NATs podem ser produzidos tanto de genes codificadores de proteínas como de genes não codificadores, sendo alguns NATs lncRNAs. Um exemplo de NAT em câncer é o lncRNA *WRAP53*, um antissenso do supressor tumoral *TP53*. *WRAP53* pode levar à indução de p53 via interação com a região 5' UTR do RNA mensageiro. Um decréscimo nos níveis de *WRAP53* leva ao concomitante decréscimo nos níveis de indução de p53 em caso de danos ao DNA (MAHMOUDI et al., 2009). Apesar da classificação como não codificador, *WRAP53* é capaz de codificar uma proteína responsável pela formação do corpo de Cajal, promovendo sobrevivência de células tumorais (MAHMOUDI et al., 2011).

Alguns estudos anteriores do nosso laboratório também caracterizaram a associação entre

a alteração de lncRNAs e a desregulação gênica. Beckedorff e colaboradores (BECKEDORFF et al., 2013) mostraram que a alteração nos níveis do lncRNA *ANRASSF1* levavam a inibição do supressor tumoral *RASSF1A* através do recrutamento de *PCR2*, assim contribuindo para a tumorigênese com modelo de células HeLa.

Mecanismos de interferência por lncRNAs são uma fonte a ser investigada de maneira mais aprofundada. Em um trabalho de nosso grupo de pesquisa, Tahira e colaboradores (TAHIRA et al., 2011), ao estudar microarranjo de 38 amostras de PDAC verificaram a expressão de conjuntos de lncRNAs, como *PPP3CB*, *MAP3K14* e *DAPK1*. A análise de enriquecimento referente a estes lncRNAs mostrou associação com a via *MAPK*, que é relacionada a transformação maligna e metástase em câncer pancreático. Estes exemplos mostram o potencial papel regulatório dos lncRNAs e a necessidade de estudos adicionais para investigar a função de lncRNAs regulatórios e seu papel na tumorigênese e progressão do câncer de pâncreas.

A agressividade e heterogeneidade do câncer de pâncreas, assim como a alta taxa de mortalidade mostram a necessidade de maior entendimento da biologia tumoral para auxiliar tanto no diagnóstico como no tratamento de PDAC. Novos métodos que visam elucidar os componentes moleculares das vias de formação tumoral surgem como uma maneira alternativa para lidar com este desafio.

Recentemente foi publicado um estudo (RAPHAEL et al., 2017) de integração genômica no qual foi feita a investigação de subtipos de lncRNAs em tumores pancreáticos. Esses achados são evidência de que há uma possível associação entre esta classe de transcritos e os efeitos decorrentes da tumorigênese, reforçando portanto as premissas deste projeto.

## 2 OBJETIVOS

### 2.1 Objetivos Gerais

Implementação de uma estratégia computacional para identificação e anotação funcional de RNAs longos não codificadores associados a subtipos moleculares de PDAC.

### 2.2 Objetivos Específicos

1. Classificação de tumores pancreáticos em subtipos moleculares segundo o perfil de expressão gênica;
2. Análise da expressão diferencial para identificação de assinaturas de genes alterados nos diferentes subtipos moleculares de tumor;
3. Análise de enriquecimento de genes diferencialmente expressos para vias biológicas de interesse nos diferentes subtipos moleculares de tumor;
4. Criação de rede modular de co-expressão para avaliar conjuntos de genes que possam estar associados biologicamente;
5. Anotação das vias biológicas e marcadores prognósticos associados a cada um dos módulos gênicos;
6. Integração dos subtipos moleculares de PDAC aos módulos de co-expressão identificados e anotação de lncRNAs funcionalmente relevantes para a regulação da expressão gênica e definição dos diferentes fenótipos tumorais.

### 3 MATERIAIS E MÉTODOS

#### 3.1 Casuística

Foram utilizados dados de RNA-seq de 150 amostras de tumores primários de adenocarcinomas pancreáticos ductais disponíveis publicamente no The Cancer Genome Atlas (TCGA) (WEINSTEIN et al., 2013). Além disso, foram utilizados dados de RNA-seq de amostras pareadas de PDAC e tecido pancreático não tumoral de 14 indivíduos provenientes do banco de tumores do Hospital AC Camargo com celularidade tumoral variando de 30-100%, além de dados clínicos (sobrevida, idade, status, gênero e classificação anatomopatológica) associados para avaliação. Os experimentos referentes ao sequenciamento das amostras e validação experimental *in vitro* fazem parte da tese de doutorado de Vinícius Paixão. O processamento dos dados fazem parte da tese de doutorado de Julio Sosa. Estes projetos são de responsabilidade do Dr. Eduardo M. R. Reis.

#### 3.2 Aquisição e Processamento de Dados de Expressão Gênica

Os dados de expressão gênica de 150 amostras de Adenocarcinoma Pancreático Ductal sequenciados para transcritos poliadenilados (poly-A RNA-seq) foram obtidos a partir do banco de dados do TCGA (WEINSTEIN et al., 2013), dos quais os reads originais haviam sido previamente alinhados com a ferramenta RSEM (LI; DEWEY, 2011) tendo como referência o genoma humano do Gencode (versão GRCh37), e as contagens de transcritos sumarizadas por gene utilizadas para as análises subsequentes. Genes com menos de uma contagem por milhão (*Counts Per Million* - CPM) em pelo menos 20% das amostras foram filtradas. As diferenças de tamanho das bibliotecas entre as amostras sequenciadas foi corrigida pelo método TMM (*Trimmed Means of M-Values*), utilizando a biblioteca edgeR da linguagem R (RITCHIE et al., 2015). A normalização das contagens de transcritos em CPM foi efetuada com a biblioteca edgeR da linguagem R (RITCHIE et al., 2015), e em seguida os dados foram transformados para escala logarítmica em base dois.

### 3.3 Aquisição de Dados dos Níveis de Proteínas

Os dados dos níveis de proteína de 123 amostras de Adenocarcinoma Pancreático Ductal capturado por *Reverse Phase Protein Arrays* (RPPA) foram obtidos a partir do banco de dados do TCGA (WEINSTEIN et al., 2013).

### 3.4 Aquisição de Dados de Mutação

Os dados de anotação de mutações em formato maf (*mutation annotation format*) de 150 amostras de Adenocarcinoma Pancreático Ductal capturado por *Whole Exome Sequencing* (WES) foram obtidos a partir do banco de dados do TCGA (WEINSTEIN et al., 2013).

### 3.5 Classificação de Amostras Tumorais em Subtipos moleculares

A análise não supervisionada via NMF foi implementada para a classificação das amostras em subgrupos. Utilizando os dados de expressão gênica normalizados e transformados em escala logarítmica, os 2000 genes mais variáveis entre as amostras foram selecionados por meio de desvio absoluto da mediana (*Median Absolute Deviation* - MAD) para servir de entrada para a classificação não supervisionada.

Primeiramente, para determinar o número de subgrupos ideal, foram efetuadas vinte bateladas paralelamente com os mesmos dados de maneira iterativa até atingir a convergência, testando soluções de dois até sete possíveis subgrupos. Através do gráfico de coeficiente cofenético, foi efetuada a segunda parte, na qual foram rodadas 100 processos até a convergência para o valor ideal de  $k$ , e a atribuição das amostras em cada subgrupo determinada pela matriz consenso gerada pela solução da classificação por NMF. Foi utilizado como parâmetro de método para iteração o algoritmo proposto por Brunet em 2004 (GAUJOUX; SEOIGHE, 2010).

### 3.6 Identificação de Assinaturas Moleculares por Análise de Expressão Diferencial

Por uma limitação computacional, a identificação de subgrupos de tumores por NMF utilizou como entrada apenas os genes com maior variação entre as amostras. Para obter assinaturas moleculares características de cada um dos subgrupos de amostras identificados por NMF foi realizada uma análise de expressão diferencial de todos os genes expressos nas amostras, com a utilização da biblioteca limma da linguagem R (RITCHIE et al., 2015).

A análise foi efetuada par a par, comparando cada um dos subgrupos versus todos os outros, usando por critério de seleção genes que apresentassem um  $\log_2 FC > |1.2|$  e  $p\text{-valor} < 0,05$ . O número de genes obtidos, assim como sua intersecção foram visualizados com a biblioteca UpsetR da linguagem R (LEX et al., 2014). Uma análise de variância para verificar genes diferentes entre os quatro subgrupos foi efetuada com a biblioteca siggenes da linguagem R (SCHWENDER, 2012).

### 3.7 Análise de Enriquecimento para Vias Biológicas

Genes hiper-expressos em cada subgrupo foram avaliados segundo seu enriquecimento para vias biológicas significativamente alteradas ( $p\text{-valor} < 0.05$ ), segundo um teste hipergeométrico tendo como bases as categorias de Ontologia de Genes para Processos Biológicos (GO:BP) (ASHBURNER et al., 2000; CONSORTIUM, 2016) e a *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (KANEHISA; GOTO, 2000), utilizando o programa gProfiler implementado na biblioteca gprofileR da linguagem R (REIMAND et al., 2007).

### 3.8 Análise dos Componentes do Microambiente Tumoral

A análise do grau de infiltração de componentes do microambiente tumoral foi feita com a ferramenta ESTIMATE da linguagem R (YOSHIHARA et al., 2013). Essa ferramenta estima a fração dos componentes estromais e dos sistema imune através da análise de enriquecimento de vias biológicas para amostras individuais, utilizando como referência para a assinatura imune a sobreposição de genes de células hematopoiéticas com genes associados a infiltração de leucócitos no tecido tumoral. A assinatura estromal foi selecionada entre células não hematopoiéticas comparando a fração celular tumoral e a fração complementar estromal após extração a laser em dados de câncer de mama, colorretal e ovário.

### 3.9 Análise de redes de Co-expressão Gênica

A análise de conjuntos de genes com perfil de expressão correlato/associado foi efetuada com a biblioteca WGCNA da linguagem R (LANGFELDER; HORVATH, 2008). Os 8000 genes mais variáveis entre as amostras foram selecionados por MAD a partir da matriz de expressão de genes normalizada e transformada em escala logarítmica. A correlação de Pearson

do nível de expressão entre os genes foi utilizada para alimentar a função de potência de adjacência, definida por:

$$a = |cor(x_i, x_j)|^\beta$$

A matriz de adjacência obtida é então utilizada para detectar os conjuntos de genes (módulos) co-expressos através de agrupamento hierárquico por distância euclidiana acoplado à uma função de dissimilaridade topológica, calculada por:

$$TOM = \sum_u \frac{a_{iu}a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

do qual os ramos formados pelo dendograma correspondem aos módulos de genes co-expressos.

### 3.10 Análise de Sobrevida

Dados clínicos referentes às amostras de PDAC foram obtidos por meio da biblioteca RTCGA da linguagem R (KOSINSKI; BIECEK, 2016). Como etapa de pré-processamento, aplicamos uma censura à esquerda de 60 dias e uma censura à direita de 3 anos nos dados dos pacientes. Para a análise univariada (regressão de Cox), o potencial prognóstico de cada gene foi avaliado estratificando a expressão em duas classes em função da mediana.

Genes com  $p < 0,05$  são selecionados para outra análise, multivariada, comparando o potencial prognóstico da expressão com outras variáveis clínicas de interesse (gênero, idade, estadiamento, histórico de diabetes e de pancreatite crônica). As curvas de sobrevida foram geradas com a biblioteca *survminer* da linguagem R (KASSAMBARA, 2018). Diferenças estatisticamente significativas entre as curvas de sobrevida ( $p < 0,05$ ) foram avaliadas usando o teste de log rank implementado na biblioteca *survival* (THERNEAU; LUMLEY, 2018).

### 3.11 Anotação estrutural e funcional de lncRNAs

A análise de predição de estrutura secundária e anotação de interação com proteínas foi feita pela consulta no banco de dados *lnc2Catlas* (REN et al., 2018). Para efetuar o cálculo

de predição de interação lncRNA-proteína, a base *lnc2Catlas* usa a ferramenta *Global Score* de Cirillo e colaboradores (CIRILLO et al., 2017), que integra a informação de estrutura local de proteínas e RNAs em um modelo de propensão geral de ligação, tendo sido calibrada com dados experimentais e otimizada para lidar com transcritos longos. A predição de possíveis interações RNA-RNA foi feita pelo catálogo obtido da base de Terai em colaboradores (TERAI et al., 2016). A anotação de potencial codificador e conservação evolutiva foi feita através da base de dados *LNCipedia* (VOLDERS et al., 2015). A busca do envolvimento de lncRNAs de interesse em outros estudos de câncer foi feita com a ferramenta *lnc2cancer* (NING et al., 2016).

## 4 RESULTADOS

### 4.1 Classificação Não Supervisionada de Tumores Pancreáticos em Subgrupos moleculares

Para identificar RNAs longos não codificadores no contexto da heterogeneidade molecular existente em tumores primários de Adenocarcinomas Pancreáticos Ductais, foi executada a classificação das amostras em subgrupos distintos a partir de diferenças de expressão gênica. A classificação não supervisionada por NMF utilizando os 2000 genes com maior variância entre as amostras resultou em uma solução para dois ou quatro possíveis subgrupos/classes distintos, sendo que para obter a maior estratificação possível das amostras, foi escolhida a solução para quatro subgrupos.

A avaliação da qualidade do modelo pode ser feita através de dois critérios: o coeficiente cofenético, que é uma medida da correlação entre a matriz consenso da distribuição das amostras em seus respectivos grupos e a distribuição das amostras em uma matriz consenso após o reordenamento por agrupamento hierárquico aglomerativo.

Os valores de coeficiente cofenético podem variar entre zero a um, sendo que valores mais próximos de “um” indicam soluções mais consistentes, robustas e significativas. o que também pode ser verificado na matriz consenso das amostras em seus respectivos grupos. O gráfico do coeficiente cofenético (**Figura 5A**) mostra, a partir da iteração para possíveis soluções de dois até sete  $k$  subgrupos, que há um pico para  $k$  em dois e quatro subgrupos.

Com a finalidade de obter maior estratificação dos dados e maior repertório de informação/contexto biológico, foi selecionada a solução que resultou em **quatro subgrupos distintos**. A matriz consenso mostrando a distribuição das amostras para cada solução em  $k$  subgrupos (**Figura 5B**) permite a inspeção visual de que a solução em quatro subgrupos é bem robusta, corroborando o resultado mostrado pelo coeficiente cofenético.

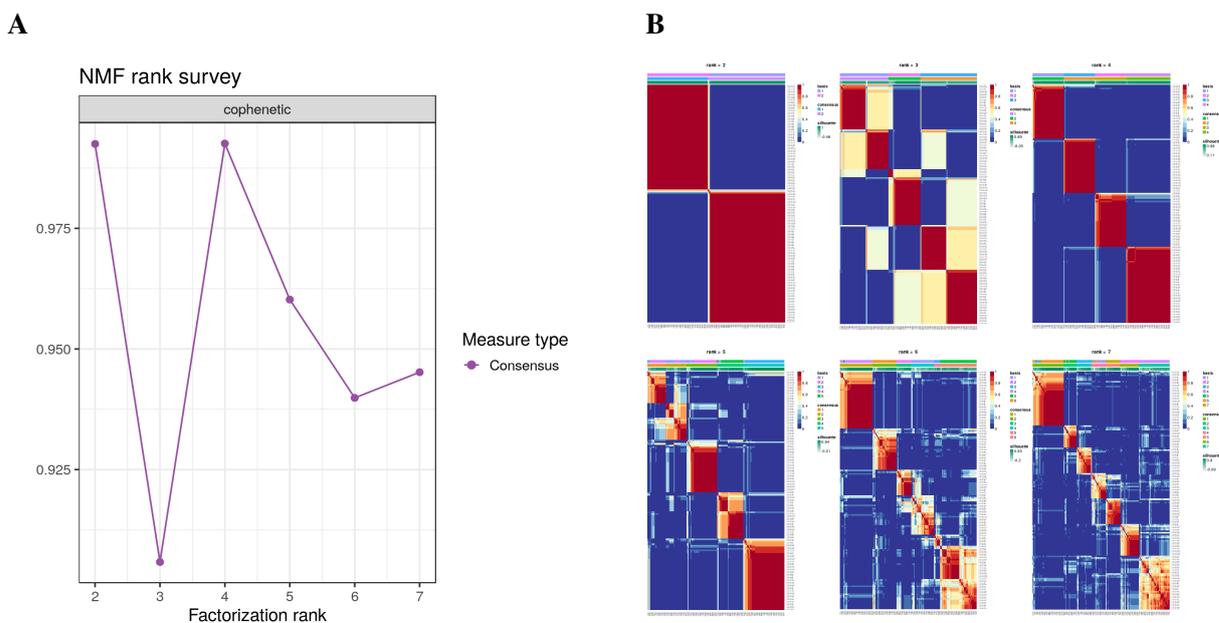


Figura 5 – Os 2000 genes com expressão mais variável (por desvio absoluto da mediana) foram utilizados como entrada para a classificação não supervisionada por NMF. **A.** Resultado gráfico da iteração de dois a sete possíveis grupos/classes. O coeficiente de correlação cofenética (gráfico superior esquerdo) foi utilizado como um indicador quantitativo da estabilidade do agrupamento obtido a partir das possíveis soluções, onde valores mais próximos de 1 refletem grupos mais consistentes. Para maior resolução, foi escolhida a solução que considera quatro subgrupos. **B.** Representação gráfica da matriz consenso mostrando o agrupamento das amostras para diferentes soluções de número de grupos.

#### 4.2 Identificação de Assinaturas Moleculares Associadas a Subtipos Moleculares de PDAC

A investigação de assinaturas moleculares características dos subgrupos foi iniciada por uma análise de variância, que identificou **4437** genes diferencialmente expressos em pelo menos um dos 4 subgrupos de amostras (**Figura 6A**), dentre os quais **60** são lncRNAs (**Figura 6B**).

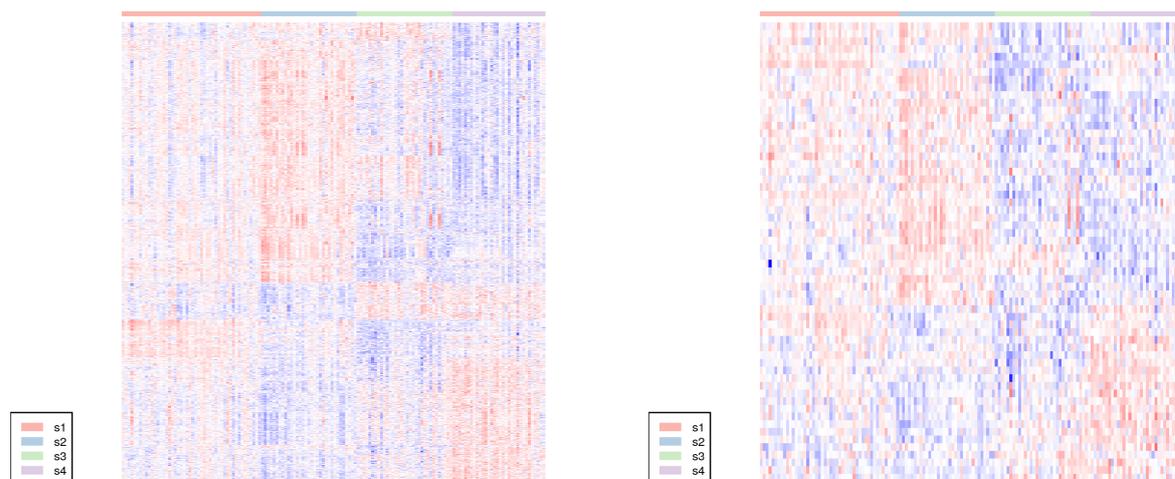
**A****B**

Figura 6 – **A.** “Heatmap” com abundância relativa dos 4437 genes (linhas) diferencialmente expressos entre os quatro subgrupos de amostras (colunas) resultantes da análise por NMF. **B.** Dos genes representados em **A.**, 60 lncRNAs. Para cada gene a expressão está representada em valores de desvio-padrão (*Z-score*) de todas as amostras.

Além disso, a assinatura de expressão referente a cada subgrupo foi identificada pela comparação de cada um dos grupos em relação a todos os outros, e o número de genes hiperexpressos sumarizado na **Figura 7B**.

A anotação dos lncRNAs com expressão significativamente distinta ( $p < 0,05$ ) em cada um dos subgrupos (**51 ao total**) é apresentada na **Tabela 1**.

Tabela 1 – Lista de lncRNAs diferencialmente expressos nos subgrupos.

Gene	logFC	P-valor	P-ajustado	Biotipo	Subgrupo
LOC388387	4,701	2.29E-17	7.41E-15	lincRNA	1
FLJ42875	1,516	5.97E-06	3.89E-04	antisense	1
C20orf56	1,503	1.25E-03	4.69E-02	lincRNA	1
C17orf73	-1,488	1.45E+00	1.86E+01	processed_transcript	1
FLJ43390	3,000	2.74E-07	6.62E-06	lincRNA	2
LOC255167	1,980	8.58E-09	3.85E-07	lincRNA	2
LOC284749	1,616	1.17E-01	4.92E-01	lincRNA	2
LOC146336	1,521	1.26E+00	3.81E+00	processed_transcript	2

LOC100188949	1,503	1.68E-06	3.12E-05	lincRNA	2
LOC100190938	1,487	4.13E-12	8.32E-10	lincRNA	2
ADAM6	1,402	1.26E-01	5.26E-01	lincRNA	2
C14orf139	1,349	1.74E-11	2.81E-09	lincRNA	2
LOC283663	1,290	3.50E-02	1.74E-01	lincRNA	2
LOC145820	1,271	1.98E-04	1.94E-03	lincRNA	2
SNORD116-4	1,266	1.57E+01	3.28E+01	processed_transcript	2
SPON1	1,252	2.71E-06	4.71E-05	processed_transcript	2
PAR5	1,251	1.08E-01	4.59E-01	sense_intronic	2
LOC96610	1,239	2.24E-01	8.64E-01	processed_transcript	2
LOC100233209	1,200	1.06E-03	8.23E-03	processed_transcript	2
C17orf55	-1,214	4.85E-02	2.31E-01	lincRNA	2
C14orf34	-1,338	1.59E+00	4.62E+00	processed_transcript	2
WBSCR26	-1,672	1.63E-06	3.04E-05	antisense	2
LOC100131726	-1,863	1.27E-01	5.29E-01	antisense	2
LOC100131726	2,312	2.68E-03	4.90E-02	antisense	3
DKFZp434J0226	1,403	1.09E-02	1.53E-01	processed_transcript	3
LOC255167	-1,218	1.03E-02	1.46E-01	lincRNA	3
FLJ42875	-1,335	3.37E-03	5.90E-02	antisense	3
NCRNA00092	-1,471	2.91E-09	2.58E-07	lincRNA	3
C17orf73	-1,586	2.56E+00	1.34E+01	processed_transcript	3
LOC338651	-1,620	1.38E-03	2.74E-02	antisense	3
LOC389332	-2,348	2.47E-09	2.26E-07	lincRNA	3
LOC146336	-2,551	1.56E-04	4.16E-03	processed_transcript	3
C20orf56	-2,886	1.09E-11	2.26E-09	lincRNA	3
FLJ43390	-3,391	2.28E-09	2.11E-07	lincRNA	3
LOC388387	-3,762	2.17E-09	2.03E-07	lincRNA	3
C17orf73	3,796	2.07E-08	1.11E-06	processed_transcript	4
LOC338651	1,667	9.07E-04	7.32E-03	antisense	4
WBSCR26	1,578	1.39E-05	2.20E-04	antisense	4
HOTAIR	1,449	1.80E+01	3.62E+01	antisense	4

ABO	1,445	1.67E-02	8.96E-02	processed_transcript	4
LOC389332	1,401	1.27E-02	7.11E-02	lincRNA	4
C14orf34	1,369	1.39E+00	4.02E+00	processed_transcript	4
LOC100127888	1,364	1.49E-02	8.13E-02	antisense	4
LOC150197	1,332	5.02E-04	4.44E-03	lincRNA	4
HOXA11AS	1,310	1.31E+01	2.75E+01	antisense	4
SLC44A4	1,212	1.20E-05	1.94E-04	processed_transcript	4
LOC255167	-1,244	8.34E-03	4.91E-02	lincRNA	4
PAR5	-1,278	9.07E-02	3.82E-01	sense_intronic	4
SPON1	-1,347	3.25E-07	9.72E-06	processed_transcript	4
SNORD116-4	-1,370	9.78E+00	2.15E+01	processed_transcript	4
H19	-1,459	4.60E-07	1.27E-05	processed_transcript	4

#### 4.3 Anotação Funcional das Assinaturas de Genes Associada aos Subgrupos

A análise de enriquecimento para vias biológicas alteradas foi efetuada com uso das bases curadas de Ontologia de Genes para Processos Biológicos e Vias KEGG, tendo como intuito avaliar o contexto biológico dos genes diferencialmente expressos identificados nos subgrupos de tumores pancreáticos. Algumas das vias enriquecidas estão listadas na **Tabela 2**.

Tabela 2 – Vias biológicas enriquecidas por subgrupo identificado.

P-valor	Identificador	Domínio	Termo	Módulo
0	GO:0046879	BP	hormone secretion	s1
0	GO:0051046	BP	regulation of secretion	s1
0	GO:0009306	BP	protein secretion	s1
0	GO:0030073	BP	insulin secretion	s1
0	GO:0061041	BP	regulation of wound healing	s1
0	GO:0007159	BP	leukocyte cell-cell adhesion	s2
0	GO:0006954	BP	inflammatory response	s2
0	GO:1903037	BP	regulation of leukocyte cell-cell adhesion	s2
0	GO:0046649	BP	lymphocyte activation	s2
0	GO:0046651	BP	lymphocyte proliferation	s2

0	GO:0072676	BP	lymphocyte migration	s2
0	GO:0001816	BP	cytokine production	s2
0	GO:0051897	BP	positive regulation of protein kinase B signaling	s2
0	GO:0008544	BP	epidermis development	s3
0	GO:0043588	BP	skin development	s3
0	GO:0030855	BP	epithelial cell differentiation	s3
0	GO:0009913	BP	epidermal cell differentiation	s3
0	GO:0030216	BP	keratinocyte differentiation	s3
0	GO:0030198	BP	extracellular matrix organization	s3
0	GO:0019752	BP	carboxylic acid metabolic process	s4
0	GO:0001676	BP	long-chain fatty acid metabolic process	s4
0	GO:0019369	BP	arachidonic acid metabolic process	s4
0	GO:0019373	BP	epoxygenase P450 pathway	s4
0	GO:0032536	BP	regulation of cell projection size	s4
0	GO:0009410	BP	response to xenobiotic stimulus	s4
0	GO:0055114	BP	oxidation-reduction process	s4

Segundo a descrição dos fenótipos identificados por Bailey e colaboradores (BAILEY et al., 2016), o subtipo escamoso foi reconhecido pela hiper expressão de *TP63*, no com vias envolvidas em processos de inflamação, resposta a hipoxia, reprogramação metabólica, sinalização de *TGF- $\beta$*  e ativação da via de *MYC*. O subtipo progenitor possui alteração em *PDX1*, com vias associadas a oxidação de ácidos graxos, biogênese de hormônios esteróides, metabolismo de drogas e glicosilação de O-ligada de mucinas. Encontra-se também alteração nas apomucinas *MUC5AC* e *MUC1*, mas não *MUC2* ou *MUC6*, que aparecem como coexpressas no subgrupo progenitor.

O subtipo Endócrino/Exócrino é definido como uma subclasse do subtipo progenitor, com hiper-regulação de fatores de transcrição como *NR5A2*, *MIST1* (também conhecido como *BH-LHA15A*) e *RBPJL*, assim como seus alvos das vias à jusante. Há também alteração de genes relacionados com diferenciação endócrina como *INS*, *NEUROD1*, *NKX2-2* e *MAFA*, assim como de genes relacionados à diferenciação terminal do tecido pancreático como *AMY2B*, *PRSSI*, *PRSS3*, *CEL* e *INS*.

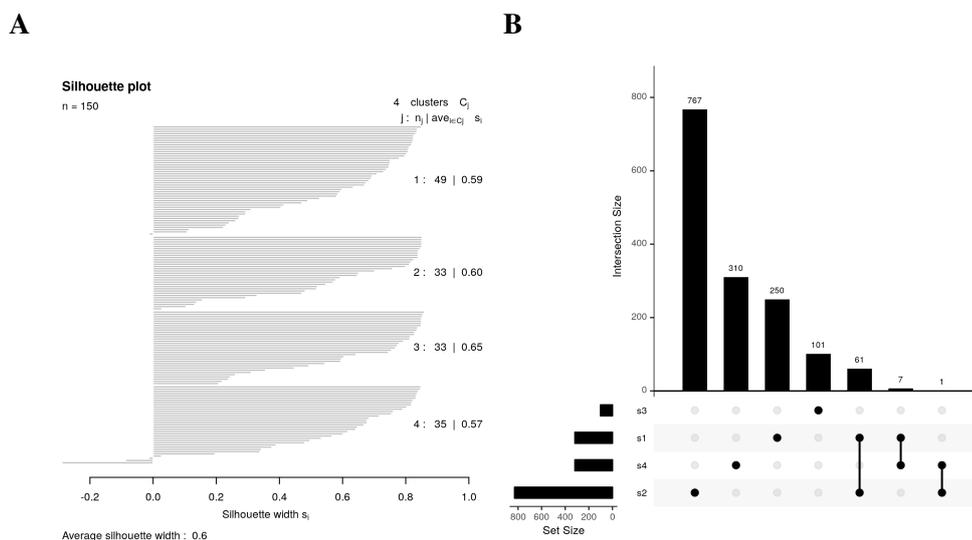


Figura 7 – **A.** Silhueta com genes representativos de cada um dos quatro subgrupos de amostras obtidos por NMF. Amostras com valores de silhueta menores que zero foram excluídas para as análises de expressão diferencial. **B.** A análise de expressão diferencial comparando cada um dos grupos versus todos os outros. As barras laterais à esquerda indicam o número total de genes hiperexpressos em cada grupo, e as barras verticais indicam os subconjuntos de genes que são exclusivos do grupo (círculos na legenda) ou que possuem sobreposição (linhas ligando os círculos) entre os grupos definidos. Como critérios para considerar um gene como diferencialmente expresso em um grupo foi utilizado  $\log_2 FC > |1, 2|$  e  $p - \text{valor} < 0, 05$ .

O subtipo Imunogênico possui grande sobreposição em características com a classe Progenitora, com a adição de evidências significativas de infiltrado imune. Possui alteração em vias associadas ao sistema imune como sinalização de células B, apresentação de antígenos, assim como vias de células T  $CD4^+$ , T  $CD8^+$  e do receptor *Toll-like*. O perfil de expressão está relacionado ao de infiltração de células B e T, tanto de células citotóxicas  $CD8^+$  e células T regulatórias ( $CD4^+ CD25^+ FOXP3^+ T_{regs}$ ), com hiper-regulação de *CTLA* e *PDI*, associados ao fenótipo de evasão adquirida do sistema imune.

Portanto, os genes diferencialmente expressos e as vias identificadas neste estudo recapitulam os fenótipos identificados por Bailey e colaboradores (BAILEY et al., 2016), corroborando a análise feita em um conjunto de dados distintos e reforçando as informações geradas pelas assinaturas encontradas para serem utilizadas em um estudo mais aprofundado da associação de RNAs não codificadores aos subtipos identificados. A partir da sobreposição observada entre as categorias enriquecidas em cada um dos 4 subgrupos definidos neste trabalho e as categorias descritas no estudo de Bailey e colaboradores, foi atribuído o subtipo molecular que melhor representa os subtipos identificados neste trabalho: s1: subtipo Exócrino, s2: subtipo Imunogênico, s3: subtipo Escamoso e s4: subtipo Progenitor.

#### 4.4 Análise do Microambiente Tumoral dos Subgrupos de PDAC

A análise de enriquecimento para assinaturas de componentes do estroma e do sistema imune servem dois propósitos: avaliar quantitativamente a composição de componentes do microambiente nas amostras e, de maneira indireta, estimar o grau de celularidade tumoral.

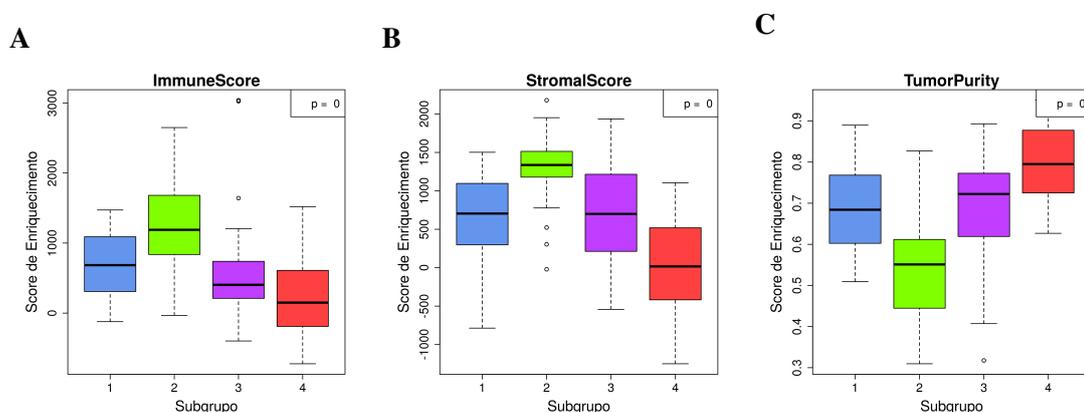


Figura 8 – Avaliação do enriquecimento para componentes do microambiente tumoral. Utilizando a ferramenta *ESTIMATE*, foi possível estimar de maneira quantitativa o grau de enriquecimento para células estromais e do sistema imune e, também, a pureza tumoral das amostras estratificadas por subgrupo identificado.

A **Figura 8** mostra o escore de enriquecimento obtido para cada uma das assinaturas de componentes imunes, estromais e a fração de celularidade, estratificado por subgrupo. O subgrupo 2, de caráter “Imunogênico”, possui maior enriquecimento para a assinatura imune, o que corrobora as vias enriquecidas pela classificação não supervisionada, indicando maior infiltração de células do sistema imune no microambiente tumoral. Além disso, os subgrupos dois (Imunogênico) e três (de fenótipo “Escamoso”) estão mais enriquecidos para componentes estromais.

A estimativa de infiltração dos componentes estromais e do sistema imune estratificado para os subgrupos nas amostras do TCGA assemelha-se com a estimativa anotada por Bailey e colaboradores em um conjunto de dados distinto, reforçando a robustez da classificação e a estimativa de infiltração dos componentes do microambiente tumoral nos subgrupos identificados em PDAC.

#### 4.5 Análise de rede de Coexpressão Gênica

A abordagem de estudar redes de genes que atuam em conjunto é de grande interesse por permitir resultados mais informativos quanto ao contexto biológico da análise em questão que

não é possível com a análise de componentes isolados. Para o caso de redes de coexpressão, o padrão de expressão dos genes pode ser usado para a determinar o grau de conectividade entre os genes. Tendo como entrada a matriz de expressão dos 8000 genes mais variáveis, primeiro foi efetuada a estimativa do valor ideal de  $\beta$  (potência da função de adjacência) a ser utilizado.

A determinação foi feita pela iteração de valores de possíveis valores de  $\beta$  e verificação do valor correspondente de  $R^2$ , que refere-se ao índice de **topologia livre de escala**. Quando mais próximo de um, mais livre de escala se torna a rede. Na prática, o ideal é a escolha de um valor de  $\beta$  no qual a angulação da curva começa a se aproximar de zero, valor representado pela linha horizontal na **Figura 9A**, sendo que, neste caso, foi selecionado um valor de  $\beta = 10$ .

Ao mesmo tempo, é importante verificar a conectividade média obtida para cada valor de  $\beta$  (**Figura 9B**), sendo que valores muito baixos podem levar a módulos contendo poucos genes conectados. A determinação do valor de  $\beta$  permite calcular a matriz de adjacência, que pode então ser utilizada para a determinação da matriz de dissimilaridade topológica, que contém a diferença de conectividade entre os genes.

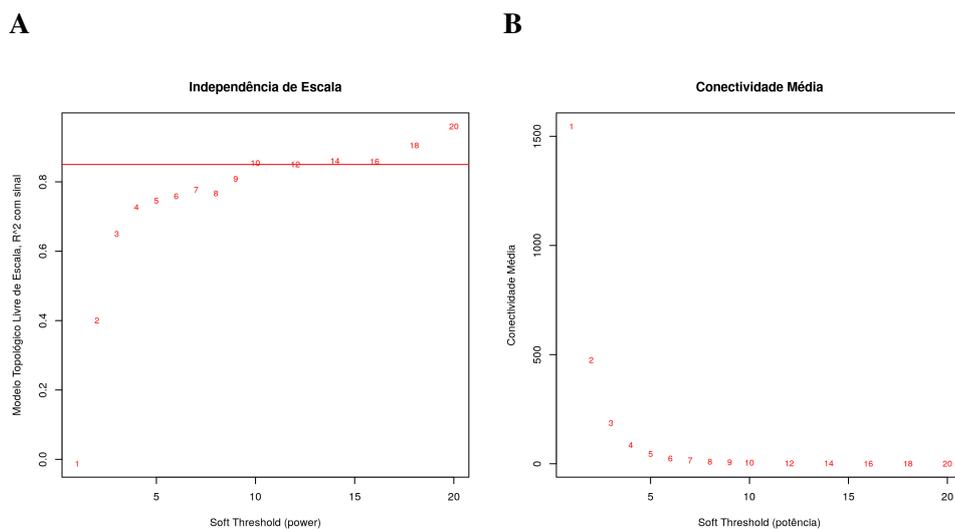


Figura 9 – **A**. Determinação dos valores ideais dos parâmetros para utilização na função de adjacência e determinação dos módulos. Aqui foi escolhido o valor de  $\beta = 10$ , que representa o menor valor de potência obtido quando a curva atinge um platô. **B**. Mensuração da conectividade média estimada para cada valor de Beta correspondente. A obtenção de módulos significativos depende da conectividade entre os genes dentro de um determinado módulo (conectividade intra-modular).

O heatmap da matriz de dissimilaridade topológica (**Figura 10A**) permite a visualização dos conjuntos de genes (módulos) identificados, um total de 13 módulos. O dendograma resultante do agrupamento hierárquico aglomerativo dos genes da matriz de dissimilaridade leva à determinação dos módulos mais conexos de genes coexpressos (**Figura 10B**).

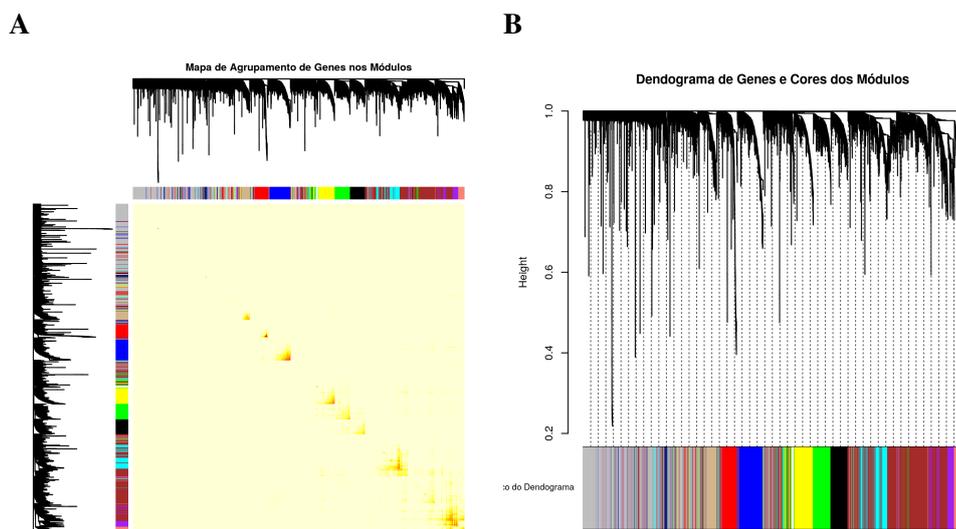


Figura 10 – **A.** Matriz de dissimilaridade topológica. A distância medida por agrupamento hierárquico aglomerativo do grau de conectividade entre os genes permite a visualização de módulos de genes coexpressos. **B.** Cada linha do dendrograma representa um gene, sendo que os ramos delimitam os módulos identificados, os quais são representados por cores distintas (barra inferior). A identificação de módulos de genes permite maior grau de contextualização dos diversos processos biológicos operantes e sua associação com os subgrupos resultantes da classificação não supervisionada por NMF.

#### 4.6 Contexto Biológico e Clínico dos Módulos de Coexpressão

A avaliação da significância biológica dos módulos de genes coexpressos foi feita primeiramente pela análise de enriquecimento para vias biológicas dos genes de cada um dos módulos. As vias significativamente identificadas para Ontologia de Genes de Processos Biológicos e Vias KEGG foram anotadas, permitindo identificar os processos celulares preferencialmente associados a cada um dos módulos de coexpressão.

Além disso, foi avaliado o enriquecimento dos módulos para genes com potencial prognóstico em função da expressão gênica. Para essa finalidade, todos os genes presentes nos módulos foram testados quanto a sua associação com a sobrevivência dos pacientes, por meio de uma análise univariada de regressão de Cox em função da expressão. Módulos que possuem genes com maior conectividade e menores p-valores e, conseqüentemente, de maior confiança quanto ao seu potencial preditivo, são mais enriquecidos para possíveis marcadores prognósticos, como pode ser visto na **Figura 11**.

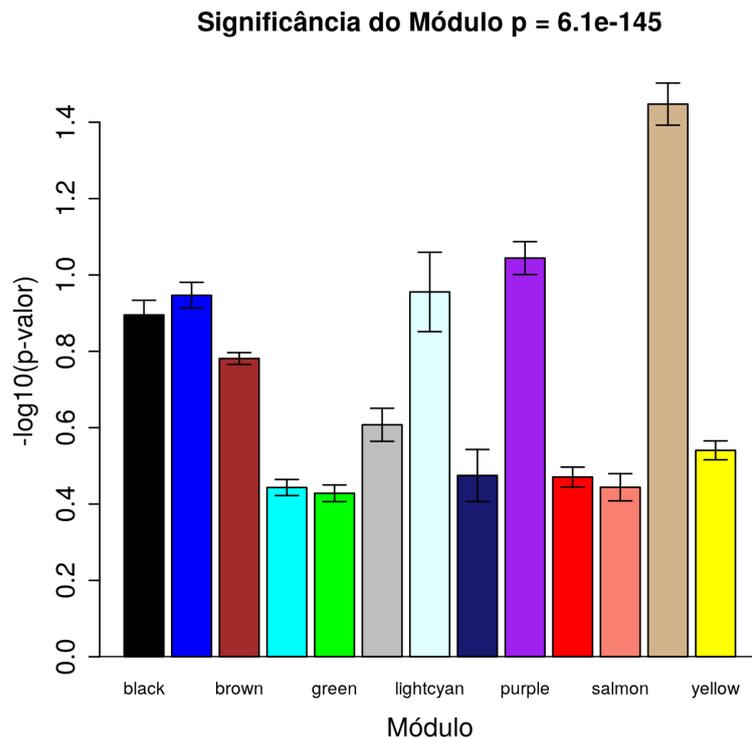


Figura 11 – Significância dos módulos de coexpressão. Cada os genes foi testados quanto ao seu potencial prognóstico por meio de análise de regressão de Cox. Neste contexto, módulos com maior número de genes significativos são mais enriquecidos para possíveis marcadores prognósticos.

As curvas de Kaplan Meier da **Figura 12** representam as taxas de sobrevida de pacientes com PDAC em função da expressão para dois genes exemplares do módulos *Brown* (*BOC* e *ZEB2*) e *Tan* (*AURKA* e *TPX2*, respectivamente, os quais estão listados na Tabela 3 e que foram selecionados para visualização.

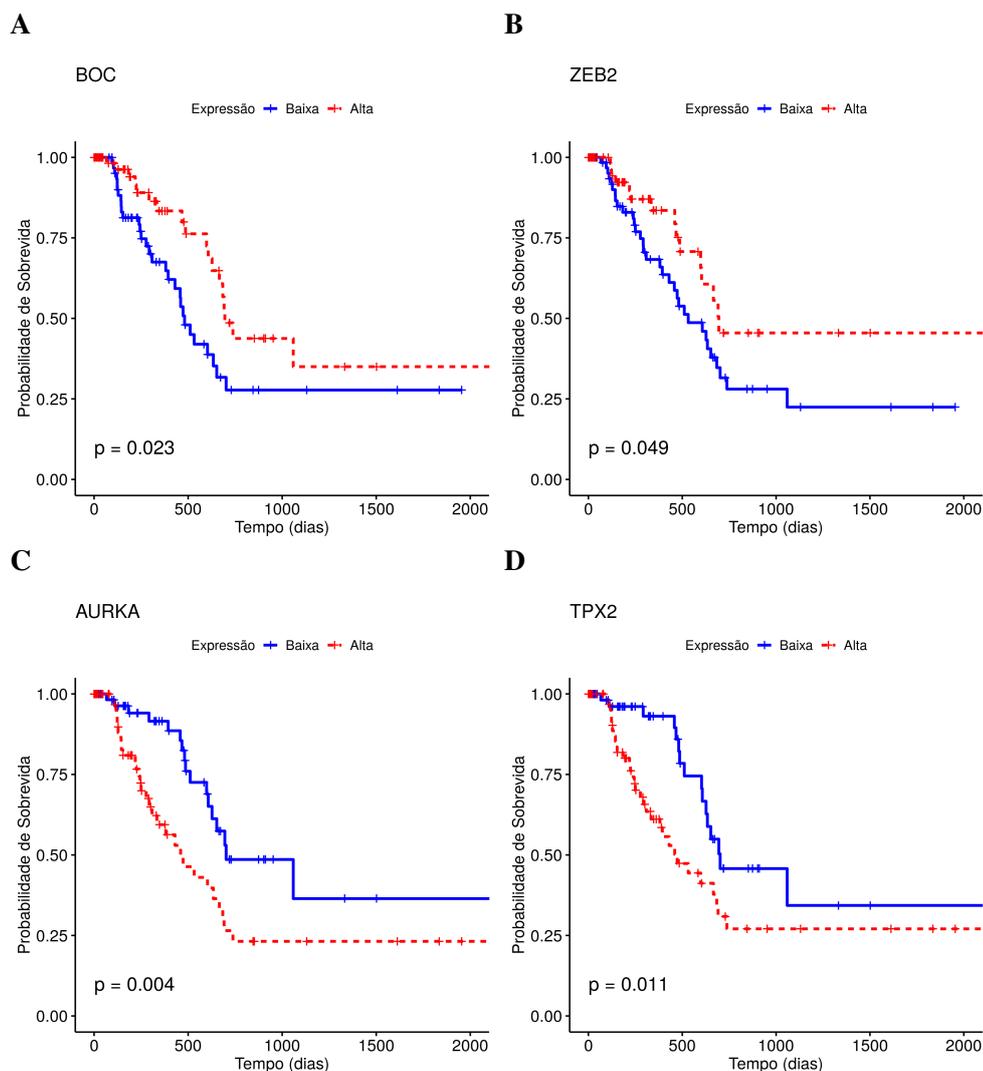


Figura 12 – Curva de Kaplan Meier para representar a taxa de sobrevivência de pacientes com PDAC em função da expressão. Dois genes exemplares dos módulos enriquecidos (*Brown* e *Tan*) e listados na Tabela 3 foram selecionados para visualização. Para evitar a inserção de grandes flutuações dos dados clínicos na avaliação foi aplicada uma censura à esquerda de 60 dias e à direita de 3 anos.

Os genes significativos provenientes da análise univariada ( $p < 0,05$ ) foram selecionados para uma análise multivariada, comparando o potencial prognóstico em função da expressão com outras variáveis clínicas comumente avaliadas, como gênero, idade, estadiamento, grau histológico, histórico de diabetes e de pancreatite crônica. Os genes anotados como significativos na análise univariada, assim como o p-valor do teste de *log-rank* da análise multivariada constam na **Tabela 3**.

Tabela 3 – regressão univariada e multivariada de Cox.

P-valor	Identificador	Domínio	Termo	Módulo
0	GO:0046879	BP	hormone secretion	s1

0	GO:0051046	BP	regulation of secretion	s1
0	GO:0009306	BP	protein secretion	s1
0	GO:0030073	BP	insulin secretion	s1
0	GO:0061041	BP	regulation of wound healing	s1
0	GO:0007159	BP	leukocyte cell-cell adhesion	s2
0	GO:0006954	BP	inflammatory response	s2
0	GO:1903037	BP	regulation of leukocyte cell-cell adhesion	s2
0	GO:0046649	BP	lymphocyte activation	s2
0	GO:0046651	BP	lymphocyte proliferation	s2
0	GO:0072676	BP	lymphocyte migration	s2
0	GO:0001816	BP	cytokine production	s2
0	GO:0051897	BP	positive regulation of protein kinase B signaling	s2
0	GO:0008544	BP	epidermis development	s3
0	GO:0043588	BP	skin development	s3
0	GO:0030855	BP	epithelial cell differentiation	s3
0	GO:0009913	BP	epidermal cell differentiation	s3
0	GO:0030216	BP	keratinocyte differentiation	s3
0	GO:0030198	BP	extracellular matrix organization	s3
0	GO:0019752	BP	carboxylic acid metabolic process	s4
0	GO:0001676	BP	long-chain fatty acid metabolic process	s4
0	GO:0019369	BP	arachidonic acid metabolic process	s4
0	GO:0019373	BP	epoxygenase P450 pathway	s4
0	GO:0032536	BP	regulation of cell projection size	s4
0	GO:0009410	BP	response to xenobiotic stimulus	s4
0	GO:0055114	BP	oxidation-reduction process	s4

---

#### 4.7 Integração de Dados de coexpressão gênica com subtipos moleculares de PDAC

A integração dos dados dos módulos de coexpressão gênica com os subgrupos de amostras obtidos pela classificação não supervisionada pode ser vista na **Figura 13**, e a sumarização das informações encontradas se encontra na **Tabela 4**.

### Correlação de Módulos com Subgrupos

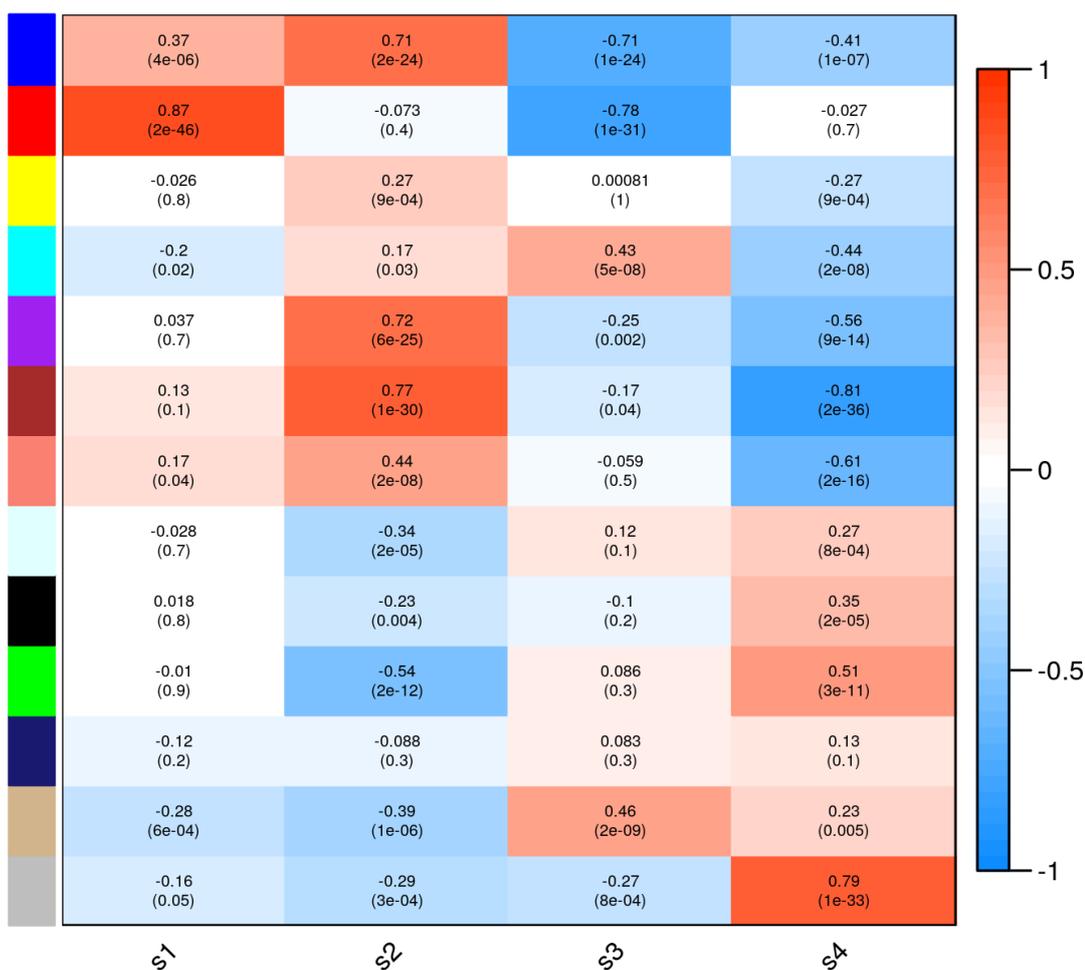


Figura 13 – Integração de dados dos subgrupos de amostras com os módulos de genes.

O módulo *Red*, mais associado ao subgrupo 1, apresenta vias enriquecidas de caráter exócrino: secreção de suco pancreático e de bile e aumento de metabolismo de ácidos graxos. O subgrupo 2 apresenta associação forte com os módulos *Brown*, com função imunogênica como ativação de células T, proliferação de leucócitos, resposta pró-inflamatória, e o módulo *Blue*, de função neuroendócrina: regulação de secreção hormonal, diferenciação de células A, alteração de sinalização neuronal.

O subgrupo 3 encontra-se associado ao módulo *Tan*, de funções de reparo de DNA e apoptose, representando alterações em processos de proliferação de massa celular interna, fosforilação de histonas, manutenção telomérica, alteração no maquinário de manutenção por excisão de bases e por recombinação homóloga. O módulo *Grey* é caracterizado pela baixa conectivi-

dade e, teoricamente, menos informativo. Sua associação com o subgrupo 4, no entanto, está enriquecido para função do tipo progenitora, como reparo tecidual (*wound healing*), morfogênese, diferenciação celular e síntese de mucinas tipo O-glicanas, é concordante com o fenótipo atribuído para este subgrupo.

Tabela 4 – Sumarização das alterações identificadas em PDAC.

Subgrupo	Módulo	Principais Funções/Vias dos Módulos	% KRAS mutado	Carga mutacional	Número de lncRNAs
s1	Red	Funções Exócrinas, Secreção Pancreática	82	173.8776	4
s2	Brown	Ativação, Proliferação e Migração de Células Imunes	79	704.8485	19
s3	Tan	Reparo de DNA, Progressão do Ciclo Celular	91	200.3333	12
s4	Grey	Morfogênese, Reparo Tecidual, Síntese de Mucinas	89	207.4286	16

#### 4.8 Anotação de lncRNAs associados com subtipos moleculares de PDAC

A identificação de RNAs longos não codificadores associados a subtipos moleculares distintos possibilitou seguir com a caracterização em nível de interação com outros RNAs, proteínas, anotações prévias da literatura, assim podendo inferir, ainda que de maneira especulativa, seu papel funcional dentro do contexto biológico no qual foram identificados.

Neste estudo chamou a atenção o subgrupo 3, de fenótipo “Escamoso”, por apresentar maior enriquecimento para genes com potencial prognóstico e por estar associado a vias de reparo de DNA e apoptose, além de expressão alterada do gene *TP63*, importante supressor tumoral em PDAC (BAILEY et al., 2016; SOMERVILLE et al., 2018). Neste subgrupo, o antissenso ***LOC100131726 (FAM83A-AS1)*** e o RNA não codificador não caracterizado ***DKFZp434J0226*** foram identificados como hiper expressos.

A análise de interação com proteínas frequentemente alteradas em câncer, segundo consulta na base de dados *lnc2catlas*, mostrou que dos dois lncRNAs, somente o lncRNA *FAM83A-AS1* possui afinidade por e possível interação com proteínas regulatórias como *FGFR2*, *AXIN1*, *PTEN*, *BRAF*, *SMAD4*, *TGFBR2*, *TP53* e *CDKN2A*, fundamentais na regulação de processos oncogênicos. No contexto de carcinoma hepatocelular *FAM83A-AS1* também foi descrito como significativamente alterado (ZHU et al., 2014).

A avaliação dos lncRNAs com expressão mais baixa no subgrupo 3 em relação aos outros subgrupos permitiu identificar possíveis interações com diversas proteínas sendo que interações mais recorrentes foram com as proteínas *BRCA2*, *AXIN1*, *MET*, *SMAD4*, e *TP53*, que estão

associadas com processos de reparo de DNA, apoptose e transição epitélio mesênquima, corroborando as principais funções alteradas neste subgrupo, assim como seu fenótipo característico.

A quantificação dessas proteínas foi avaliada em função da estratificação nos subgrupos identificados, e os níveis da proteína BRCA2, com a afinidade mais alta pelos lncRNAs hipo-expressos no subgrupo 3 visualizada no gráfico da **Figura 14**.

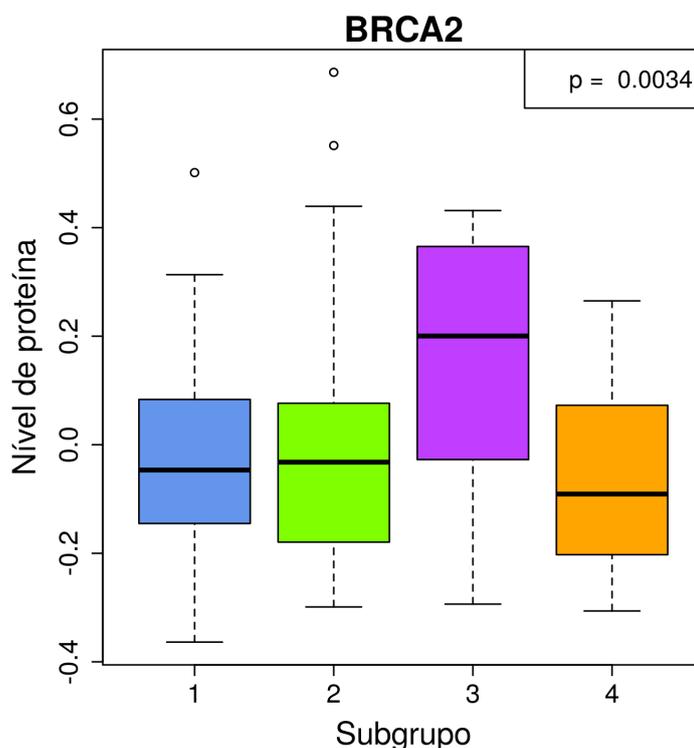


Figura 14 – Nível de abundância da proteína BRCA2 estratificada por subgrupo. A abundância do produto proteico é significativamente maior no subgrupo 3, de fenótipo escamoso, o que está associado a menor expressão de lncRNAs com alta afinidade por esta proteína, representando um possível papel de complexo regulador desempenhado por RNAs não codificadores.

A **Figura 15** mostra a curva de sobrevida de pacientes estratificada por subgrupo que, apesar de não significativa ( $p = 0,086$ ), reproduz a tendência prognóstica em função dos fenótipos identificados e em acordo com os subtipos identificados e descritos por Bailey e colaboradores (BAILEY et al., 2016).

As análises de Raphael e colaboradores (RAPHAEL et al., 2017) apontaram para significância em contexto clínico dos dados somente em um subgrupo das amostras de pureza tumoral mais alta (definido in silico) de todo o conjunto de dados e estratificadas em função dos níveis de de proteína, ao passo que não foi relatado significância clínica em função da classificação em subgrupos através dos níveis de expressão gênica, indicando consistência com os resultados

apontados neste estudo.

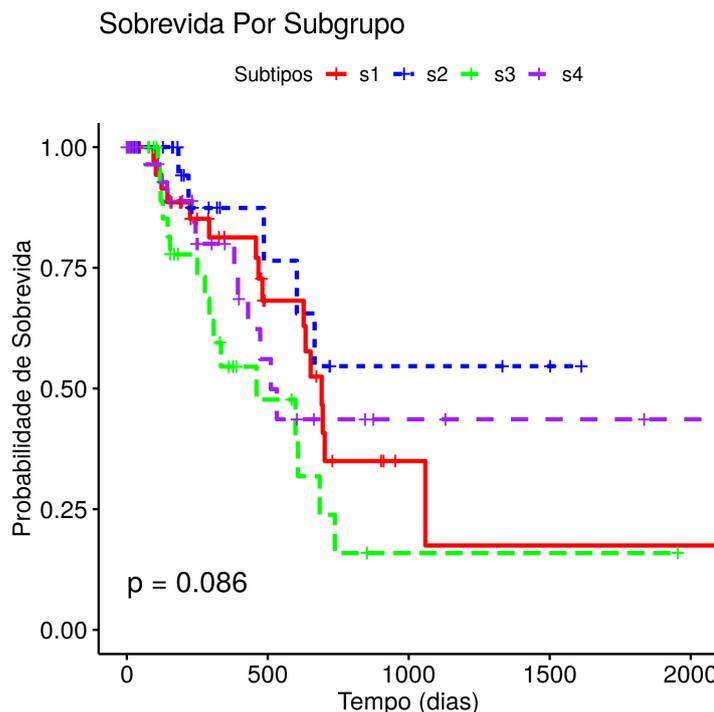


Figura 15 – Curva de Kaplan Meier representando a taxa de sobrevida de pacientes com PDAC em função de cada um dos subgrupos identificados. Apesar do resultado estratificado por subgrupo de pacientes não ser significativo, a tendência prognóstica é similar à apontada na literatura por Bailey e colaboradores (BAILEY et al., 2016). Para evitar a inserção de grandes flutuações dos dados clínicos na avaliação foi aplicada uma censura à esquerda de 60 dias e à direita de 3 anos.

#### 4.9 Análise dos lncRNAs Identificados em Amostras do AC Camargo

Os lncRNAs anotados como diferencialmente expressos entre os subgrupos foram avaliados em dados de RNA-seq de amostras sequenciadas em nosso laboratório. A **Figura 16** contém a visualização do perfil de expressão dos RNAs não codificadores diferencialmente expressos no subgrupo 3 em função dos outros subgrupos identificados. Dentre os lncRNAs mostrados na Figura 16, *LINC00261* ( $\log FC = -2,346$ ), *LINC00982* ( $\log FC = -1,640$ ), *LINC00483* ( $\log FC = 2,808$ ) e *LINC00671* ( $\log FC = -5,628$ ) também estão significativamente alterados ( $\log FC > |1,2|$ ) na comparação de expressão diferencial entre amostras normais e tumorais.

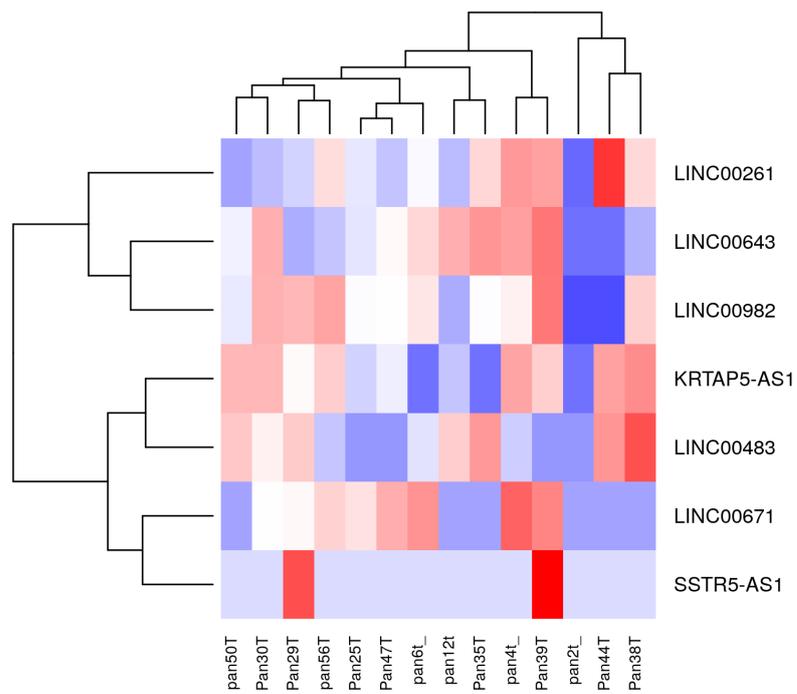


Figura 16 – Padrão de expressão dos lncRNAs identificados como diferencialmente expressos no subgrupo 3 e com anotação em amostras obtidas do hospital AC Camargo.

## 5 DISCUSSÃO

### 5.1 Desafios Impostos pela Heterogeneidade Tumoral

Os avanços em biologia celular e molecular forneceram um repertório informacional cada vez maior sobre a biologia humana que vai desde o nível molecular ao nível fisiológico. Esses avanços foram em grande parte centrados em uma abordagem reducionista, realizando experimentos para realizar a anotação dos componentes biológicos, sejam esses genes, proteínas, células ou tecidos. A biologia de sistemas é uma disciplina crescente que emprega a análise integrada na caracterização de sistemas biológicos complexos, no qual as interações entre os componentes de um sistema são descritos matematicamente para estabelecer um modelo computável (EDELMAN; EDDY; PRICE, 2010).

A natureza extremamente complexa e heterogênea do câncer, com desenvolvimento a partir de múltiplos eventos moleculares e celulares, somado às interações formadas com o microambiente tumoral em vários níveis fisiológicos entre diversas células e tecidos. Além disso, alterações no genoma de uma célula, somadas a atuação das células estromais componente dos microambiente tumoral, podem vir a conferir habilidades proliferativas e de evasão da supressão tumoral, sendo que o evento de aquisição do fenótipo maligno de uma célula requer várias mutações que venha a atuar sinergisticamente (HANAHAN; WEINBERG, 2011).

A confluência dos efeitos gerados pelas modificações genéticas e epigenéticas ilustra a relação complexa, não linear, entre os estados molecular e fenotípico de uma célula cancerosa, enfatizando a necessidade de integração de dados heterogêneos por modelos *in silico*. O grande número de modelos existentes para estudo de câncer reflete a amplitude de eventos moleculares, celulares e fisiológicos decorrentes. Abordagens mais refinadas fazem uso de técnicas estatísticas aplicadas a dados de expressão gênica provenientes de sequenciamento de alta-capacidade para a identificação de assinaturas moleculares e os fenótipos associados em câncer.

Essas assinaturas podem apontar para uma determinada função aberrante de genes ou vias, podendo ser usadas para prever o tipo, estadiamento ou grau histológico de biópsias de amostras tumorais. Métodos estatísticos avançados podem também ser usados para inferir a estrutura ou relação entre moléculas em redes regulatórias de importância no câncer. Além disso, modelos estequiométricos ou cinéticos de reações em redes de reações bioquímicas podem ser de uso

na simulação mecanística de comportamento metabólico ou da transdução de sinal no câncer (EDELMAN; EDDY; PRICE, 2010).

O desafio de entender a biologia tumoral é de grande interesse, particularmente no caso do PDAC, visto a alta mortalidade e franca carência de marcadores acurados ou tratamentos efetivos. A natureza altamente estocástica de tumores pancreáticos, somado a heterogeneidade molecular, celular, fenotípica e clínica torna a demarcação dos componentes responsáveis e atuantes na tumorigênese e manutenção tumoral ainda mais árdua.

Abordagens *in silico* de deconvolução de assinaturas moleculares em PDAC são de grande interesse, por permitirem um modo de extração dos componentes moleculares associados aos diferentes fenótipos existentes com implicações clínicas distintas.

Este estudo utilizou NMF, uma metodologia de classificação não supervisionada *in silico* de amostras de tumores pancreáticos em subtipos moleculares distintos. Os processos celulares alterados em cada um dos subgrupos de amostras derivados da classificação não supervisionada levou a identificação de genes diferencialmente expressos, a notar o subtipo Exócrino/Endócrino, contendo *NR5A2*, *BHLHA15*, *RBPJL*, *INS*, *NEUROD1*, *NKX2-2*, *AMY2B*, *PRSS1*, *PRSS3*, *CEL* e *INS*, Imunogênico, com as vias de células B, T CD4<sup>+</sup> e T CD8<sup>+</sup>, Escamoso, com alteração em *TP63* e Progenitor, contem as mucinas *MUC1*, *MUC13*, *MUC16* e *MUC17*, recapitulando parcialmente as assinaturas moleculares descritas na por Bailey e colaboradores (BAILEY et al., 2016).

As diferenças existentes entre as assinaturas resultantes devem-se, possivelmente, a diferença da metodologia de sequenciamento (poli-A vs RNA-total) além das diferenças entre as coortes em termos de número de pacientes, etnia, geografia e espectro do grau de celularidade tumoral. A corroboração da metodologia aplicada de deconvolução de sinal em um conjunto de dados distintos reforça a identificação dos fenótipos e, por consequência, possibilita um estudo direcionado - subtipo-específico - da biologia tumoral. Além de apresentar uma assinatura de genes codificadores de proteínas característicos, os subgrupos contém também uma assinatura característica de RNAs longos não codificadores associados.

## 5.2 Influência do Microambiente no Contexto de Subgrupos Tumoriais

A presença dos componentes do microambiente tumoral, formado por células como fibroblastos, adipócitos, linfócitos T e B, células dendríticas, mielóides, endoteliais, entre outras,

formam uma rede de constante interação com as células tumorais, podendo auxiliar no desenvolvimento, proliferação, migração, evasão do sistema imune e resistência a quimioterápicos (BALKWILL; CAPASSO; HAGEMANN, 2012).

A análise de enriquecimento para os componentes celulares estromais e imunes pode permitir entender melhor o contexto de desenvolvimento e manutenção dos diferentes subgrupos encontrados, além de auxiliar na elucidação do papel funcional que lncRNAs possam estar desempenhando em diferentes compartimentos do tumor e seu microambiente.

O alto enriquecimento do subgrupo 2 (Imunogênico) para componentes do sistema imune é consistente com a assinatura dos genes e o fenótipo designado para este grupo, sendo que indica maior ativação de uma resposta imune e, conseqüentemente, maior suscetibilidade a tratamentos quimioterápicos e melhor prognóstico geral.

O subgrupo 3 (Escamoso), por outro lado, apresenta enriquecimento significativo para componentes estromais, o que pode ser associado a uma maior resistência a tratamentos quimioterápicos e conseqüente pior prognóstico. Isso se deve à expressão e conseqüente secreção intensa de componentes da matriz extracelular pelas células estromais altamente ativas, que acaba por tornar a região peritumoral extremamente fibrosa, criando um microambiente de difícil acesso tanto por células do sistema imune como de quimioterápicos e por sua vez gerando grande interesse nos componentes moleculares, como os lncRNAs, atuantes neste subtipo tumoral.

### 5.3 O Contexto de lncRNAs em Subtipos Moleculares de PDAC

A motivação de identificar, associar, caracterizar e contextualizar os lncRNAs vem do fato de que ainda há pouca informação sobre essa espécie de RNAs, no que tange tanto na identificação contexto-específico quanto na função que venham a desempenhar. O potencial que os lncRNAs possuem enquanto reguladores de diferentes processos celulares os torna de grande interesse para a biologia. Esse potencial dos lncRNAs é ainda pouco explorado no contexto da formação e manutenção tumoral, sendo a determinação de seu papel crucial para a elucidação dos diversos processos celulares atuantes.

Em 2017, Raphael e colaboradores (RAPHAEL et al., 2017) analisaram dados de expressão de lncRNAs em amostras de PDAC provenientes do banco de dados do TCGA, e por meio de agrupamento hierárquico consenso a partir dos 360 lncRNAs mais variáveis obtiveram dois grupos distintos, um “Clássico” e um “Basal”, denominados em função da sobreposição com a

mesma classificação feita separadamente com mRNAs.

A subsequente análise de expressão diferencial entre os dois grupos levou a identificação de lncRNAs característicos, como *EVADR*, que também havia sido associado com adenocarcinomas e inclusive PDAC anteriormente (GIBB et al., 2015), e *LINC00261*, quase duas vezes mais expresso no subtipo clássico que no basal. Já foi demonstrado o papel de *LINC00261* na regulação da expressão de *FOXA2* através do recrutamento de *SMAD2/3* ao promotor de *FOXA2* (JIANG et al., 2015).

Além disso, *LINC00261* já foi implicado como possuindo um papel funcional em tumores pancreáticos (MULLER et al., 2015) e na formação do pâncreas (JIANG et al., 2015; ZORN; WELLS, 2009). Outro lncRNA anotado por Raphael e colaboradores é *GATA6-AS1*, também duas vezes mais expresso no subtipo clássico; já foi demonstrada a ativação transcricional deste antissenso em células tronco embrionárias durante o processo de diferenciação em endoderma (SIGOVA et al., 2013).

Neste trabalho, o *LINC00261* (*C20orf56*) também foi identificado como hiperexpresso no subtipo exócrino ( $\log FC = 1,503$ ) e hipoexpresso no subtipo escamoso ( $\log FC = -2,886$ ). No conjunto de dados avaliado neste estudo a sobrevida não foi significativa ( $p = 0,274$ ), ao passo que a comparação dos níveis de *LINC00261* em tumores de pulmão de não pequenas células (*Non Small Cell Lung Cancer* - NSCLC) em relação à amostras normais por LIU, XIAO e XU (LIU; XIAO; XU, 2017) mostrou uma diminuição, validada por PCR em tempo real (*quantitative Real-time PCR* - qRT-PCR), assim como uma correlação com estadiamento, acometimento de metástase e conseqüente pior prognóstico.

Em vista da sobreposição da classificação das amostras nos subtipos basal descrito por Mofit e escamoso descrito por Bailey (mostrado no trabalho de Raphael e colaboradores), o padrão de expressão de *LINC00261* corrobora os resultados identificados neste contexto, tornando este RNA não codificador de interesse para maiores estudos no contexto de PDAC.

#### 5.4 Integração de Dados

A abordagem de considerar genes como parte de uma rede integrada de sistemas, ao invés de alvos isolados, têm grande apelo por ser muito intuitiva do ponto de vista biológico. Redes de grafos permitem a representação de sistemas biológicos, sejam estas interações entre proteínas, metabólitos ou, como no caso deste estudo, da correlação do perfil de expressão gênica.

Redes de coexpressão permitem avaliar de maneira quantitativa o grau de associação, ou conectividade, entre os genes. A avaliação do contexto biológico e clínico de módulos de genes interconexos contribui de maneira mais robusta para o entendimento da biologia tumoral.

A identificação de 13 módulos gênicos distintos, representados por cores, permitiu a caracterização biológica através do enriquecimento para processos/vias característicos, como de reparo de DNA e progressão do ciclo celular, incluindo as vias de, mas não limitado à “Resposta a UV”, “Recombinação Homóloga” e “Reparo por excisão de base”, levando à inferência de que este grupo de tumores pode ter desenvolvido maior resistência a tratamentos quimioterápicos.

Além disso, a análise de enriquecimento para potenciais marcadores prognósticos apontou para o módulo *Tan*, como sendo um módulo de grande interesse por conter um número considerável de genes mais conexos e com potencial prognóstico significativo. Entre os genes identificados, como mostra a curva de sobrevida da Figura 11, níveis mais altos de expressão de *AURKA* e *TPX2* implicam em um aumento de mais de duas vezes no risco à sobrevida em relação à pacientes com a expressão mais baixa.

Essa informação é corroborada pelo grande enriquecimento para componentes estromais do microambiente tumoral. Esse subgrupo apresenta os lncRNAs *FAM83A-AS1* ( $\log FC = 2,312$ ) e *DKFZp434J0226* ( $\log FC = 1,403$ ) hiperexpressos entre tumores. Em um estudo de comparação dos níveis de expressão em amostras normais e tumorais de carcinoma hepatocelular por microarranjo de DNA, *FAM83A-AS1* foi identificado como significativamente alterado, com expressão mais baixa em tumores, validado por RT-PCR (ZHU et al., 2014). A análise de regressão de cox em função da expressão identificou *FAM83A-AS1* como potencial marcador prognóstico ( $p = 0,045$ ), sendo que o aumento de sua expressão implica em um aumento de 79% no risco de sobrevida para o paciente.

Entre os lncRNAs hipoexpressos no subgrupo 3, *FLJ42875*, *LOC338651*, *C20orf56* e *LOC388387* possuem alta afinidade pela proteína BRCA2, um importante regulador de processos tumorais. A expressão mais baixa desses lncRNAs, e a concomitante abundância significativamente maior do produto proteico de *BRCA2* (Figura 13) neste subgrupo em relação aos outros subgrupos leva a hipótese primária de que estes lncRNAs podem, conjuntamente, desempenhar uma função de complexo repressor de BRCA2.

A análise de exoma mostrou que há poucas amostras com mutações no gene *BRCA2* (6 ao total), sendo que o subgrupo 3 não possui nenhuma amostra com mutações detectadas para este gene. Essa desregulação epigenética nos tumores pode estar conferindo a este subgrupo

maior disponibilidade para uso do maquinário de reparo de DNA e, conseqüentemente, menor suscetibilidade a letalidade sintética e maior resistência a quimioterápicos.

## 6 CONCLUSÕES

A aplicação da metodologia *in silico* de classificação não supervisionada segundo o perfil de expressão gênica de amostras de PDAC por NMF levou a identificação de quatro subtipos moleculares distintos, de maneira a recapitular os fenótipos descritos na literatura realizados em outros conjuntos de amostras. A análise de expressão diferencial permitiu a identificação de lncRNAs associados aos diferentes subtipos, como *FAM83A-AS1*, hiper-expresso no subgrupo 3 (de fenótipo escamoso), associado a um pior prognóstico, em relação aos outros subgrupos e com potencial de interação com as proteínas FGFR2, AXIN1, PTEN, BRAF, SMAD4, TGFBR2, TP53 e CDKN2A, moléculas reconhecidamente importantes por seu papel na transdução de sinal e supressão tumoral em vários tipos de câncer, incluindo o de pâncreas. Entre os lncRNAs com expressão significativamente baixa no subgrupo 3 em relação ao outros subgrupos, *FLJ42875*, *LOC338651*, *C20orf56* e *LOC38838* possuem potencial de interação de alta afinidade à proteína BRCA2, um importante regulador em processos como reparo de DNA e resistência a quimioterápicos, levando à hipótese de um possível mecanismo de complexo repressor desempenhado por estes lncRNAs e sua associação com pior prognóstico. Além disso, a integração de dados com módulos de co-expressão gênica levou à associação do subgrupo 3 a funções de reparo de DNA e progressão do ciclo celular, assim como a associação à possíveis marcadores prognósticos, como *AURKA* e *TPX2*, genes envolvidos na regulação da progressão do ciclo celular. A metodologia aplicada neste trabalho contribui com a adição de novos candidatos a reguladores de processos celulares em câncer de pâncreas.

## REFERÊNCIAS

- AIELLO, N. M. et al. Metastatic progression is associated with dynamic changes in the local microenvironment. *Nat. Commun.*, v. 7, 2016.
- ALIZADEH, A. A. et al. Toward understanding and exploiting tumor heterogeneity. *Nature medicine*, v. 8, p. 846–53, 2015.
- AMUNDADOTTIR, L. T. Pancreatic cancer genetics. *Int. J. Biol. Sci.*, v. 12, p. 314–325, 2016.
- ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, v. 25, p. 25–9, 2000.
- BAILEY, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, v. 531, p. 47–52, 2016.
- BALKWILL, F. R.; CAPASSO, M.; HAGEMANN, T. The tumor microenvironment at a glance. *J. Cell Sci.*, v. 125, p. 5591–6, 2012.
- BECKEDORFF, F. C. et al. The intronic long noncoding rna anrassf1 recruits prc2 to the rassf1a promoter, reducing the expression of rassf1a and increasing cell proliferation. *PLoS*, v. 9, 2013.
- BHAN, A.; SOLEIMANI, M.; MANDAL, S. S. Long noncoding rna and cancer: A new paradigm. *Cancer Res.*, v. 77, p. 3965–3981, 2017.
- BRUNET, J. P. et al. Metagenes and molecular pattern discovery using matrix factorization. *PNAS*, v. 101, p. 4164–9, 2004.
- CERASE, A. et al. Xist localization and function: new insights from multiple levels. *Genome Biol.*, v. 16, 2015.
- CHAND, S. et al. The landscape of pancreatic cancer therapeutic resistance mechanisms. *Int. J. Biol. Sci.*, v. 12, p. 273–82, 2016.
- CIRILLO, D. et al. Quantitative predictions of protein interactions with long noncoding rnas. *Nature Methods*, v. 14, 2017.
- COLLISSON, E. A. et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.*, v. 17, p. 500–503, 2011.
- CONSORTIUM, T. G. O. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.*, v. 45, p. 331–338, 2016.
- DAGOGO-JACK, I.; SHAW, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.*, v. 15, p. 81–94, 2018.
- DJEBALI, S. et al. Landscape of transcription in human cells. *Nature*, v. 489, p. 101–8, 2012.
- EDELMAN, L. B.; EDDY, J. A.; PRICE, N. D. In silico models of cancer. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, v. 2, p. 438–459, 2010.
- GALL, T. M.; WASAN, H.; JIAO, L. R. Pancreatic cancer: current understanding of molecular and genetic aetiologies. *Postgrad Med J*, v. 91, p. 594–600, 2015.

- GAUJOUX, R.; SEOIGHE, C. A flexible r package for nonnegative matrix factorization. *BMC Bioinformatics*, v. 11, 2010.
- GHARIBI, A.; ADAMIAN, Y.; KELBER, J. A. Cellular and molecular aspects of pancreatic cancer. *Acta Histochemica*, v. 118, p. 305–16, 2016.
- GIBB, E. A. et al. Activation of an endogenous retrovirus-associated long non-coding rna in human adenocarcinoma. *Genome Med*, v. 7, 2015.
- GUPTA, R. A. et al. Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, v. 464, p. 1071–6, 2010.
- HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. *Cell*, v. 144, p. 646–74, 2011.
- HARROW, J. et al. Gencode: the reference human genome annotation for the encode project. *Genome Res.*, v. 22, p. 1760–74, 2012.
- HRUBAN, R. H. et al. Progression model for pancreatic cancer. *Clin. Cancer Res.*, v. 6, p. 2969–72, 2000.
- INCA. <<http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pancreas>>, 2016.
- JIANG, W. et al. The lncrna deanr1 facilitates human endoderm differentiation by activating foxa2 expression. *Cell Rep.*, v. 11, p. 137–48, 2015.
- JONES, P.; BAYLIN, S. B. The epigenomics of cancer. *Cell*, v. 128, p. 683–92, 2007.
- KALIMUTHU, S. N.; CHETTY, R. Gene of the month: Smarcb1. *J. Clin. Pathol.*, v. 69, p. 484–489, 2016.
- KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, v. 28, p. 27–30, 2000.
- KAPRANOV, P. et al. The majority of total nuclear- encoded non-ribosomal rna in a human cell is "dark matter"un-annotated rna. *BMC Biol.*, v. 8, 2010.
- KASSAMBARA, A. survminer: Drawing survival curves using 'ggplot2'. *R package version 0.4.3*, 2018.
- KHAN, M. A. et al. Molecular drivers of pancreatic cancer pathogenesis: Looking inward to move forward. *Int. J. Mol. Sci.*, v. 18, 2017.
- KOPP, F.; MENDELL, J. T. Functional classification and experimental dissection of long noncoding rnas. *Cell*, v. 172, p. 393–407, 2018.
- KOSINSKI, M.; BIECEK, P. Rtcga: The cancer genome atlas data integration. *R package version 1.10.0*, 2016.
- LANGFELDER, P.; HORVATH, S. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, v. 9, 2008.
- LEX, A. et al. Upset: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, v. 20, p. 1983–1992, 2014.
- LI, B.; DEWEY, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, v. 12, 2011.

- LIU, Y.; XIAO, N.; XU, S. F. Decreased expression of long non-coding rna linc00261 is a prognostic marker for patients with non-small cell lung cancer: a preliminary study. *European Review for Medical and Pharmacological Sciences*, v. 21, p. 5691–5695, 2017.
- MAGLIANO, M. P. di et al. Common activation of canonical wnt signaling in pancreatic adenocarcinoma. *PLoS One*, v. 2, 2007.
- MAHMOUDI, S. et al. Wrap53, a natural p53 antisense transcript required for p53 induction upon dna damage. *Mol. Cell*, v. 33, p. 462–71, 2009.
- MAHMOUDI, S. et al. Wrap53 promotes cancer cell survival and is a potential target for cancer therapy. *Cell Death Dis.*, v. 2, 2011.
- MAKOHON-MOORE, A.; IACOBUZIO-DONAHUE, C. A. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nat. Rev. Cancer*, v. 16, p. 553–65, 2016.
- MANTOVANI, A. et al. Cancer-related inflammation. *Nature*, v. 454, p. 436–44, 2008.
- MOFFITT, R. A. et al. Virtual microdissection identifies distinct tumor - and stroma - specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.*, v. 47, p. 1168–78, 2015.
- MULLER, S. et al. Next-generation sequencing reveals novel differentially regulated mrnas, lncrnas, mirnas, sdrnas and a pirna in pancreatic cancer. *Mol. Cancer*, v. 11, p. 137–48, 2015.
- NING, S. et al. Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. *Nucleic Acids Res.*, v. 44, p. 980–5, 2016.
- PARALKAR, V. R.; WEISS, M. J. Long noncoding rnas in biology and hematopoiesis. *Blood*, v. 121, p. 4842–6, 2013.
- PREIS, M.; KORC, M. Signaling pathways in pancreatic cancer. *Crit. Rev. Eukaryot. Gene Expr.*, v. 21, p. 115–29, 2011.
- PROVENZANO, P. P. et al. Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma. *Cancer Cell*, v. 21, p. 418–29, 2012.
- QUINN, J. J.; CHANG, H. Y. Unique features of long non-coding rna biogenesis and function. *Nat. Rev. Genet.*, v. 17, p. 47–62, 2016.
- RANSOHOFF, J. D.; WEI, Y.; KHAVARI, P. A. The functions and unique features of long intergenic non-coding rna. *Nat. Rev. Mol. Cell Biol.*, v. 19, p. 143–157, 2017.
- RAPHAEL, B. J. et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*, v. 14, p. 185–203, 2017.
- REIMAND, J. et al. g:profiler - a web-based toolset for functional profiling of gene lists from large-scale experiments. *NAR*, v. 35, p. 193–200, 2007.
- REN, C. et al. Lnc2catlas: an atlas of long noncoding rnas associated with risk of cancers. *Scientific reports*, v. 8, 2018.
- RINN, J. L. et al. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell*, v. 129, p. 1311–23, 2007.
- RINN, J. L.; CHANG, H. Y. Genome regulation by long noncoding rnas. *Annu. Rev. Biochem.*, v. 81, p. 145–66, 2012.

- RITCHIE, M. E. et al. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.*, v. 43, 2015.
- SCHLITTER, A. M. et al. Molecular, morphological and survival analysis of 177 resected pancreatic ductal adenocarcinomas (pdacs): Identification of prognostic subtypes. *Sci. Rep.*, v. 7, 2017.
- SCHWENDER, H. siggenes: Multiple testing using sam and efron's empirical bayes approaches. *R package version 1.50.0*, 2012.
- SIGOVA, A. A. et al. Divergent transcription of long noncoding rna/mrna gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. U S A*, v. 110, p. 2876–81, 2013.
- SOMERVILLE, T. D. D. et al. Tp63-mediated enhancer reprogramming drives the squamous subtype of pancreatic ductal adenocarcinoma. *Cell Rep.*, v. 25, p. 1741–1755, 2018.
- TAHIRA, A. C. et al. Long noncoding intronic rnas are differentially expressed in primary and metastatic pancreatic cancer. *Mol. Cancer*, v. 10, 2011.
- TERAI, G. et al. Comprehensive prediction of lncrna-rna interactions in human transcriptome. *BMC Genomics*, v. 17, 2016.
- THERNEAU, T.; LUMLEY, T. survival: Survival analysis. *R package version 2.43-3*, 2018.
- TINDER, T.; SUBRAMANI, D.; BASU, G. Muc1 enhances tumor progression and contributes toward immunosuppression in a mouse model of spontaneous pancreatic adenocarcinoma. *J. Immunol.*, v. 181, p. 3116–25, 2008.
- USZCZYNSKA-RATAJCZAK, B. et al. Towards a complete map of the human long non-coding rna transcriptome. *Nat. Rev. Gen.*, v. 19, p. 535–548, 2018.
- VERSTEEGE, I. et al. Truncating mutations of hsnf5/ini1 in aggressive paediatric cancer. *Nature*, v. 394, p. 203–6, 1998.
- VOLDERS, P. J. et al. An update on Incipedia: a database for annotated human lncrna sequences. *Nucleic Acids Res.*, v. 43, p. 174–80, 2015.
- WANG, K. C. et al. A long noncoding rna maintains active chromatin domain to coordinate homeotic gene expression. *Nature*, v. 472, p. 120–4, 2011.
- WEINSTEIN, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, v. 45, p. 1113–20, 2013.
- WIDSCHWENDTER, M. et al. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nat. Rev. Clin. Oncol.*, v. 15, p. 292–309, 2018.
- WORMANN, S. M. et al. Loss of p53 function activates jak2-stat3 signaling to promote pancreatic tumor growth, stroma modification, and gemcitabine resistance in mice and is associated with patient survival. *Gastroenterology*, v. 151, p. 180–193, 2016.
- YOSHIHARA, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, v. 4, 2013.
- ZHU, J. et al. The long noncoding rna expression profile of hepatocellular carcinoma identified by microarray analysis. *PLoS One*, v. 9, 2014.
- ZORN, A. M.; WELLS, J. M. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.*, v. 25, p. 221–51, 2009.