Programa de Pós-graduação Interunidades em Bioinformática - USP

Tese de Doutorado

“Análise integrada de alterações somáticas e
transcrpcionais em tumores de pâncreas”

Aluno: Diogo Vieira da Silva Pellegrina

Orientador: Eduardo Moraes Rego Reis

Data: 28/02/2020

Sumário

Resumo	4
Abstract	6
Agradecimentos	8
Introdução	9
O adenocarcinoma ductal do pâncreas	9
Alterações somáticas em PDAC	11
Alterações transcricionais no PDAC	12
Integração de dados de exoma e transcritoma	14
LncRNAs regulatórios no câncer de pâncreas	16
Anotação funcional de lncRNAs através da análise de redes de co-expressão gênica	18
Objetivos	22
Objetivos Gerais	22
Objetivos Específicos	22
Materiais e Métodos	23
Amostras clínicas de PDAC	24
Sequenciamento do exoma e transcritoma	25
Processamento dos dados de exoma e chamada de variantes	27
Enriquecimento de categorias gênicas	33
Montagem de redes de co-expressão de genes codificadores de proteína e lncRNAs	34
Validações Experimentais	36
Resultados	38
Análise do exoma	38
Processamento de dados de sequenciamento e chamada de variantes	38
Priorização e anotação de genes com mutações relevantes no PDAC	46
Vias moleculares enriquecidas entre genes significativamente alterados	51
Análise de assinaturas mutacionais no PDAC	55
Construção de redes de co-expressão	57
Integração entre dados de alterações somáticas e transcricionais	62
Comparação de vias moleculares enriquecidas	62
Associação entre características do exoma e módulos da rede de co-expressão	63
Inferência funcional de lncRNAs a partir de rede de co-expressão gênica	67

Discussão	71
Análise de alterações somáticas simples no PDAC	71
Seleção de candidatos a gene “drivers” no PDAC	74
Vias moleculares enriquecidas em genes mutados no PDAC	80
Assinaturas mutacionais detectadas no PDAC	82
Análise de integração dos dados exoma e rede de co-expressão	83
Análise do transcrito usando redes de co-expressão	90
Conclusões	95
Referências	99
Anexos	113

Resumo

O adenocarcinoma ductal do pâncreas (PDAC) é uma neoplasia caracterizada por uma alta heterogeneidade genômica e a relação entre mutações somáticas recorrentes e os padrões de expressão gênica nestes tumores é ainda pouco caracterizada. Foi desenvolvida uma rotina computacional para processar e analisar dados do sequenciamento do exoma de 14 casos de pacientes com PDAC e identificar mutações somáticas simples (SNVs, *indels*). Dados de transcriptoma das mesmas amostras foram usados para montagem de uma rede de co-expressão de genes codificadores de proteína e RNAs não codificadores longos (lncRNAs). Análises de enriquecimento de genes foram usadas para identificar vias moleculares entre genes recorrentemente mutados, genes diferencialmente expressos ou genes localizados em módulos específicos da rede. A alta conectividade de transcritos com genes de vias moleculares específicas foi usada para inferência funcional de lncRNAs. Foram identificadas 28.683 mutações simples em 12.438 genes, incluindo 3.314 alterações somáticas já anotadas em tumores humanos e 976 genes com papel oncogênico ou supressor tumoral no PDAC. Entre 1.603 candidatos a genes “drivers” destacam-se 18 com mutações recorrentes, incluindo componentes do complexo Ubiquitina Ligase E3, fatores de transcrição, modificador de histonas, remodeladores de cromatina, transdução de sinal e apoptose. A sobreposição de vias moleculares enriquecidas em mutações somáticas e alterações transcricionais identificou 83 vias em comum, incluindo ‘*Ras protein signal transduction*’, ‘*apoptotic process*’, e ‘*cell adhesion*’. Para 11 genes com mutações somáticas com alta prevalência foi encontrada uma correlação entre a presença da mutação e alterações na expressão de módulos da rede. Entre esses se destaca a associação de mutações em KRAS com a expressão de módulos enriquecidos nas vias ‘*Hedgehog signalling*’, ‘*WNT signalling*’, ‘*DNA repair*’ e ‘*Cell Cycle*’, e mutações em TP53 e a expressão alterada em módulos contendo genes das vias de sinalização de ErbB. Foram também observadas associações significativas (p

$<1 \times 10^{-4}$) entre a expressão gênica de módulos da rede e assinaturas mutacionais e a sobrevida dos pacientes. Foram identificados lncRNAs diferencialmente expressos no PDAC e com alta conectividade a genes associados a vias moleculares específicas como '*Cell Adhesion*', '*O-linked glycosylation of mucins*' e '*DNA Repair*'. Validações experimentais corroboraram a função regulatória destes lncRNAs nestas vias. A existência de casos sem mutação em KRAS e de vias em que todos os casos avaliados possuem ao menos um gene da via mutado sugerem que mutações em outros componentes destas vias possam fenocopiar o efeito oncogênico de mutações em KRAS. Este trabalho confirmou a alta heterogeneidade de mutações somáticas no PDAC e revelou novos genes "drivers" candidatos representados na população brasileira.

Abstract

The pancreatic ductal adenocarcinoma (PDAC) is a neoplasia with a characteristically high genomic heterogeneity and without much data on the interactions between somatic mutations and the tumoral gene expression. A computational routine was developed to process and analyze exome sequencing data from 14 samples from PDAC patients and identify its simple somatic mutations (SNVs and indels). Transcriptome data from those same samples was used to build a coexpression network of protein coding and long noncoding genes (lncRNAs). Pathway enrichment analyses were used to identify molecular pathways with frequently mutated genes, differentially expressed genes, or genes found on specific network modules. lncRNAs were found in the coexpression network showing strong connectivity in certain pathways, for which it was possible to infer a regulatory function. 28.683 simple mutations in 12.438 genes, including 3.314 somatic alterations already cataloged in human tumors and 976 genes characterized as tumor suppressors or oncogenes in PDAC. From the 1.603 driver gene candidates 18 are found recurrently, including components of the Ubiquitin Ligase E3 complex, transcription factors, histone modifiers, chromatin remodelers, genes involved in signal transduction and apoptosis. The 83 molecular pathways that are both enriched by somatic mutations and differential expression, including 'Ras protein signal transduction', 'apoptotic process', and 'cell adhesion'. For 11 of the most prevalent genes a correlation was found between being mutated and the gene expression on network modules. Among those is highlighted the KRAS mutations correlation with altered expression on modules enriched by 'Hedgehog signalling', 'WNT signalling', 'DNA repair' and 'Cell Cycle' pathways, and TP53 mutations with altered expression on modules enriched by the 'ErbB signaling pathway'. Significant associations ($p < 1 \times 10^{-4}$) were also found between module gene expression and both mutational signatures and survival. Differentially expressed lncRNAs were identified highly connected to specific molecular pathways such as 'Cell Adhesion',

'O-linked glycosylation of mucins' and 'DNA Repair'. Experimental validations supported the proposed regulatory functions of the lncRNAs in those pathways. The existence of PDAC cases without KRAS mutations and of pathways in which every study case has a mutation in at least one gene suggests that some mutations in genes from those pathways might phenocopy the oncogene effect of mutations in KRAS. This work confirmed the high heterogeneity of somatic mutations in PDAC and revealed new driver gene candidates in a sample of the Brazilian population.

Agradecimentos

Este trabalho não seria possível sem a orientação do prof. Eduardo Reis e a colaboração de vários membros do laboratório, entre eles Thalita Bueno, Bianca Dazzani, Julio Sosa, Diogo Pessoa, e em especial Ester Bertoldi e Vinicius Paixão. E também sem o servidor LBI mantido pelo prof. João Setubal e Carlos Moraes.

Também agradeço ao prof. Israel Tojal e Renan Valieris pela análise do SigneR e a profa. Dirce Carraro e Cláudia Alessandra de Paula pelo sequenciamento de amplicons de KRAS, além dos outros membros das bancas dos exames de qualificação, prof. Carlos Menck, Profa. Ana Krepischi, e Prof. Pedro Galante que contribuíram com observações muito importantes para o sucesso deste trabalho.

Por fim também agradeço a David Pires, Deborah Stern e Eduardo Dobay colaboradores extra-oficiais que me ajudaram tantas vezes, e à Roberta Scavone minha colaboradora da vida, que fez um ótimo sumário.

Este projeto foi financiado por uma bolsa da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e um auxílio à pesquisa da FAPESP (auxílio regular 2013/13844-2).

Introdução

O adenocarcinoma ductal do pâncreas

O câncer de pâncreas é uma das neoplasias mais letais e apenas 6% dos indivíduos afetados sobrevivem mais do que 5 anos após o diagnóstico (Siegel et al., 2014). Embora não se inclua entre as neoplasias mais prevalentes, representa uma das principais causas de morte por câncer no Brasil e no mundo (IARC 2008), com uma taxa de incidência (nº novos casos / 100.000 indivíduos) próxima à taxa de mortalidade (Hariharan et al., 2008; Hidalgo 2010). Apenas uma pequena fração tem origem familiar (5 a 10%), sendo a maior parte dos casos compostos por tumores esporádicos (Hidalgo 2010). Os adenocarcinomas ductais que acometem a região exócrina do órgão (PDAC, *Pancreatic ductal adenocarcinoma*) constituem o tipo mais frequente de neoplasia pancreática, representando mais de 80% dos casos (Jones et al., 2008).

A única opção terapêutica curativa do PDAC é a ressecção cirúrgica ainda em estágios iniciais, quando os tumores se encontram confinados à glândula. Contudo, a doença apresenta sintomas sutis e que costumam ser atribuídos a outras enfermidades mais comuns, fazendo com que a maior parte dos novos casos seja diagnosticada já em estágios avançados, resultando na baixa sobrevida da doença. O tratamento conservador é a pancreatectomia, embora apenas uma fração das lesões precursoras

evolua para tumores invasivos, um procedimento associado à elevada morbidade.

O câncer de pâncreas, como outras neoplasias, é uma doença genética em que mutações que promovem a divisão celular, que inibem o reparo do DNA, entre outras, promovem um sistema de *feedback* positivo em que as células alteradas aumentem em número e com o tempo se tornam cada vez mais mutadas (McGranahan and Swanton 2015). Estas alterações genéticas desregulam inúmeros processos celulares, afetando também a expressão gênica e a função do tecido (McGranahan and Swanton 2015). Muitos estudos já indicaram que o único tratamento curativo do PDAC se dá pela remoção cirúrgica do tumor ainda em estágios iniciais (Kleeff et al., 2016), e que o surgimento do fenótipo metastático do câncer de pâncreas ocorre ao longo de vários anos (Yachida et al., 2010), o que salienta a importância da intervenção terapêutica em estágios iniciais quando a doença é curável. O estadiamento do PDAC é feito pelo patologista a partir dos protocolos estabelecidos pelo *AJCC Cancer Staging Manual* em que se observa o tamanho do tumor, o estado dos linfonodos locais, e a existência de metástases (Appel et al., 2012). Muitas informações sobre tumores de pâncreas (além de outros tipos) pode ser obtida no ICGC (ICGC, 2010), uma plataforma que mantém dados obtidos por diversos consórcios que extraíram e realizaram experimentos de larga escala em amostras tumorais. Estão disponíveis dados clínicos, históricos familiares, dados de expressão gênica por microarranjos e dados de sequenciamento, dados de alterações somáticas simples e estruturais, dados de metilação, entre outros. É importante notar, no entanto, que estes dados não descrevem todas as amostras,

cada uma participa de uma fração destes experimentos, havendo poucas amostras em que há disponível mais de um tipo de dado.

Alterações somáticas em PDAC

Análises moleculares já realizadas indicam que maior parte das mutações detectadas no genoma de tumores em pacientes com PDAC são adquiridas ao longo da vida. Mutações somáticas ativadoras no oncogene KRAS estão presentes na maior parte dos casos (Hruban et al., 1993; Lohr et al., 2005), seguidas por mutações inativadoras nos genes supressores de tumor CDKN2A, SMAD4, e TP53 (Biankin et al., 2012). Estudos de sequenciamento com alta resolução da porção codificadora do DNA genômico (exoma) confirmaram a mutação frequente desses oncogenes ou supressores tumorais e revelaram a ausência de outras mutações somáticas com alta prevalência no PDAC. Estes estudos genômicos também mostraram que os tumores pancreáticos se caracterizam por uma grande heterogeneidade inter e intratumoral e diversidade de alterações somáticas com baixa ou média frequência alélica (Cros et al., 2018). No entanto, um fator comum aos tumores estudados se dá pelas vias moleculares alteradas, vias como apoptose, sinalização *Hedgehog*, sinalização de KRAS, e sinalização Wnt/Notch estão presentes em quase todos os tumores encontrados (Jones et al., 2008).

Os diferentes fatores que originam as mutações somáticas em uma célula, como raios UV, erros de replicação, agentes químicos mutagênicos, entre outros, o fazem de

forma diferente, dependendo da base nitrogenada do nucleotídeo e também da identidade do nucleotídeo vizinho. Dessa forma diferentes agentes mutagênicos podem gerar padrões mutacionais específicos e característicos. Um fator complicador para a identificação desses padrões mutacionais é que múltiplos fatores estão atuando simultaneamente, em diferentes intensidades, resultando em um padrão complexo de alterações somáticas encontradas no tumor. A análise de quais nucleotídeos se encontram mais mutados em um tumor pode trazer pistas sobre os agentes mutagênicos que deram origem às alterações encontradas. Uma abordagem utilizada para determinar quais processos mutagênicos estão atuando, e em que proporção, é a fatorização de matrizes não negativas (NMF). A NMF separa o conjunto de alterações somáticas em diferentes sinais, gerando assinaturas mutacionais, que podem estar presentes em diferentes proporções em uma dada amostra. O COSMIC mantém um repertório de assinaturas mutacionais encontradas em diferentes tumores, e as classifica de acordo com o tipo de agente mutagênico que é predominantemente responsável por sua formação (Alexandrov et al., 2020).

Alterações transcricionais no PDAC

Apesar de ser de origem genética, o PDAC produz alterações muito evidentes na expressão gênica, análises anteriores do transcrito observam grande número de genes alterados em vias chave para o desenvolvimento tumoral, em Zhao et al., 2018, uma metanálise que englobou 1200 amostras de PDAC, encontrou centenas de vias diferencialmente expressas, entre elas as 50 mais significativas é possível encontrar

vias classicamente alteradas pelo desenvolvimento tumoral como ciclo celular, diferenciação celular, regulação negativa de apoptose, glicosilação de mucinas, angiogênese, resposta imune, resposta à radiação gama, entre outros. (Zhao et al., 2018)

Uma limitação dos estudos de alterações no transcrito do PDAC realizados se deve à heterogeneidade intertumoral e intratumoral, associada à diversidade de tipos celulares presentes nos fragmentos de tecido avaliados (Collisson et al., 2011, Moffitt et al., 2015). Em tumores de outros tecidos, como mama por exemplo (Sorlie et al., 2003), foram observados subtipos tumorais definidos a partir de diferenças moleculares, que possuem características distintas e respondem de forma distinta a tratamentos. Alguns estudos citados a seguir têm investigado a existência de subtipos moleculares de PDAC a partir de diferentes metodologias, usando dados de expressão gênica, mutações somáticas, entre outros.

Usando dados de expressão gênica obtidos com microarranjos de DNA, Collisson e colaboradores descreveram pela primeira vez a existência de subtipos moleculares de PDAC com perfis de expressão gênica distintos (“classical”, “quasimesenchymal”, “exocrine-like”) associados a diferentes padrões de evolução clínica e resposta terapêutica (Collisson et al., 2011). Alguns anos depois, utilizando uma abordagem de deconvolução de dados de transcrito, Moffitt e colaboradores identificaram componentes da expressão gênica atribuíveis às células epiteliais malignas e ao estroma adjacente que compõe a massa tumoral do PDAC (Moffitt et al.,

2015). Nesse estudo foram identificados dois subtipos tumorais denominados “classical” e “basal-like”, sendo este último associado a um pior prognóstico. Interessantemente, foram descritos também dois subtipos moleculares de estroma (“normal” e “activated”) que possuem valor prognóstico independente do subtipo tumoral (Moffitt et al., 2015). Foi feita uma terceira proposta de classificação molecular de subtipos de PDAC, desta vez a partir dos dados genômicos de 150 amostras de pacientes (Raphael et al., 2017) que classificou os tumores PDACs em dois sub-grupos em função da pureza tumoral, e descreveu diferenças observáveis nestes dois sub-grupos, como prevalência de diferentes mutações e diferente padrão de expressão gênica. Uma meta-análise mais recente dos dados disponíveis (Zhao et al., 2018) mostrou que existe pouca concordância entre os subtipos moleculares descritos pelos diferentes trabalhos citados acima, (Raphael et al., 2017, Collisson et al., 2011, Moffitt et al., 2015). É plausível que a classificação molecular de subtipos de PDAC seja melhor refinada à medida que sejam obtidos dados adicionais de um número maior de amostras e por diferentes abordagens experimentais. Os estudos de alterações somáticas citados acima focalizaram a análise em regiões codificadoras de proteína do genoma. Não é de nosso conhecimento a existência de trabalhos que tenham focado na identificação de mutações somáticas em regiões regulatórias ou em genes não codificadores de proteínas.

Integração de dados de exoma e transcrito

As associações funcionais entre as alterações somáticas e transcricionais

detectadas em tumores humanos, e no câncer de pâncreas em particular, ainda foram pouco documentadas de forma simultânea e integrada. Um dos estudos pioneiros que integrou dados de mutação e expressão gênica foi realizado em um conjunto de 14 amostras de tumores de mama, onde deleções, amplificações focais e *indels* detectadas no exoma foram validadas utilizando dados do transcrito (Craig et al., 2013). No contexto do câncer de pâncreas, dados de genoma de PDAC foram usados para identificar possíveis oncogenes associados a eventos de duplicação gênica (Samuel et al., 2013). Depois de verificadas quais as regiões genômicas encontravam-se recorrentemente duplicadas, uma análise transcritômica foi utilizada para selecionar para validação experimental os genes que apresentassem expressão aumentada em função do ganho do número de cópias (Samuel et al., 2013). Em outro trabalho que utilizou as mesmas técnicas (Zhao et al., 2010), foi feita uma busca por genes supressores de tumor. A partir dessa análise integrada foram priorizados candidatos não só aqueles genes que perderam heterozigose na amostra tumoral, mas os que, além disso, apresentaram a diminuição da expressão do transcrito na amostra tumoral em relação ao controle não tumoral. Uma limitação deste estudo é que, como os outros citados acima, foi interrogado com espaço amostral reduzido (frequentemente com apenas um indivíduo), tendo como principal intenção servir de prova de conceito deste tipo de análise. Recentemente, o grupo TCGA investigou de forma integrada dados de alterações somáticas e transcricionais no PDAC a fim de estabelecer uma proposta de classificação em subtipos moleculares sem, no entanto, inferir acerca do efeito funcional das mutações sobre a expressão de genes ou vias moleculares

(Raphael et al., 2017).

LncRNAs regulatórios no câncer de pâncreas

O grupo de pesquisa em que este trabalho está inserido foi pioneiro na descrição da expressão aberrante de RNAs não codificadores longos (lncRNAs) no PDAC (Tahira et al., 2011), tendo sido postulada a sua potencial utilização como biomarcadores para diagnóstico (Esteller 2011; Reis e Verjovski-Almeida 2012). Porém a maior parte dos lncRNAs ainda não foi estudada funcionalmente. RNAs não codificadores longos compreende uma classe de transcritos maiores que 200 nucleotídeos (um tamanho arbitrário que os separa de micro RNAs, piwi RNAs, e outros RNAs pequenos com tamanho tipicamente < 50 nt) que possuem significância biológica não pela tradução de sua sequência em aminoácidos, mas por exercer funções estruturais ou regulatórias, através da interação com outros RNAs, com a cromatina, ou com proteínas, que são dependentes de sua estrutura primária, secundária ou terciária.

Um número crescente de lncRNAs têm sido descritos como relevantes para a manutenção de fenótipos malignos no PDAC (Huang et al., 2016). Um dos artigos mais citados, (Kim K et al., 2013) investigou o lncRNA intergênico HOTAIR, sabidamente envolvido em tumores de outros tecidos, e demonstrou que este se encontra superexpresso em tumores de pâncreas, em comparação ao pâncreas normal. HOTAIR interage com o complexo repressor *Polycomb 2* (PRC2) que promove a

supressão de expressão gênica através da trimetilação de histonas H3K27. A abundância de HOTAIR está correlacionada à menor sobrevida dos pacientes. Ainda nesse trabalho, a análise funcional de HOTAIR em linhagens tumorais de pâncreas revelou que este lncRNA é um regulador positivo de fenótipos associados à agressividade do tumor, tais como proliferação celular, progressão do ciclo celular e inibição da apoptose (Kim K et al., 2013). Em outro trabalho foi demonstrado que a expressão do lncRNA MALAT1, que se liga a ribonucleoproteínas no núcleo envolvidas na regulação do ciclo celular, está correlacionada com o tamanho, estágio dos tumores, invasão celular, e inversamente correlacionado com sobrevida dos pacientes (Liu et al., 2014). Outro lncRNA que interage fatores de transcrição é o HOTTIP, que interage com o complexo de modificação de cromatina MLL1. O silenciamento de HOTTIP regula a expressão de 1271 genes, entre eles vários implicados na progressão tumoral, como AURKA e HOXA1 (Chenget al., 2015). O UCA1 é um lncRNA superexpresso em tumores, foi observado que sua expressão é associada à resistência de quimioterápicos, invasão de tecidos e desenvolvimento tumoral (Wang et al., 2017). Ensaio de imunoprecipitação de cromatina e de RNA mostraram que UCA1 estimula a expressão do fator de transcrição EZH2, e posteriormente se liga a ele fazendo com que este por sua vez ative a expressão de uma série de genes da via do ciclo celular, incluindo CCND1, ATK e GSK-3B (Wang et al., 2017). Esses exemplos ilustram a relevância da anotação e estudo funcional de lncRNAs, sendo importante notar que maior parte dos lncRNAs já descritos não foram caracterizados funcionalmente.

Anotação funcional de lncRNAs através da análise de redes de co-expressão gênica

Embora novas evidências venham se acumulando a respeito da função de diferentes lncRNAs, a função da maioria dos lncRNAs permanece desconhecida (Martin et al 2012). Considerando que lncRNAs podem exercer papéis regulatórios na expressão de genes codificadores, é plausível nestes casos esperar-se observar uma associação entre a expressão de um dado lncRNA regulatório e seus mRNAs alvos. Não existem ainda métodos computacionais eficientes para a predição da função de lncRNAs a partir da sua sequência primária ou predição de estruturas secundárias ou terciárias. Uma abordagem alternativa utilizada para inferir associações funcionais entre lncRNAs e mRNAs é a construção de redes de co-expressão (Cogill e Wang, 2014). Uma rede de co-expressão é uma forma abrangente com que se pode observar o comportamento do transcrito frente a uma certa condição experimental. Em uma rede, os genes formam nós (vértices) em um grafo não orientado em que as ligações (arestas) são determinadas por uma medida de correlação de suas expressões gênicas. É de se esperar em sistemas biológicos que alguns genes estejam muito mais ligados que outros, adquirindo um comportamento de centralidade (*hub*), e quando esse gene é um lncRNA pode-se supor que este gene esteja agindo como um regulador daqueles com os quais ele esteja ligado (Qi Liao et al., 2011). A partir do conceito de “*guilt by association*”, utiliza-se o conhecimento da função dos mRNAs regulados para inferir o processo biológico ou função molecular em que o lncRNA atua.

(Pellegrina et al., 2017, Jiang et al., 2019). Trabalhos que utilizaram redes de co-expressão para estudar a função de lncRNAs em câncer.

A análise de redes ponderadas de correlação (*Weighted correlation network analysis* - WGCNA) implementada como um pacote em R, é um algoritmo criado com o intuito de analisar dados de expressão gênica como uma rede de co-expressão (Zhang e Horvath, 2005). A rede de correlação construída pelo WGCNA é uma rede em que cada um dos nós, que representam genes ou transcritos, é ligado a todos os outros em diferentes intensidades. Esta intensidade é calculada como o módulo da correlação de Pearson entre a expressão dos nós elevada a uma constante beta que serve como um ajuste que determina quão conexa é a rede como um todo, dando um peso maior aos valores mais próximos de 1, e achatando os valores próximos a zero. Isto é feito para moldar a topologia da rede de forma a evitar que todos os genes sejam muito conexos, tirando o foco dos *hubs* mais importantes, e evitando que todos os genes sejam muito distantes, evitando que apenas os *hubs* restem com alguma conexão. Quando uma rede segue esta distribuição equilibrada de conectividade ela é dita *scale-free* (livre de escala) (Zhang e Horvath 2005). Esta medida ajustada de correlação será referida como “similaridade”.

Nas redes geradas pelo WGCNA, os genes são agrupados em “módulos” a partir das medidas de correlação de sua expressão com a de todos os outros genes, sendo atribuído a cada gene um valor de conectividade que descreve quanto este é central em uma rede, informação a partir da qual pode-se supor que a expressão deste exerce algum tipo de influência sobre os outros genes da rede (Zhang e Horvath 2005).

Esta conectividade é calculada conforme a equação abaixo, onde a_{ij} é a similaridade entre dois genes.

$$C_i = \frac{\sum_{u \neq i} \sum_{\{v | v \neq i, v \neq u\}} a_{iu} a_{uv} a_{vi}}{\sum_{u \neq i} a_{iu} ((\sum_{v \neq i} a_{vi}) - a_{ui})}$$

C_i é, portanto, a soma da similaridade de todos os genes entre si, normalizados por quanto cada um deles é similar ao gene i (se fosse uma rede sem pesos, onde a ligação entre genes é binária, este seria o número de conexões feitas pelos genes ligados a i), normalizado pela soma das similaridades de i vezes quão mais similar i é que dos outros genes da rede. Em uma analogia usando interações sociais entre pessoas, C_i é 1 se todos os seus amigos são amigos entre si, e 0 se nenhum deles se conhece (Zhang e Horvath 2005).

A fim de reduzir a dimensionalidade dos dados, para cada um dos módulos o programa faz uma análise de componentes principais (*Principal Component Analysis* - PCA) utilizando os dados de expressão dos genes do módulo. Para cada amostra utilizada na construção da rede é atribuído um *eigengene* que corresponde ao primeiro componente do PCA e que representa a expressão dos genes do módulo na amostra. Com esta síntese dos dados é possível, por exemplo, calcular a correlação de Spearman entre os *eigengenes* e características fenotípicas das amostras (ex. histologia, sobrevida, sexo, outros), permitindo a identificação dos módulos da rede de

co-expressão mais associados a estas características.

Objetivos

Objetivos Gerais

Identificar mutações somáticas no exoma e integrar com dados de alterações na expressão gênica de uma coleção de 14 casos de PDAC para identificação de vias preferencialmente alteradas por alterações somáticas e transcricionais.

Anotar funcionalmente lncRNAs regulatórios relevantes para o desenvolvimento tumoral a partir da análise de redes de co-expressão..

Objetivos Específicos

- Implementar rotinas computacionais para:

- 1) Identificar e anotar alterações somáticas (SNVs e *indels*) no exoma de 14 amostras de tumores PDAC, incluindo mutações em genes codificadores de proteínas e RNAs não codificadores longos.
- 2) Identificar vias moleculares enriquecidas em alterações transcricionais e/ou somáticas potencialmente importantes para a manutenção do fenótipo tumoral no PDAC.
- 3) Gerar rede de co-expressão gênica a partir de dados do transcrito das mesmas amostras de PDAC e identificar módulos da rede com expressão gênica associada a alterações somáticas, a malignidade e a sobrevida dos casos de PDAC avaliados.
- 4) Identificar lncRNAs de elevada conectividade na rede associados a vias moleculares enriquecidas nos diferentes módulos.

Materiais e Métodos

A figura abaixo descreve o fluxograma geral das etapas de geração e análise computacional dos dados ao longo do trabalho.

Figura 1: Fluxograma do trabalho; os programas usados nos procedimentos computacionais estão indicados nos hexágonos, os procedimentos experimentais estão representados por retângulos e as elipses representam bancos de dados usados como referência para anotação dos dados.

Amostras clínicas de PDAC

A análise de alterações genômicas deste trabalho foi realizada utilizando amostras de tumor e tecido não tumoral de pâncreas de pacientes diagnosticados com adenocarcinoma ductal de pâncreas (PDAC) e submetidos à ressecção cirúrgica no

Hospital A.C.Camargo. A ausência de tratamento neoadjuvante pré-operatório foi um critério de inclusão no estudo. As amostras de PDAC foram classificadas de acordo com o sistema TNM, recomendado pela Organização Mundial de Saúde (Bosman et al., 2010). Para aumentar o poder estatístico da análise foram selecionadas amostras de PDAC com características histopatológicas e clínicas homogêneas. Os fragmentos de tecidos analisados foram congelados em nitrogênio líquido imediatamente após a ressecção cirúrgica e fazem parte da coleção do Biobanco do A.C.Camargo, que organiza e mantém amostras com diversos controles de qualidade. A autorização para a utilização das amostras deste estudo já foi devidamente concedida pela Comissão de Ética da Instituição e se encontra registrada na Plataforma Brasil (CAAE: 15059213.0.0000.5432).

Foram obtidas 14 amostras de PDAC com estadiamento clínico I e II, que correspondem aos estágios em que os tumores são mais frequentemente operáveis, bem como tecido adjacente histologicamente normal a cada tumor, servindo como controle. Fragmentos de PDAC e de tecido adjacente foram selecionados pela médica patologista Dra. Maria Dirlei Begnami (Hospital AC Camargo). Os cortes histológicos dos fragmentos foram corados com hematoxilina/eosina e examinados por um patologista para confirmação histopatológica. Apenas foram processados para extração de RNA e DNA fragmentos de tecido tumoral em que foi estimado conter 80% ou mais de células neoplásicas e em que houve ausência de infiltrados inflamatórios. Todas as amostras analisadas possuem um relatório anatomopatológico associado, assim como

informações sobre a evolução clínica do paciente, que estão disponíveis no Anexo 1.

Sequenciamento do exoma e transcrito

Para distinguir entre mutações somáticas no exoma e polimorfismos foi feito o sequenciamento simultâneo de DNA das amostras tumorais e de DNA genômico dos pacientes, isolado de leucócitos de sangue circulante ou de tecido pancreático não tumoral em casos que não havia sangue disponível (Anexo 1). Para análise do transcrito foi feito o sequenciamento de tecido tumoral e tecido adjacente ao tumor sem evidência histológica de malignidade obtidos dos mesmos casos acima.

A preparação das bibliotecas para RNA-seq e Exome-seq foi realizada por Vinícius Ferreira da Paixão durante seu projeto de doutorado (Paixão, 2020). As bibliotecas de DNA foram preparadas com o *Nextera Rapid Capture Expanded Exome* (Illumina Inc., cat. n° FC-140-1004) que cobre 62 Mb do exoma humano, incluindo genes codificadores de proteínas, lncRNAs, além de miRNAs e um grande número de regiões UTR (*untranslated regions*). Como comparação o *Nextera Rapid Capture Exome* (Illumina Inc., cat. n° FC-140-1000) interroga apenas 37MB. As bibliotecas foram sequenciadas na plataforma HiSeq 2500 do serviço de sequenciamento do INCT Genoma Humano (IB-USP) com sequenciamento *paired-end* (2 x 75nt), sendo obtidos em média 80 milhões de *reads* para as amostras tumorais e tecido pancreático não tumoral adjacente dos mesmos pacientes. Algumas bibliotecas foram sequenciadas em uma segunda corrida para atingir uma profundidade mínima de 40x de cobertura de

sequenciamento.

As bibliotecas de RNAseq foram preparadas usando o kit *TruSeq Stranded Total RNA with Ribo-Zero* (Illumina Inc., cat. n° 20020596), para capturar a informação do RNA total (poliadenilado e não poliadenilado) depletado de RNA ribossomal, isolado das 14 amostras de PDAC e de tecido pancreático sem malignidade dos mesmos indivíduos. A tecnologia de RNA-seq apresenta diversas vantagens em relação à tecnologia utilizada na maior parte dos estudos predecessores do transcriptoma de PDAC, os microarranjos de DNA, tais como maior sensibilidade, faixa dinâmica de detecção, melhor capacidade em distinguir formas alternativas de *splicing*, expressão alélica, além de detectar RNAs sem a necessidade de se restringir a uma lista de sondas previamente selecionadas (Wang et al., 2009). Este último aspecto é fundamental para uma análise abrangente da atividade transcricional em regiões não anotadas do genoma que originam RNAs não codificadores até agora não detectados. A utilização da abordagem de RNA-seq permitiu assim a detecção e quantificação relativa do conjunto completo de RNAs não codificadores expressos em tecidos pancreáticos, tanto de tumores como de tecidos adjacentes saudáveis.

O processamento dos dados de sequenciamento e a análise do transcriptoma das amostras foram realizados pelo colega de laboratório Julio Omar Sosa durante seu doutorado (Sosa, 2019). Os dados brutos do transcriptomas foram alinhados à referência hg19 do genoma humano usando o programa TopHat2 (Kim et al., 2013) resultando em uma média de 14,7 milhões de *reads* alinhados para cada amostra. A

montagem dos diferentes transcritos expressos nos tecidos pancreáticos foi feita usando o programa StringTie (Pertea et al., 2015). Para excluir as novas isoformas encontradas que possivelmente fossem artefatos de sequenciamento foram removidos os transcritos novos com éxons com tamanho aberrante (maiores que 2.420) acima dos observados em genes anotados no GENCODE, e também os monoexônicos encontrados em regiões intergênicas ou em regiões *antisense* a genes conhecidos, deste filtro restaram 40.347 transcritos. Esta reconstrução do transcrito foi utilizada neste trabalho, sendo a quantificação relativa de cada transcrito em cada amostra estimada em TPM (*Transcripts Per Million*) usando o programa RSEM (Li e Dewey, 2011). A expressão gênica diferencial entre amostras de tumor e de tecido não tumoral foi feita usando o algoritmo DESeq2 (Love et al., 2014).

Processamento dos dados de exoma e chamada de variantes

Todas as análises informáticas deste trabalho são documentadas e integradas em *scripts* em Python/BASH paralelizáveis que executam os diferentes programas, organizam a estrutura dos dados de modo que a saída de um processo seja a entrada do seguinte, sumarizam resultados, e criam todos os gráficos de maneira automática.

A figura 2 abaixo apresenta um fluxograma com as principais etapas da análise computacional do exoma. Os dados do sequenciamento de exoma passam, primeiramente, pelo bcl2fastq (Illumina, 2017), em que são excluídos ou cortados *reads*

considerados ou de baixa qualidade, ou provenientes de adaptadores, enquanto converte os milhões de arquivos bcl em um arquivo em formato fastq para cada amostra. O mapeamento dos *reads* sobre o genoma de referência (NCBI - GRCh37) foi feito utilizando-se do BWA (bio-bwa.sourceforge.net/). A partir deste momento as análises foram realizadas de acordo com o guia de boas práticas para identificação e análise de variantes utilizando dados de sequenciamento em larga-escala (*GATK Best practices*) elaborado pelo Broad Institute (Van der Auwera et al., 2013), que é amplamente utilizado e validado pela comunidade.

Best Practices for Somatic SNPs in Whole Genomes and Exomes

Figura 2: Fluxograma de melhores práticas recomendadas pelo GATK (Benjamin et al., 2019) e utilizadas neste trabalho para o tratamento dos dados e extração de

variantes do exoma.

Para cada uma das amostras, os *reads* do exoma obtidos foram realinhados e recalibrados conforme recomendado pelo manual de melhores práticas do GATK. No entanto, para fazer a detecção de mutações somáticas, tanto inserções e deleções (*indels*) como variantes de um único nucleotídeo (SNV, *single nucleotide variants*) foram testados três algoritmos desenvolvidos paralelamente por diferentes grupos (i.e., VarScan (Koboldt et al., 2012), Mutect2 (Cibulskis et al., 2013), e LoFreq (Wilm et al., 2012). As mutações encontradas por cada um dos algoritmos e suas intersecções estão apresentados na figura 3.

Figura 3: Diagrama de *Venn* com as intersecções entre mutações detectadas pelos três algoritmos de chamada de variantes testados.

Inicialmente era planejado considerar as mutações encontradas por quaisquer dois dos três algoritmos. Porém chamou a atenção a baixa sobreposição entre as

variantes chamadas por VarScan e LoFreq (2.238 mutações, 1,2%). Além disso, uma análise mais detalhada dos resultados mostrou que tanto LoFreq quanto VarScan excluíram variantes sabidamente relevantes em PDAC, como a mutação oncogênica *driver* frequente G12D em KRAS (Prior et al., 2012). Com base nestes resultados, optamos por utilizar apenas os variantes identificados pelo Mutect2 nas análises subsequentes (Benjamin et al., 2019).

Foram utilizados os parâmetros padrão recomendados na documentação do Mutect2 para a chamada de variantes, com exceção do parâmetro opcional que determina o número máximo de *reads* alternativos que podem ser detectados na amostra controle. O valor padrão é tal que a alteração é considerada germinativa se houver um ou mais *reads* alternativos na amostra controle, e com esse valor estrito alterações em regiões com alta profundidade de sequenciamento e que continham alguns *reads* alternativos na amostra controle eram excluídas mesmo tendo uma frequência de *reads* alternativos menor que 1%. Isso é um problema pois é esperado que existam erros de sequenciamento, especialmente e em regiões cobertas centenas de vezes em que há portanto uma probabilidade maior de que um destes erro seja localizado. Por essa razão, o parâmetro `-max_alt_allele_in_normal_fraction` foi adicionado de forma que as alterações seriam excluídas apenas se a porcentagem de *reads* alternativos detectados no controle ultrapassasse 10%.

Algumas das alterações chamadas como variantes somáticas podem na realidade corresponder a polimorfismos encontrados no genoma humano. Além do uso

do dbSNP (Kitts et al., 2002) para excluir eventuais polimorfismos da análise de chamada de variantes, também foram excluídas alterações descritas em outros bancos do tipo, como o gnomAD (Karczewski et al., 2019) e o ABraOM (Naslavsky et al., 2017), este último representando polimorfismos presentes em uma amostragem de indivíduos idosos da população brasileira.

Para a anotação das variantes encontradas foi utilizado o VEP (Yourshaw et al., 2015) que indica a que gene pertence a variante, e o tipo da mutação observada: i) regiões não codificadoras, ii) mutações *missense* (substituições simples de aminoácidos), iii) mutações *nonsense* (como *frameshifts*, e alterações de início ou fim da tradução), iv) mutações silenciosas. Para variantes localizadas na região codificadora de proteínas é indicado também qual o impacto esperado na estrutura da proteína. As variantes e genes mutados foram também anotados em relação a bancos de dados COSMIC (*Catalogue of Somatic Mutations in Cancer*, Forbes et al., 2016), uma base de dados curada que cataloga mutações observadas em humanos a partir de milhares de artigos científicos, e CCGD (*Candidate Cancer Gene Database*, Abbot et al., 2015) que cataloga resultados de ensaios genéticos baseados em elementos transponíveis (*transposon-based forward genetic screens*) em camundongos, em que são observados quais genes contribuem para a tumorigênese dos camundongos quando afetados pela inserção de um elemento transponível.

O passo final para a anotação das alterações somáticas foi utilizar o algoritmo OncodriveFML (Mularoni et al., 2016) para priorizar mutações com impacto funcional

para a transformação ou manutenção de fenótipos tumorais, as distinguindo de mutações passageiras. Este método analisa o impacto das mutações na estrutura de proteínas (apenas quando se trata de uma região codificadora), verifica se elas podem interferir na ação de fatores de transcrição ou de micro RNAs e procura por sinais de seleção positiva através da distribuição maior que ao acaso destes sinais no conjunto dos dados de todos os pacientes envolvidos no estudo, e após avaliar globalmente essas características atribui um p-valor para rejeitar ou não a hipótese nula de que a mutação no gene é uma mutação aleatória passageira.

O número de mutações somáticas simples identificadas no estudo foi comparado com o número de variantes somáticas identificadas por projetos conduzidos por consórcios internacionais que realizaram o sequenciamento com alta cobertura do exoma (TCGA/PAAD-US, <https://portal.gdc.cancer.gov/>) e do genoma completo (ICGC/PACA-CA, <https://dcc.icgc.org/>) de centenas de casos de PDAC.

Enriquecimento de categorias gênicas

Um tipo de ferramenta muito importante para que seja possível categorizar, visualizar e interpretar as listas de genes criadas nas diferentes etapas deste trabalho é o enriquecimento de vias. Este tipo de ferramenta que já teve várias implementações, como o gProfiler (Reimand et al., 2016) usado por este estudo, parte de uma lista de genes e verifica quantos destes se encontram em catálogos de vias moleculares. Estes

catálogos constituem-se de listas organizadas hierarquicamente que agrupam genes de acordo com seus envolvimento em processos biológicos, de acordo com características estruturais de suas proteínas, de acordo com o tipo de reações enzimáticas que promovem, ou de acordo com outras características como localização subcelular, por exemplo. Um teste estatístico de Fisher (teste hiper-geométrico) verifica se a proporção de genes em comum entre a lista selecionada e a listas de referência é significativa ou se pode ser explicada pelo acaso. Para este trabalho foram utilizadas as anotações funcionais dos bancos de dados *Gene Ontology (Biological Process)* (Gene Ontology Consortium, 2019), KEGG (Kanehisa et al., 2000), *Reactome* (Fabregat et al., 2018), e *WikiPathways* (Kelder et al., 2009).

Quando necessário sintetizar ainda mais os resultados devido ao grande número de vias significativas encontradas foi usado o algoritmo de clusterização *k-means* para sumarizar as vias encontradas em função de quais genes estas contém. O menor número de grupos possível que não resultasse em grande perda de informação foi escolhido utilizando o critério bayesiano de informação (Pelleg e Moore, 2000).

Montagem de redes de co-expressão de genes codificadores de proteína e lncRNAs

O programa WGCNA (Zhang e Horvath, 2005) foi usado para criar uma rede com os dados de expressão gênica dos 14 pares de amostras de tumores de pâncreas e tecido não tumoral adjacente. Após a normalização dos dados pela função rLog

utilizando o pacote DESeq2 (Anders e Hubber, 2010), os dados de contagem dos 40.347 transcritos detectados foram filtrados para eliminar dados com alto ruído da análise. Foram excluídos os transcritos que apresentassem i) intensidade média nas amostras menor que 0,075 TPM ou ii) desvio padrão menor que 0,75. Após essa etapa de filtragem, restaram 23.951 transcritos (26,5% dos detectados), sendo que 23,8% destes são lncRNAs, que em seguida foram usados para a construção de uma rede de co-expressão. Os valores de filtragem foram escolhidos empiricamente de forma a minimizar o tamanho da rede, e ao mesmo tempo incluir a maior parte (> 95%) dos 1.787 transcritos diferencialmente expressos (pvalor 0.05, fold-change 2) entre as amostras de tumor e tecido não tumoral, dos quais 1.069 são codificadores de proteína e 718 são lncRNAs. Após essa etapa de filtragem, restaram 23.951 transcritos (26,5% dos detectados), sendo que 23,8% destes são lncRNAs, que em seguida foram usados para a construção de uma rede de co-expressão.

Para o cálculo dos valores de similaridade usados na construção da rede, a matriz de correlação entre todos os pares de transcritos foi ajustada usando um fator beta exponencial igual a 8, escolhido por resultar na formação de agrupamentos gênicos mais próximos a uma distribuição livre de escala (dados não mostrados). Em seguida os transcritos foram agrupados em função da similaridade segundo o algoritmo *dynamicTOM overlap* formando 23 “módulos”. Estes módulos são identificados seguindo um esquema de cores atribuídas aleatoriamente. A figura abaixo apresenta os módulos definidos na análise de rede de co-expressão.

Figura 4. Rede de co-expressão gênica em 14 pares de amostras de tumor e tecido não tumoral de pâncreas gerada pelo WGCNA (Zhang e Horvath, 2005). A conectividade dos 23.951 transcritos considerados na análise foi utilizada como medida de distância para a construção do dendrograma. O algoritmo *dynamicTOM overlap* foi usado para fazer o agrupamento dos transcritos em módulos da rede. Na barra horizontal, transcritos com a mesma cor se encontram no mesmo módulo. Foram anotados 23 módulos no total.

Para inferir a função dos lncRNAs diferencialmente expressos utilizando a informação da rede de co-expressão, foi feita uma análise de enriquecimento de vias utilizando transcritos com funções conhecidas já anotadas presentes em cada um dos módulos. Posteriormente, foram selecionados para inspeção manual módulos da rede enriquecidos em vias moleculares relevantes para a tumorigênese a fim de localizar lncRNAs com alta conectividade local (apresentando um comportamento *hub* levando apenas em conta os genes da via escolhida) indicando assim que este possa exercer algum tipo de regulação sobre a via ou alguns de seus membros.

Validações Experimentais

Amostras de tecido tumoral e controle, bem como amostras de tumores enxertados em camundongos imunodeficientes (PDX, preparadas pelo colaborador de laboratório Luís Bruno da Cruz e Alves de Moraes), e de linhagens celulares (PANC-1, MIA Paca 2, Capan 2, AsPC 1, e BxPC 3) foram utilizadas para quantificar os lncRNAs candidatos selecionados a partir da análise de redes de co-expressão em PCR quantitativo em tempo real (RT-qPCR) por outros membros do laboratório (Doutorando Vinicius Ferreira da Paixão e a técnica Bianca Dazzani). O DNA extraído das amostras de tecido tumoral foram utilizadas pela pós-doutora Cláudia Alessandra Andrade de Paula (CIPE - AC Camargo Cancer Center) para fazer o ressequenciamento com alta profundidade, da ordem de dezenas de milhares de vezes, de um *amplicon* que cobre os códons 12 e 13 do gene KRAS.

Resultados

Análise do exoma

Processamento de dados de sequenciamento e chamada de variantes

Inicialmente foi avaliada a cobertura e profundidade de sequenciamento (n° *reads*/base) dos éxons capturados nas bibliotecas do exoma de tumor e tecido não tumoral/sangue (controle) dos 14 casos de PDAC. A cobertura mínima das amostras foi de 40x do exoma de 62Mb avaliado.

Figura 5: Número de *reads* alinhados no genoma para cada amostra avaliada na análise de exoma. Barras azuis para as amostras controles e vermelhas para os tumores.

Apesar da exigência de que houvesse pelo menos uma cobertura média de 40

vezes por amostra, é plausível que esta cobertura não se distribua uniformemente para todos os 55.766 genes interrogados na plataforma de captura utilizada. De fato, ao medirmos a cobertura média de cada gene e construirmos um histograma, observamos tanto nas bibliotecas de tumor quanto controles a existência de genes com cobertura muito maior que a média, e que aproximadamente 10% dos genes têm cobertura menor que 20 vezes, (figura 6)

Figura 6: Histogramas com a cobertura média de genes nas amostras tumorais (painel superior) e controle (painel inferior). Em preto o histograma dos genes em função da cobertura média nas amostras. O eixo X está em escala log. O eixo Y da

direita mede a fração dos genes que tem cobertura igual ou maior ao valor do eixo X, mostrado pela linha vermelha.

Como observado na figura acima alguns genes possuem uma cobertura média muito superior a outros. Os genes REXO1L2P, e USP17L6P possuem cobertura média de aproximadamente 2500 vezes, e os genes GAGE12D, CT45A1, LOC100132247, e REXO1L1 possuem cobertura de aproximadamente mil vezes. Quanto aos genes mais citados em relação a desenvolvimento tumoral, BRCA2 possui uma cobertura média de aproximadamente 60 vezes, TP53 e KRAS aproximadamente 50 vezes, MAPK1 e SMAD4 aproximadamente 40 vezes, e CDKN2A 27 vezes. Os genes TTN e MUC13, conhecidos por conterem grande número de mutações aleatórias devido ao fato de possuírem longas sequências codificadoras, possuem cobertura média de aproximadamente 60 vezes. O Anexo 2 mostra a profundidade de cada gene em cada amostra.

Após a análise do Mutect2 foram detectadas um total de 28.683 mutações em 12.438 genes, distribuídas entre as diferentes amostras, sendo deste total 2818 indels (com 6,7 bases de extensão média) e 25530 SNVs. O gráfico de dispersão abaixo (figura 7) compara o número de mutações simples encontradas (SNVs e *indels*) e o total de *reads* alinhados em cada amostra. O eixo X está em escala log, podendo ser notado que há uma separação entre dois grupos de amostras, algumas com números bem maiores de alterações (aparentemente mais genomicamente instáveis), e outras com números menores de alterações (aparentemente mais genomicamente estáveis).

É importante notar que não há correlação significativa entre a profundidade de sequenciamento e o número de alterações somáticas encontradas. O Anexo 3, no formato MAF (*Mutation Annotation Format*), mostra todas as alterações encontradas.

Figura 7: No eixo Y o número de *reads* alinhados para cada uma das amostras, no eixo X o número de alterações encontradas pelo Mutect2 em escala log. Uma função logarítmica foi ajustada aos pontos pelo método dos mínimos quadrados, a correlação dos dados é de 0,31, p-valor 0,27.

Para estimar a sensibilidade da chamada de mutações a partir da análise do exoma, foi feito o sequenciamento com alta profundidade (pelo menos 10^4 x *reads* por amostra) de uma região do oncogene KRAS, cobrindo os códons 12 e 13 que são um conhecido *hotspot* de mutações. Na figura 8 abaixo é possível comparar as frequências alélicas obtidas no sequenciamento de alta profundidade e a detectada nos dados de

exoma gerados neste trabalho.

Figura 8: Comparação da frequência alélica de *reads* alterados/*reads* selvagens (eixo y) nos códons 12 e 13 obtida pelo sequenciamento do exoma (colunas laranjas) ou pelo sequenciamento específico de *amplicon* em altíssima profundidade (colunas marrom) para cada amostra de tumor. A barra de cores inferior indica se a amostra foi anotada como mutada (verde) ou não (vermelho) em KRAS pelo programa Mutect2.

Como observado na figura acima, as frequência alélicas de variantes de KRAS determinadas pelo sequenciamento do exoma em nosso estudo foram comparáveis às estimadas pelo sequenciamento com altíssima cobertura. Apenas uma amostra considerada sem mutação pelo Mutect2, a amostra 29T, que possui uma frequência alélica maior de 5% no sequenciamento de alta profundidade mas não possui uma

frequência apreciável nos dados do exoma (apenas 1 *read* alterado de um total de 113). O Mutect2 excluiu a mutação de KRAS na amostra 44T por considerá-la germinativa (20% dos *reads* alterados na amostra controle). De fato, o sequenciamento do *amplicon* desta amostra indicou uma frequência alélica próxima a 50%, o que está de acordo com esta hipótese, e por isso foi excluída pelo Mutect2. As outras amostras que foram consideradas mutadas pelo sequenciamento de alta profundidade mas não pelo Mutect2 possuem baixa frequência alélica, próxima ao limite de detecção do algoritmo. As amostras 04T e 47T foram consideradas mutadas pelo Mutect2 mas em outros códons do KRAS.

Devido à instabilidade genômica característica dos tumores pancreáticos era esperado que fosse encontrado um grande número de mutações de baixa prevalência entre as amostras (Collisson et al., 2011, Moffitt et al., 2015, Raphael et al., 2017). Como pode ser visto no painel A da figura 9, as 28.683 alterações somáticas detectados se encontram em coordenadas genômicas que em sua maioria estão alteradas em um único paciente. Para obter suporte independente para a relevância de mutações presentes nesse conjunto, foi verificada a presença de alterações que ocorressem nas mesmas coordenadas de mutações já anotadas no banco de dados COSMIC. Observamos que 3.314 alterações, 12% do total detectado neste estudo já se encontram documentadas no COSMIC em algum tipo de tumor (Figura 9B).

Figura 9: Histograma com número de alterações somáticas (SNVs e *indels*) detectadas (em escala log10) em função do número de amostras em que a mesma coordenada mutada foi detectada. Painel A; total de mutações detectadas. Painel B; apenas mutações já descritas no COSMIC.

Para avaliar se o número de mutações simples e genes mutados identificados em nosso *pipeline* era compatível com o observado em outros estudos de análise genômica já realizados em PDAC, comparamos o número médio de mutações simples e genes mutados por paciente detectado em nosso trabalho com os valores correspondentes reportados pelos consórcios ICGC (PACA-CA) e TCGA (PAAD-US). Os resultados estão apresentados na Tabela 1. No caso do PACA-CA, onde foi realizado o sequenciamento do genoma completo, foram consideradas apenas as alterações somáticas que se encontram nas mesmas regiões do exoma avaliadas em nosso trabalho (62 Mb, *Nextera Rapid Capture Expanded Exome*). Observamos um número médio de mutações por paciente em nosso estudo (2048,8) comparável com o documentado pelo consórcio ICGC/PACA-CA (2285,3). Enquanto a maior parte das

variantes detectadas pelo ICGC/PACA-CA (56%) já está documentada no COSMIC, apenas 12% das variantes identificadas em nosso estudo já estão anotadas nessa base de dados.

No caso do PAAD-US, não estão documentados na publicação os intervalos do exoma utilizado na análise e optamos por considerar todas as mutações simples reportadas (Raphael et al., 2017). É notável o número significativamente menor de mutações reportadas pelo consórcio TCGA/PAAD-US (267,2) em relação ao nosso estudo e ao PACA-CA. A utilização de procedimentos mais estritos para a chamada de variantes pelo PAAD-US pode explicar o menor número de alterações detectadas nesses dados em relação ao nosso estudo e às reportadas pelo ICGC/PACA-CA. Curiosamente, apenas 24% das mutações detectadas pelo PAAD-US já estão reportadas no COSMIC. O número médio de genes mutados por paciente em nosso estudo (69,7; 8% do total) é comparável ao detectado pelo PACA-CA. (81,5; 6%) e PAAD-US (10,1; 6%).

	Número médio de alterações somáticas detectadas por paciente		
	Este trabalho (n=14)	*ICGC/PACA-CA (n=268)	TCGA/PAAD-US (n=177)
Mutações simples	2048,8	2285,3	267,2
Mutações encontradas no COSMIC	236,7 (12%)	1272,1 (56%)	63,3 (24%)
Genes mutados	888,4	1377,2	169,0
Genes mutados encontrados no CCGD	69,7 (8%)	81,5 (6%)	10,1 (6%)
Mutações encontradas no COSMIC e CCGD	20,8	77,8	4,4

Tabela 1: Número médio por paciente (indicado por 'n') de mutações simples (SNVs ou *indels*) detectadas neste trabalho e reportadas em projetos do ICGC (PAAD-US, PACA-CA). Estão também listadas o número de mutações representadas nos bancos COSMIC e CCGD.

* Dados de genoma completo filtrados aceitando apenas os SNVs que se encontram nos intervalos do exoma usado neste trabalho.

O Anexo 4 contém a informação para os 12.438 genes com mutações detectadas nas amostras de PDAC, incluindo o q-valor associado, em quantos pacientes é alterado, se está no CCGD, e quantas de suas alterações estão reportadas no COSMIC.

Priorização e anotação de genes com mutações relevantes no PDAC

A partir dos 12.438 genes mutados em pelo menos um paciente, o programa OncodriveFML foi utilizado para priorizar genes com mutações somáticas que contribuem para o desenvolvimento tumoral e distingui-las de mutações passageiras utilizando como critérios a prevalência nas amostras, a distribuição local das mutações (*hotspots*), a predição de alteração estrutural na proteína ou RNA, entre outros (Mularoni et al., 2016). Utilizando um p-valor ajustado de 0,05 foram selecionadas 4.660 genes mutados. Para filtrar ainda mais a lista de genes com mutações relevantes, optamos por excluir genes que, nos dados de transcritômica destas amostras, tivessem expressão média menor que o valor arbitrário de $TPM > 0.1$, seguindo a lógica de que se um gene não é expresso nem no tecido sadio nem no tumoral, a sua mutação possivelmente não tem consequências funcionais para o fenótipo tumoral. Com isso reduziu-se o número de genes candidatos para 3.643. Posteriormente foi utilizado um critério ainda mais estrito de $p_{adj} < 5e-6$ e $TPM > 0.1$ que resultou em 1.603 genes (~12% dos genes mutados) contendo 5.747 mutações simples (SNVs e *indels*), uma média de 3,6 mutações por gene, que foram utilizados nas análises subsequentes, sendo 44,6% deste total em regiões não codificadoras.

Em seguida a ferramenta de anotação VEP (Yourshaw et al., 2015) foi utilizada para anotar as alterações somáticas quanto a alteração na proteína codificada pelo gene. Foram identificadas 1.883 mutações em regiões não codificadoras, 609 mutações *nonsense*, e 2.745 mutações *missense*. As mutações silenciosas foram removidas das análises subsequentes.

Figura 10: Anotação das mutações usando VEP classifica como as alterações simples afetam a estrutura primária das proteínas codificadas.

Outra forma de anotação utilizada para validar as mutações somáticas simples identificadas neste estudo foi a comparação das coordenadas com as mutações já anotadas no banco de dados COSMIC (Forbes et al., 2016). Adicionalmente, para priorizar genes funcionalmente relevantes no PDAC os dados foram comparados com genes anotados no CCGD (Abbot et al., 2015). O CCGD (*Candidate Cancer Gene Database*) cataloga resultados de triagens genéticas baseados na inserção aleatória de elementos transponíveis (*transposon-based forward genetic screens*) em camundongos, onde foram identificados genes que contribuem para a tumorigênese dos camundongos quando perturbados pela inserção de um elemento transponível. Cerca de 12% das mutações detectadas (n=749) e 17% dos genes mutados (n=277) nas amostras de PDAC já foram anotadas no COSMIC ou no CCGD, respectivamente. Adicionalmente, observamos que 291 mutações detectadas em nosso estudo estão representadas simultaneamente no COSMIC e em genes reportados pelo CCGD. A lista anotada das mutações simples e genes mutados detectadas neste trabalho estão

sumarizadas no Anexo 4.

A fim de priorizar os genes com alterações somáticas mais prevalentes nas amostras analisadas, aos critérios descritos acima foi aplicado um filtro adicional que exigia a presença em ao menos 4 dos 14 casos de PDAC avaliados no estudo, resultando em 18 genes (Figura 11). Neste conjunto encontram-se genes já reportados na literatura como alterados no câncer de pâncreas como KRAS, ARID1B, e UBR5 (Ying, et al., 2016; Shearer et al., 2015), e genes já descritos com alterações somáticas em outros tipos de tumores como ZFHX3 (Walker et. al, 2015). Ao conjunto anterior foram incluídos genes com alterações somáticas simples no PDAC destacados em estudos que avaliaram dezenas de amostras de pacientes (Waddell et al., 2015; Raphael et al., 2017), mesmo quando não selecionados nos filtros mencionados acima (Figura 11). Nesse conjunto adicional incluem-se genes sabidamente relevantes no PDAC como TP53 (mutado em 5 amostras), que não está incluído no CCGD, BRCA2 (mutado em 4 amostras) e CDKN2A (mutado em 3 amostras), que estão abaixo do limiar de significância considerado ou não foram anotados no CCGD, além de genes significativos em nossa análise mas mutados em menos do que 4 amostras, como SMAD4, BRCA1, ATM e BRAF.

Figura 11: *Heatmap* com genes recorrentemente mutados no PDAC identificados neste estudo e/ou na literatura. Além de 18 genes i) significativamente mutados ($q\text{-valor} < 5 \times 10^{-6}$), ii) expressos nas amostras (TPM médio > 0.1), iii) já listados no CCGD e COSMIC e iv) nos quais tenham sido detectadas alterações somáticas em

ao menos 4 pacientes, foram incluídos 42 genes que não satisfaziam todos os critérios acima mas que foram descritos em trabalhos reportando alterações genômicas recorrentes no PDAC (Waddell et al., 2015, em laranja; Raphael et al., 2017, em marrom). As mutações (ordenadas pelo q-valor) detectadas em cada um dos 14 pacientes (colunas da direita, ordenadas em função do número total de mutações) estão anotadas como “*nonsense*” (vermelho), “*missense*” (verde) ou e não codificadoras (azul). A coluna “q-valor” indica a significância estatística das mutações no gene atribuída pelo programa OncodriveFML. A coluna “Classe CCGD” indica o potencial oncogênico relativo do gene (A = top 10% , B = 11 a 25%, C = 26 até 50%, e D = menor que 50%). A coluna “Fração COSMIC” indica a proporção de mutações detectadas nos genes que são confirmadas no catálogo de mutações COSMIC.

Vias moleculares enriquecidas entre genes significativamente alterados

Para melhor compreender o impacto funcional das alterações somáticas detectadas na biologia do PDAC foi feita uma análise de enriquecimento de vias usando gProfiler (Reimand et al., 2016). Como entrada foi fornecida a lista de 1.603 genes considerados significativamente mutados utilizando os critérios de seleção descritos acima, ordenada em ordem crescente de q-valor. Utilizando anotações funcionais dos repositórios *Gene Ontology (Biological Process)*, KEGG, *Reactome*, *WikiPathways*, e *Human phenotype ontology*, foram encontrados 696 termos significativamente enriquecidos ($p < 0,05$, Anexo 5), sendo os mais significativos apresentados na figura 12 abaixo.

20 vias mais enriquecidas pelos dados de alterações somáticas

Figura 12: Vinte vias mais enriquecidas pelos genes com mutações somáticas no PDAC, identificadas com o programa gProfiler. O p-valor ajustado referente ao enriquecimento da categoria (eixo Y da esquerda) está representado por uma linha vermelha. O número de genes alterados nas vias está indicado pelas barras azuis (eixo Y direito).

A fim de sumarizar a informação das vias enriquecidas entre os genes mutados em PDAC foi usado o algoritmo *k-means* para agrupar termos semelhantes em função

da sobreposição dos genes que os compõem. Foram obtidos seis grupos contendo vias moleculares predominantemente relacionadas aos termos abaixo:

- 373 vias, se destacando as relacionadas a desenvolvimento neuronal, vias envolvidas em câncer, e citoesqueleto
- 127 vias, se destacando as relacionadas a crescimento celular, desenvolvimento de tecidos e motilidade celular
- 48 vias, se destacando as relacionadas a fosforilação de proteínas, e sinalização celular
- 44 vias, se destacando as relacionadas a sistema imune e proteólise
- 43 vias, se destacando as relacionadas a apoptose, e resposta a estímulos químicos
- 9 vias, todas relacionadas a transporte de macromoléculas

Várias das vias consideradas significativamente enriquecidas não apresentaram mutações em todas as amostras avaliadas. Contudo, é plausível que as vias críticas para a malignidade do PDAC apresentem mutações na maior parte dos pacientes, ainda que em diferentes genes da via. Portanto, foram selecionadas as vias moleculares que apresentam mutações em pelo menos um gene da via, em todas as amostras avaliadas, resultando em 360 vias moleculares (Anexo 5). Espera-se que a probabilidade de que uma dada via apresente mutações em todas amostras aumente em função do número de genes que a compõe, e para reduzir este viés, foram avaliadas as vias enriquecidas pelo menor número de genes, e que todas as amostras tenham ao menos um desses genes mutados. Entre as vias selecionadas com mutações em todas as amostras encontram-se as vias de '*Positive regulation of MAP*

kinase activity e *G protein signaling pathways*, que contêm KRAS, um gene “driver” na oncogênese do PDAC (Figura 13).

Figura 13: *Heatmap* mostrando genes alterados que se encontram nas vias ‘*Positive regulation of MAP kinase activity*’ e ‘*G protein signaling pathways*’, os pacientes estão listados nas colunas, ordenados pelo número total de mutações, estando os mais mutados à esquerda. Os genes envolvidos estão listados nas linhas e ordenados pelo p-valor do OncodriveFML, sendo o menos passageiro acima. Em cada linha estão em branco os pacientes que não possuem alterações neste gene, os que possuem alterações estão coloridos de acordo com a classificação destas pelo

programa VEP.

Também estão enriquecidas em genes com alterações somáticas em todos os casos as vias '*Neurotrophin signaling pathway*', '*Chronic myeloid leukemia*' (11 genes), '*FLT3 Signaling*', '*Focal Adhesion-PI3K-Akt-mTOR-signaling pathway*' e outras que podem ser mais profundamente estudadas como relevantes na patofisiologia do PDAC (Anexo 5).

Análise de assinaturas mutacionais no PDAC

As alterações somáticas de ponto (SNVs) identificadas foram analisadas na tentativa de identificar assinaturas mutacionais associadas a processos mutagênicos distintos. Usando o pacote *signeR* para avaliar padrões de substituição de bases predominantemente encontrados entre as alterações somáticas detectadas nas amostras e identificar assinaturas mutacionais *de novo* (Rosales et. al, 2016). Foi possível observar a existência de duas assinaturas distintas (Figura 14). A assinatura S1 é caracterizada predominantemente pela alteração frequente de C>T, em dinucleotídeos CpG. A assinatura S2 é caracterizada por alterações do tipo C>A. Como esperado, a contribuição de cada amostra para as assinaturas identificadas é proporcional ao número de mutações identificadas nas amostras (Figura 15)

Figura 14: As SNVs detectadas nas 14 amostras de PDAC contribuem para duas assinaturas mutacionais distintas.

Figura 15: Intensidade das assinaturas S1 (vermelho) e S2 (amarelo) em cada uma das amostras (eixo vertical esquerdo). O número de mutações simples detectadas (azul) em cada amostra também está indicado (eixo vertical direito).

Ao comparar as assinaturas encontradas com as assinaturas anotadas no

COSMIC se nota que a S1 é mais semelhante a assinatura COSMIC 1 (similaridade cosseno = 0,96), associada a mutações esporádicas devido a desaminação de citosinas metiladas que se acumulam em indivíduos idosos. A assinatura S2 se mostrou mais semelhante (similaridade cosseno = 0,83) a assinatura COSMIC 29, descrita como frequentemente associada a alterações somáticas relacionadas ao tabaco. No entanto, não foram observadas correlações entre as razões S1/S2 atribuídas às amostras e a idade ou o status de tabagismo (fumante/não fumante) dos casos avaliados.

	S1	S2	
COSMIC 1	0.96	0.14	Mutações iniciadas pela desaminação espontânea de 5-metilcitosina
COSMIC 29	0.32	0.83	Associado ao hábito de mascar tabaco

Tabela 2: Similaridade entre as assinaturas encontradas neste estudo (S1 e S2) e as assinaturas anotadas pelo COSMIC. São mostradas apenas as assinaturas do COSMIC com maior similaridade para cada uma das assinaturas deste trabalho (COSMIC 1 e COSMIC 29), e uma breve descrição a respeito de quais os processos mutagênicos provavelmente envolvidos.

Construção de redes de co-expressão

A análise utilizou dados de expressão e a reconstrução do transcrito de tecido tumoral e não tumoral de pâncreas que foram processados pelo doutorando e colega de laboratório Omar Julio Sosa (Sosa, 2019) desde os primeiros passos de controle de qualidade até a análise de expressão diferencial. A construção de redes de

co-expressão utilizou a ferramenta WGCNA (Zhang e Horvath, 2005), implementada como pacote em R, que cria uma rede em que todos os transcritos são conectados entre si em função de parâmetros obtidos através do módulo das correlações, e atribui aos transcritos valores como “conectividade” e “similaridade” em rede que definem quais os transcritos são os mais centrais (à rede como um todo ou em relação a algum *subcluster*), a partir de quão mais correlacionado ele é com seus vizinhos de rede em comparação com quão correlacionados os próprios vizinhos são entre si. O racional empregado nesta análise é que a observação de uma conectividade local acima do esperado permite propor uma associação funcional ou regulatória entre lncRNAs ainda não caracterizados e grupos específicos de transcritos codificadores de proteína com função conhecida.

A rede de co-expressão construída pelo WGCNA a partir da clusterização dos 23.951 transcritos resultou em 22 *subclusters* (chamados de módulos), identificados por cores (Figura 16 e Anexo 6). Um conjunto de 14.897 transcritos (62% do total) com baixa conectividade não foi incluído em nenhum módulo da rede.

Figura 16: *Heatmap* representando a matriz de correlação, com a qual são calculadas a similaridade e a conectividade entre os transcritos na rede de co-expressão. Um método de clusterização separou automaticamente os dados em 22 módulos de transcritos agrupados por sua co-expressão, na barra superior estão coloridas as colunas conforme o módulo na qual foram agrupadas. Em cinza os 62% dos transcritos que não pertencem a módulo algum.

Vias moleculares enriquecidas por genes presentes nos módulos da rede foram identificadas usando gProfiler (Reimand et al., 2016). Para cada um dos módulos gerados pelo WGCNA, foram fornecidas listas com os genes ordenados em função da sua conectividade local. As vias enriquecidas consideradas significativamente

enriquecidas (p-ajustado < 0,05) foram inspecionadas manualmente e foram selecionadas vias consideradas mais relevantes para a malignidade ou função do tecido pancreático. A figura 17 abaixo apresenta o número de genes e lncRNAs em cada módulo da rede, a correlação média dos transcritos com o fenótipo tumoral e vias moleculares representativas que se encontram enriquecidas em cada módulo. Nota-se que enquanto alguns módulos concentram grande quantidade de genes diferencialmente expressos entre as amostras de tumor e de controle, outros são pouco alterados. O módulo com maior porcentagem de diferencialmente expressos (sendo quase todos com expressão diminuída em tumores) é enriquecido em transcritos associados ao termo chamado '*Pancreatic Secretion*' (p-ajustado = 2×10^{-12}), o que está de acordo com a perda da função pancreática no tecido neoplásico desdiferenciado. O módulo magenta foi o que apresentou maior correlação positiva com o fenótipo tumoral, e entre as vias enriquecidas neste módulo se encontram vias relacionadas glicosilação de proteínas em geral e mucinas em particular, proteínas de membranas conhecidamente relevantes no PDAC (Kaur et al., 2013). Outros módulos positivamente correlacionados ao fenótipo tumoral estão enriquecidos em vias relacionadas a *Drug metabolic Process* (módulo azul cobalto), *Cell Cycle* e *DNA repair* (módulo verde-limão), *Cell adhesion* e *Immune system processes* (módulo vermelho), entre outras relevantes para a malignidade (Figura 17).

Figura 17: Principais vias moleculares enriquecidas nos módulos da rede de co-expressão gênica de tecidos pancreáticos. O comprimento das barras representa o número de transcritos contidos em cada módulo, com uma linha vertical indicando quantos são os não codificadores (eixo inferior). Para cada módulo é indicado por um círculo a média das correlações de Spearman entre os transcritos e o fenótipo tumoral

(1 para amostras tumorais; 0 para amostras controle), sendo os círculos pretos correlações negativas e os círculos brancos correlações positivas (eixo superior).

Integração entre dados de alterações somáticas e transcricionais

Comparação de vias moleculares enriquecidas

Utilizando a mesma metodologia descrita na análise dos dados do exoma, foi feita a análise de enriquecimento de vias do transcritoma. Utilizando como entrada os 8.054 genes diferencialmente expressos nas amostras tumorais em relação a tecido não tumoral adjacente ($p\text{-valor} < 0,05$) foram encontrados 163 termos significativamente enriquecidos ($q\text{-valor} 0,05$). A intersecção destes com os 696 termos enriquecidos entre genes com mutações somáticas resultou em 83 vias moleculares encontradas em ambos os casos (Figura 18). Entre estas 83 encontram-se vias relacionadas com as funções *Ras protein signal transduction*, *apoptotic process*, *small GTPase mediated signal transduction* e *cell adhesion*. Nestas 83 vias foram encontrados 425 genes somaticamente alterados, e 4.474 genes com expressão alterada, sendo que 216 genes que apresentam simultaneamente alterações somáticas e transcricionais. A lista das categorias enriquecidas e os genes representados está disponível no Anexo 7.

Figura 18: Diagrama de Venn com a intersecção entre as vias moleculares enriquecidas pelos genes com alterações somáticas ou transcricionais no PDAC.

Associação entre características do exoma e módulos da rede de co-expressão

Para inferir a existência de associações funcionais entre as mutações somáticas e alterações transcricionais nos módulos da rede de co-expressão, foram calculadas as correlações de Spearman entre a presença de alterações somáticas em genes recorrentemente mutados com a expressão gênica de módulos da rede. Para sumarizar a expressão de todos os genes do módulo foram utilizados valores de *eigengenes* representativos de cada amostra de tumor e tecido não tumoral. Em um dado gene, a presença ou não de mutações nas diferentes amostras é representada por um vetor binário que assume valores de 1 (mutado) ou 0 (não mutado). A medida de correlação entre um vetor contendo predominantemente valores iguais a 0 com outro vetor qualquer representando características moleculares ou fenotípicas (ex. expressão gênica, sobrevida, etc.) não será informativa. Assim dos 425 genes com mutações somáticas encontrados em vias localizadas na intersecção da figura 18, apenas os 40 genes com mutações detectadas em 6 ou mais pacientes foram

selecionados para verificar a existência de correlação entre a presença de mutações e os *eigengenes* da expressão nos módulos da rede, como descrito abaixo (Figura 19). Entre os 40 genes com alta prevalência de mutações somáticas nas amostras avaliadas, 11 apresentaram correlação altamente significativa ($p < 1 \times 10^{-3}$) com *eigengenes* dos módulos da rede de co-expressão: ARHGEF4, SIPA1L1, ZFHX4, CSMD3, CELSR1, OBSCN, TIAM1, TP53, TLN2, KRAS, TBC1D1 (Figura 19). A mesma abordagem foi utilizada para detectar associações entre os *eigengenes* da expressão nos módulos da rede e i) a frequência alélica de mutações nos códons 12 e 13 de KRAS obtida por sequenciamento de *amplicons*, ii) o número de mutações detectadas no tumor, iii) a contribuição relativa das assinaturas mutacionais S1/S2 (razão signeR) e iv) a sobrevida dos pacientes, censurada à direita em 48 meses. Assim, como na figura 17, os módulos na figura 19 foram ordenados de acordo com a correlação dos *eigengenes* com o fenótipo tumoral/não tumoral das amostras.

Figura 19: Associação entre a expressão de módulos da rede de co-expressão, alterações somáticas e sobrevida de casos de PDAC. O *heatmap* mostra o p-valor da correlação entre os *eigengenes* de cada um dos módulos da rede de co-expressão (mesmas cores da figura 17 indicadas na primeira linha), representados nas colunas, e entre diferentes variáveis descritas nas linhas. Estão apresentados nas linhas de cima

para baixo os valores de correlação dos módulos (colunas) com i) presença/ausência de mutação nos 11 genes selecionados; ii) frequência alélica encontrada no sequenciamento de alta profundidade dos códons 12 e 13 de KRAS (linha 'razão KRAS alternativo'; iii) número total de mutações (SNVs e *indels*) detectadas em cada paciente; iv) representação relativa das assinaturas mutacionais S1 e S2 (linha 'razão signeR'); v) sobrevida dos pacientes. vi) A linha inferior apresenta a correlação com o fenótipo tumor/tecido não tumoral (linha 'Fenótipo'). A escala de cores diferencia genes significativamente correlacionados positivamente (verde) ou negativamente (vermelho).

Entre os 11 genes significativamente associados com módulos de expressão, na maior parte dos casos a presença de mutação nos genes KRAS, ARHGEF4, SIPA1L1, ZFHX4, CSMD3, OBSCN, TIAM1, TP53 e TLN2 está associada ao aumento da expressão gênica em 7 módulos enquanto mutações em 3 genes (CELSR1, TP53 e TBC1D1) estão associadas à diminuição da expressão de 6 módulos da rede.

As correlações medidas entre a expressão dos módulos e a frequência alélica de mutações em KRAS estimada a partir do sequenciamento com alta cobertura de *amplicons* foram em geral mais significativas do que as calculadas usando o vetor binário de presença/ausência de mutações em KRAS detectadas pela análise do exoma (Figura 19, KRAS *amplicon* vs. KRAS). Este resultado pode ser atribuído à maior sensibilidade resultante da maior cobertura de sequenciamento nesta técnica. Todos os módulos com expressão gênica significativamente correlacionada com mutações em KRAS detectadas no exoma são confirmados pelos dados de sequenciamento de *amplicons*, o que indica a ausência de resultados falsos-positivos.

Não foi observada uma associação significativa ($p < 0,01$) entre o número de mutações nas amostras e a expressão gênica nos módulos. Por outro lado, para 6 dos 21 módulos da rede foram observadas correlações significativas entre a expressão gênica do módulo e a intensidade relativa das assinaturas S1 ou S2 das amostras (Figura 19, razão signeR). Este resultado pode indicar que a expressão gênica de algumas vias moleculares possa ser afetada de forma distinta por diferentes processos mutagênicos que atuam no PDAC.

A comparação da expressão dos *eigengenes* dos módulos com a sobrevida dos pacientes indicou correlações significativas ($p < 0,01$) entre dois módulos, onde a expressão aumentada de genes envolvidos com '*nonsense mediated decay*' e '*rRNA processing*' (módulo azul celeste) ou a expressão diminuída de genes envolvidos com '*ubiquitin mediated proteolysis*' (módulo marrom) estão associados à maior sobrevida dos pacientes avaliados em nosso estudo.

Inferência funcional de lncRNAs a partir de rede de co-expressão gênica

Para cada um dos módulos com enriquecimento de vias moleculares de interesse relacionadas a malignidade, foram selecionados os lncRNAs diferencialmente expressos (p-valor ajustado de 0,01 e *fold change* igual ou maior que 2 vezes) e os mRNAs codificadores de proteínas com função conhecida na via selecionada

(diferencialmente expressos ou não). Para cada uma das vias consideradas de interesse, foi calculada a mediana da conectividade local (normalizada levando em conta apenas os membros da via) dos genes conhecidos, sendo este o valor esperado dos membros da via. Para algumas das vias foram observados lncRNAs com uma conectividade local acima da esperada para os membros da via calculada acima, que baseado em um princípio de “guilty by association” pode significar a existência de uma relação funcional entre os lncRNAs e um ou mais genes da via. A figura 20 abaixo apresenta resultados para algumas vias selecionadas.

O primeiro dos *heatmaps* na Figura 20 abaixo (painel A) foi construído usando uma das vias enriquecidas pelo módulo vermelho com genes envolvidos na via ‘*Cell Adhesion*’ ($p = 2,56 \cdot 10^{-45}$), neste módulo observa-se que um grande número de lncRNAs possuem conectividade maior que a esperada. Outra via com lncRNAs de alta conectividade que chamou a atenção por ter relevância no contexto da malignidade do câncer de pâncreas foi ‘*O-linked glycosylation of mucins*’ ($p = 1,03 \cdot 10^{-8}$) do módulo violeta (painel B, Figura 20). Um dos lncRNAs de alta conectividade local nesta via é o CCAT1, já descrito como envolvido no desenvolvimento tumoral (Nissan et al., 2012), já para o lncRNA de maior conectividade local na via, LINC01559, não foi encontrado na literatura nenhum trabalho descrevendo sua função. Uma outra via contendo lncRNAs diferencialmente expressos de alta conectividade é a via ‘*DNA repair*’ ($p = 1,52 \cdot 10^{-5}$, módulo verde limão). Um dos lncRNAs de alta conectividade nesta via é o UCA1 (Painel C, Figura 20), que já foi descrito tendo sua expressão associada à resistência de quimioterápicos, invasão de tecidos e desenvolvimento tumoral (Wang et al. 2017).

Figura 20: Pannel de *heatmaps* mostrando a interação na rede de co-expressão entre lncRNAs diferencialmente expressos (colunas) e genes (linhas) associados a vias

moleculares enriquecidas no módulo a qual estas vias pertencem. A, '*Cell Adhesion*'; B '*O-linked glycosylation of mucins*'; C, '*DNA Repair*'. Na região central dos *heatmaps* a escala de cinza indica o valor do módulo da correlação (variando entre 0 e 1) entre os lncRNAs e genes codificadores de proteína. As barras laterais, indicam em azul os que estão mais expressos em amostras controle e em vermelho os mais expressos em amostras tumorais, quando a diferença não é significativa a tendência é indicada em tons pastéis. A barra superior em escala verde-roxa relaciona a conectividade local do transcrito não codificador com a mediana da conectividade local dos transcritos codificadores nesta via, em branco os que têm a conectividade na via igual à mediana, em verde os com conectividade local superior, e em roxo os de conectividade inferior à mediana entre os membros da via.

Discussão

Análise de alterações somáticas simples no PDAC

Este estudo utilizou uma plataforma de captura que interroga 62Mb do exoma humano para analisar 55.766 genes em busca de alterações somáticas em uma coorte de 14 casos de PDAC. Além da região codificadora destes genes, nossa análise do exoma de PDAC interrogou também regiões UTR de 18.465 genes. Utilizando uma cobertura média de 40x e utilizando um pipeline bem estabelecido na literatura para chamada de variantes somáticas (GATK), foram identificadas 28.683 mutações simples, que incluem SNVs e *indels*, em 12.438 genes, sendo 96% deste são codificadores de proteínas.

Consórcios internacionais como o ICGC (<https://dcc.icgc.org/>) e TCGA (<https://portal.gdc.cancer.gov/>) já geraram dados de alterações somáticas em várias dezenas de casos, utilizando estratégias diferentes para captura exoma ou através do sequenciamento de genomas completos e para o processamento e análise dos dados para a chamada de variantes. Apesar dessas diferenças metodológicas, como um controle de qualidade do *pipeline* analítico utilizado, foi comparado o número de mutações identificadas em nossa análise com a frequência de mutações identificadas por estes grupos. Um número comparável de mutações simples foi observado em nosso estudo (2.049 mutações/amostra) e no ICGC/PACA-CA (2.285 mutações/amostra), que realizou o sequenciamento com alta-cobertura do genoma

completo de 268 casos de PDAC. Por outro lado, a mesma análise usando dados de 177 amostras do projeto PAAD-US revelou um número cerca de 10 vezes menor de mutações por amostra (267 mutações/amostra). Diferentes fatores podem explicar essas diferenças. Embora tenha sido utilizado o mesmo programa para a chamada de variantes (Mutect), no projeto PAAD-US foi usada uma versão anterior do programa (Mutect1). Outra explicação plausível é que apenas variantes identificadas por dois de três diferentes grupos de análise independentes foram reportadas. No artigo em que os dados do TCGA são apresentados não existe uma descrição das diferenças nos processos analíticos das equipes (Raphael et al., 2017), e é possível que divergências entre os métodos usados expliquem o menor número de variantes reportadas.

Não foi observada a existência de associação entre o número de mutações dos tumores e cobertura de sequenciamento, o que sugere a saturação no número de alterações somáticas detectadas nas amostras. Para investigar em maior detalhe a sensibilidade dos dados gerados para detectar mutações somáticas nas amostras de PDAC e aumentar a confiança da chamada de variantes realizada, foi feito o sequenciamento de alta profundidade de *amplicons* dos *códons* 12 e 13 de KRAS, que se encontram recorrentemente mutados na doença (Jones et al., 2008; Biankin et al., 2012; Waddell et al., 2015). Foi observada uma correlação significativa ($r=0,58$, $p=0,03$) entre as frequências alélicas dos *códons* KRAS selvagens e mutados detectadas no sequenciamento do exoma e *amplicons*. Embora o sequenciamento de *amplicons* tenha confirmado as mutações em KRAS em todos os casos positivos detectados na análise do exoma, houveram três casos em que não foram detectadas mutações no

sequenciamento do exoma (pacientes 06, 50 e 29) mas que contêm mutações nos *códons* 12 de KRAS detectadas no sequenciamento de *amplicons*. Todos os 3 casos falso-negativos apresentaram baixa frequência alélica (~2, ~5 e ~8%, respectivamente). Esse resultado indica que o nível de cobertura alcançado no sequenciamento do exoma ainda pode ser insuficiente para detectar mutações legítimas em amostras nos casos que apresentem menor celularidade tumoral ou alta heterogeneidade clonal na massa tumoral. Um caso quarto caso em que foi chamada mutação no códon 12 de KRAS no sequenciamento de *amplicon* (com frequência alélica de 50%) mas não no exoma (paciente 44) pode ser explicada pela alta frequência de *reads* variantes na amostra controle (> 20%), que excluiu essa mutação na chamada de variantes. Possíveis explicações para este resultado são a existência de mutação germinativa no gene ou uma contaminação da amostra de DNA genômico usada (isolada de sangue do paciente).

Nota-se que as amostras avaliadas distribuem-se de forma bimodal em relação ao número de mutações encontradas, o que está de acordo com a existência de amostras com maior e menor instabilidade genômica já reportadas na literatura (Waddell et al., 2015). Outra característica comumente reportada na literatura é a alta heterogeneidade genômica intertumoral do PDAC, caracterizada por poucos genes recorrentemente mutados e mutações infrequentes em um alto número de genes (Jones et al., 2008; Biankin et al., 2012). De fato, 98% das mutações detectadas em nosso estudo foram observadas em apenas um paciente. Embora a maior parte das variantes somáticas detectadas possa ser atribuída a mutações aleatórias resultantes

da instabilidade genômica característica da célula maligna, é possível priorizar um subconjunto de 3.314 mutações (12% do total) que já foram observadas anteriormente em diferentes tipos de tumores e que fazem parte da lista curada de alterações somáticas do banco de dados COSMIC (Forbes et al., 2016). Outro banco de dados usado para catalogar as mutações encontradas com maior probabilidade de serem relevantes foi o CCGD, que lista genes em que foi observado experimentalmente como a inserção de transposons aumenta ou diminui a viabilidade de tumores em modelos animais (Abbot et al., 2015). Do total de genes mutados neste trabalho, 8% (976 genes) estão listados no CCGD com algum grau de impacto oncogênico ou supressor tumoral em um modelo murino de PDAC.

Seleção de candidatos a gene “drivers” no PDAC

Assim como em outros tumores, é provável que a maior parte das mutações presentes em tumores e que se acumulam durante o desenvolvimento tumoral são “passageiras” e portanto não tem uma função central em sustentar os fenótipos malignos. Por outro lado, quanto mais vezes uma mesma mutação é observada em diferentes tumores, menor a probabilidade desta observação ser casual. Caso uma mutação tenha propensão a iniciar um processo de tumorigênese, é provável que esteja presente na maior parte da população de células transformadas e resulte em uma maior frequência alélica no tumor (McGranahan and Swanton 2015), e portanto seja detectada mais facilmente. Seguindo a mesma lógica, a observação do acúmulo de mutações em um

dado gene, ainda que em diferentes coordenadas, em um grande número de tumores, aumenta a probabilidade de que as alterações somáticas neste gene tenham um papel relevante no desenvolvimento tumoral. Para priorizar genes com mutações frequentes e/ou genes recorrentemente mutados em diferentes amostras, foi utilizada uma abordagem computacional implementada no programa OncodriveFML que busca genes com uma distribuição de mutações maior que ao acaso, levando em conta o número de mutações, o tamanho do gene, quão agrupadas as mutações estão neste gene (“hotspot”), entre outros fatores. (Mularoni et al., 2016).

Os estudos já realizados de alterações somáticas no PDAC se concentraram na análise das regiões codificadoras do genoma (Biankin et al., 2012; Waddell et al., 2015; Raphael et al., 2018). Utilizando o programa OncodriveFML que utiliza uma abordagem estatística para detectar sinais de seleção positiva em regiões codificadoras e não codificadoras do genoma, 1.603 foram considerados significativamente alterados neste trabalho, sendo 1.600 codificadores e 3 lncRNAs. Após a seleção de genes pelo OncodriveFML a porcentagem de genes que possui sobreposição com oncogenes/supressores tumorais anotados pelo CCGD aumentou de 8% para 17% (277 genes), indicando que este conjunto está enriquecido em genes funcionalmente relevantes para a malignidade no PDAC.

Cerca de metade das mutações simples observadas (2.745 mutações *missense*, 52% do total) resultaram na substituição de aminoácidos na proteína, 12% (609 mutações) resultaram em alterações drásticas na sequência da proteína codificada (perda do códon de iniciação, códon prematuro/disrupção do códon de

parada, mudança de fase de leitura, ganho/perda de sítios de splicing) e 36% (1.883 mutações) ocorreram em regiões não codificadoras do gene. Para priorizar uma lista reduzida de candidatos, foram considerados critérios de seleção adicional apenas os genes que possuísssem ao menos uma mutação no COSMIC, e que apresentassem mutação em pelo menos 4 pacientes, resultando em uma lista de 18 genes (Figura 10). Esta lista inclui KRAS, o oncogene recorrentemente mutado no PDAC (Hruban et al., 1993, Waddell, 2015) e cuja mutação ativadora já foi demonstrada ser suficiente para a transformação maligna no câncer de pâncreas em modelo murino (Hingorani et al., 2003). Os demais genes incluem genes reportados como mutados e/ou funcionalmente relevantes em outros tipos de tumores, mas poucos já foram descritos no PDAC. Interessantemente, três genes, UBR5, UBR4 e HECTD4, fazem parte do complexo Ubiquitina Ligase E3. UBR5 é extremamente conservado e essencial para o desenvolvimento de mamíferos e é um importante regulador de sinalização celular. Embora frequentemente mutado em tumores, seu papel na malignidade ainda é pouco compreendido (Shearer et al., 2015). UBR4 e HECTD4 são menos estudados; o silenciamento de UBR4 (e UBR5) foi associado à viabilidade tumoral, vascularização e proliferação em células de carcinoma hepatocelular (Leboeuf et al., 2017); HECTD4 tem a expressão regulada em câncer de vias biliares por miR-98-5p e HEIH, oncogene ligado à invasão e proliferação celular (Wan et al., 2020).

Entre os genes mutados encontra-se fatores de transcrição, modificadores de histonas e remodeladores de cromatina. KMT2A, uma metil-transferase específica da histona H3K4, que interage está ligado à diversos tumores e interage com o supressor

tumoral MEN1 (Lin et al. 2016); ARID1B e ARID2, ambos componentes do complexo SWI/SNF, foram observados com alterações somáticas em PDAC e envolvidos em reparo de DNA e regulação da transcrição através de remodelagem da cromatina (Ying, et al., 2016), sendo ARID1B diferencialmente expresso em nossos dados de transcriptoma (p-ajustado = 4%). INTS8, membro do complexo Integrator, parece regular a expressão de SMAD4 e ser regulado por TGF- β 1 em carcinoma hepatocelular (Tong et al., 2019). ZFHX3, da família *Zinc finger homeobox* e é frequentemente mutado em outros tumores, sendo estas mutações associadas à pior sobrevida dos pacientes (Walker et. al, 2015), este gene está diferencialmente expresso em nossos dados de transcriptoma (p-ajustado = 3%). DIP2C, *Disco Interacting Protein 2C* é um potencial gene supressor de tumores, possui mutações frequentes em tumores de mama e expressão alterada em diversos tumores, foi observada em tumores de mama uma correlação de expressão negativa com EGFR e positiva com receptores relevantes para este tecido (Li et al., 2017).

Foram ainda identificados genes envolvidos em transdução de sinal intracelular, morte celular programada, organização do citoesqueleto e matriz extracelular; PTPRK, membro da família das tirosina-fosfatases de proteínas, observada com mutações em diversos tumores e indicada como supressor tumoral na via de apoptose (Sun et al. 2013). CDC42BPB, uma serina/treonina proteína quinase, Lochhead et al., 2010 aponta este gene como portador de possíveis mutações ativadoras relevantes para adenocarcinoma intestinal, Sin et al., 2017 aponta este gene como marcador diagnóstico de câncer de bexiga, e Nagaraj et al., 2011 indica este gene como possível

candidato de droga para prevenir crescimento de câncer colorretal. LRBA, uma proteína superexpressa em tumores de pâncreas, envolvida em vias de apoptose, transdução de sinal, organização de citoesqueleto, regulação transcricional, entre outras, mas sem um mecanismo conhecido (Alkhairy et al., 2016). BIRC6, é uma proteína inibidora de apoptose, e sua expressão está associada a pior diagnóstico em vários tumores (Lamers et al., 2012; Hu et al., 2015; Luk et al., 2017), e está significativamente alterada em nossos dados de transcriptoma (p -ajustado = 8×10^{-4}). SASH1, um provável supressor tumoral envolvido em apoptose, proliferação celular, e invasão em diversos tumores (Rimkus et al., 2006; Chen et al., 2010; Zong et al., 2016). SYNE2, da família das Nesprinas, proteínas que fazem a ligação entre o citoesqueleto e o nucleoesqueleto, que interagem com p51, um inibidor de quinase ciclina-dependente (CDK) e fator downstream de TP53 associado em vários tumores regulando migração celular (Han et al., 2016). FRAS1, uma proteína da matriz extra-celular, associada à invasão e migração de linhagens celulares geradas à partir de tumores de pulmão (Zhan et al., 2014), este gene está diferencialmente expresso em nossos dados de transcriptoma (p -ajustado = 2×10^{-4}). MON2, é uma proteína envolvida no transporte de vesículas à partir do complexo de golgi (Gillingham et al., 2006) e apesar de não haver trabalhos que relacionem este gene a câncer, este é observado em nossos dados diferencialmente expresso com p -ajustado de 2×10^{-3} .

Alguns oncogenes e supressores tumorais já descritos no PDAC não figuram nesta lista estrita, entre eles os genes SMAD4, BRCA1, ATM e BRAF, que assim como observado em outros trabalhos (Waddell et al., 2015; Raphael et al., 2018), são

mutados em menos de 20% das amostras. TP53 é outro gene bem estabelecido como supressor tumoral no PDAC (Cicenas et al., 2017) e que não foi incluído nesta lista por não ter sido anotado no banco de dados CCGD.

Três genes não codificadores de proteína foram considerados significativamente mutados nas amostras de PDAC pelo OncodriveFML, WASH3P, PMS2P3, e CMAHP, todos eles pseudogenes de genes codificadores de proteínas. Embora outros trabalhos citem mutações frequentes em WASH3P (WASP Family Homolog 3, pseudogene) e CMAHP (Cytidine Monophospho-N-Acetylneuraminic Acid Hydroxylase, Pseudogene) em tumores de cabeça e pescoço e leucemia mieloide aguda respectivamente (Meucci et al., 2016; Tuborgh et al., 2013), nenhum destes trabalhos investigou o impacto funcional destas alterações. PMS2P3 (PMS1 Homolog 2, Mismatch Repair System Component Pseudogene 3) é um pseudogene de um gene codificador de proteína envolvido em reparo de DNA e não foi reportado alterado em tumores (Li et al., 2015).

A heterogeneidade intertumoral é reconhecidamente um fator que explica o grande número de alterações somáticas anotadas em neoplasias em geral e a pequena sobreposição entre mutações detectadas em diferentes amostras (McGranahan and Swanton 2015). Porém é plausível que entre os genes recorrentemente mutados identificados em nosso trabalho incluam-se genes “driver” na oncogênese do PDAC que sejam mais prevalentes na população brasileira e que portanto não foram identificados em estudos que focaram em amostras de populações com composição étnica distinta.

Vias moleculares enriquecidas em genes mutados no PDAC

Além de identificar individualmente genes com mutações somáticas relevantes para a malignidade, buscamos identificar vias moleculares enriquecidas que se encontram enriquecidas entre os genes mutados, considerando que mutações em genes distintos de uma mesma via podem ter impacto semelhante para sustentar fenótipos do PDAC. Os 1.603 genes com mutações somáticas estão enriquecidos em centenas de termos de ontologias que podem ser agrupados em seis grupos de vias moleculares relacionadas a: i) desenvolvimento neuronal, câncer e citoesqueleto; ii) crescimento/motilidade celular e desenvolvimento de tecidos; iii) sinalização celular; iv) sistema imune e proteólise; v) apoptose e resposta a estímulos químicos; vi) transporte de macromoléculas.

Embora as alterações em KRAS sejam prevalentes na maior parte dos pacientes (Biankin et al., 2012), a existência de uma pequena porção de pacientes sem qualquer alteração detectada nesse gene (Raphael, 2018) refuta a hipótese de que alterações somáticas em KRAS seriam essenciais ao desenvolvimento tumoral. Uma hipótese plausível é que a oncogênese em pacientes com KRAS selvagem dependa de mutações em genes em vias que regulem ou sejam reguladas por KRAS. Para buscar evidências que suportem essa hipótese, selecionamos 360 vias moleculares que apresentam mutações em pelo menos um gene da via, em todas as amostras avaliadas. Várias das vias encontradas contém um grande número de genes (ex. '*Cellular Differentiation*' em que há 164 genes mutados) de forma que a probabilidade

de que todas as amostras possuam ao menos um gene mutado seja alta. Por outro lado, essa probabilidade deve ser menor em vias contendo menos genes, o que pode indicar vias críticas para a malignidade. De fato, entre vias moleculares com menor número de genes e com mutações em todas as amostras encontramos vias que contém KRAS, associadas a '*Positive regulation of MAP kinase activity*' com apenas 16 genes alterados, e '*G protein signaling pathways*' com apenas 10 genes. Outras vias com poucos genes que incluem KRAS são '*Neurotrophin signaling pathway*' (9 genes), '*Chronic myeloid leukemia*' (11 genes), '*FLT3 Signaling*' (13 genes), '*Focal Adhesion-PI3K-Akt-mTOR-signaling pathway*' (14 genes), '*Rap1 signaling pathway*' (19 genes), entre outras. Uma perspectiva futura deste trabalho é calcular a probabilidade de que uma via se encontre mutada para todos os pacientes e atribuir um p-valor para esta significância baseado no p-valor original do enriquecimento, no número de pacientes, e no número de mutações nestes genes.

Algumas destas vias já foram associadas à malignidade no PDAC como Neurotrofina (Renz et al., 2018), RAP1 (Lorenz et al., 2018), e PI3K-Akt-mTOR (Murthy et al., 2018). A via '*G protein signaling pathways*', uma das vias das quais KRAS é membro, é notável por ser possível encontrar pelo menos uma mutação por paciente mesmo só considerando os genes contidos no CCGD, ou seja todos os pacientes possuem mutações nesta via em genes para os quais já há evidência de potencial oncogênico. A observação de tumores KRAS selvagem que apresentam mutações em outros genes dessas vias deixa em aberto a possibilidade de que estas mutações possam fenocopiar o efeito oncogênico de mutações em KRAS (Eser et al., 2013).

Algumas vias onde todos os pacientes se encontram mutados em pelo menos um gene não incluem KRAS porém contêm outros genes importantes para a malignidade do PDAC, como TP53 e BRCA1 (Cicenas et al., 2017) (“SUMO E3 ligases SUMOylate target proteins”, 14 genes) e ATM (Armstrong et al., 2019) (“Cell Cycle”, 16 genes). É plausível que também nestas vias existam mutações em genes distintos que tenham efeito equivalente em suportar fenótipos malignos, interagindo ou não com KRAS.

Assinaturas mutacionais detectadas no PDAC

As alterações somáticas pontuais (SNVs) identificadas no exoma das amostras de PDAC foram analisadas na tentativa de identificar assinaturas mutacionais associadas a processos mutagênicos distintos. Usando o pacote *signeR* (Rosales et al., 2016) para avaliar padrões de substituição de bases predominantemente encontrados foi possível identificar “de novo” duas assinaturas distintas S1 e S2, a primeira é caracterizada predominantemente pela alteração frequente de C > T, em dinucleotídeos CpG, a segunda por alterações do tipo C > A.

Ao comparar as assinaturas encontradas com as assinaturas anotadas no COSMIC, foi observado que a assinatura S1 é mais semelhante a assinatura COSMIC 1, associada a mutações esporádicas devido a desaminação de citosinas metiladas que se acumulam em indivíduos idosos, e já reportada como uma das principais assinaturas encontradas em outros trabalhos que avaliaram assinaturas mutacionais

em PDAC (Aguirre et al., 2018; Connor et al., 2019). A assinatura S2 se mostrou mais semelhante à assinatura COSMIC 29, descrita como frequentemente associada a alterações somáticas relacionadas ao tabaco. Não foram encontrados trabalhos na literatura descrevendo a presença desta assinatura de exposição ao tabagismo com o PDAC. Não foram observadas correlações entre o perfil mutacional S1/S2 das amostras e covariáveis clínicas como idade ou o status de tabagismo (fumante/não fumante) das mesmas. Uma possível explicação para este resultado negativo é o pequeno número de amostras avaliado.

Análise de integração dos dados exoma e rede de co-expressão

Para buscar por mutações somáticas que possam influenciar a expressão gênica no PDAC, foram inicialmente feitas análises com dados de transcrito dos mesmos 14 casos para i) identificar vias moleculares enriquecidas entre genes diferencialmente expressos no tumor relativamente ao tecido não tumoral adjacente e ii) construir uma rede de co-expressão gênica e identificar módulos contendo genes diferencialmente expressos. A comparação de vias moleculares enriquecidas entre genes com alterações somáticas (696 vias) com vias enriquecidas entre genes diferencialmente expressos (163 vias) resultou em 83 vias moleculares em comum. Entre essas encontram-se as vias '*Ras protein signal transduction*', '*apoptotic process*', e '*cell adhesion*', já descritas na literatura como diferencialmente expressas no PDAC (Zhao et al. 2018). Um trabalho independente observou que estas mesmas vias estão

presentes em um conjunto de 12 vias de sinalização frequentemente mutados na maior parte dos 24 tumores avaliados (Jones et al., 2008). Semelhante ao observado em nosso estudo, os genes dessas vias que se encontram mutados variam entre os diferentes tumores analisados, o que sugere que mutações em diferentes genes da via possam resultar em um efeito comum na oncogênese. No trabalho de Jones e colaboradores foi notado que os genes frequentemente mutados nas 12 vias moleculares contém uma proporção maior de genes diferencialmente expressos que os demais (Jones et al., 2008). Nossa análise reforça a evidência de que alterações somáticas e transcricionais em genes em vias de sinalização específicas são importantes para a tumorigênese e manutenção do PDAC.

Foram também observadas vias moleculares associadas a sinalização ou processos celulares gerais que se encontram simultaneamente enriquecidas em genes com alterações somáticas e transcricionais, como 'regulation of intracellular signal transduction', 'cell-matrix adhesion', 'immune effector process', entre outras, que sugerem a existência de efeitos transcricionais nestas vias resultantes de mutações somáticas.

Uma segunda abordagem de integração de dados visou associar alterações somáticas com módulos da rede de co-expressão gênica de tumores de pâncreas. As vias moleculares simultaneamente enriquecidas em alterações somáticas e transcricionais contém 425 genes mutados. Para cada um destes genes foi calculada a correlação de Spearman entre a presença/ausência de mutações somáticas nos 14 casos avaliados e a expressão dos transcritos representados nos 22 módulos da rede

de co-expressão gênica, gerada com dados do transcrito das amostras de tumor e tecido não tumoral dos mesmos casos. Pode-se esperar que as alterações somáticas na região codificadora ou regulatória de um gene influenciam a sua expressão, mas não o contrário. Embora a ausência de expressão de genes envolvidos no reparo de DNA possa influenciar o número total de mutações em uma dada amostra é esperado que estas mutações ocorram aleatoriamente no genoma, não havendo uma probabilidade maior de ocorrência de mutações nos loci que codificam para estes genes. Portanto, pode-se inferir a partir destas correlações que as mutações estejam alterando a expressão gênica no módulo da rede, mas a recíproca possivelmente não é verdadeira.

Para 11 genes (KRAS, ARHGEF4, SIPA1L1, ZFHX4, CSMD3, CELSR1, OBSCN, TIAM1, TP53, TLN2 e TBC1D1) foram observadas correlações com significância estatística ($p\text{-valor} < 0,0001$) entre a existência ou não de mutações e a expressão dos módulos da rede (figura 19). Entre os genes frequentemente mutados e significativamente correlacionados com módulos específicos da rede de co-expressão destacam-se KRAS e TP53, genes frequentemente mutados em PDAC (Raphael et al., 2017) e que já tiveram um papel *driver* demonstrado na oncogênese. A presença de KRAS oncogênico se mostrou significativamente correlacionada com a expressão média de genes (*eigengenes*) em módulos que contêm vias moleculares ou processos biológicos importantes para a malignidade do PDAC como '*Hedgehog signalling*', '*WNT signalling*' e '*DNA repair*' (Kleeff et al., 2016). Em um estudo comparando a expressão gênica em linhagens celulares murinas isogênicas com e sem a mutação G12D em

KRAS foram identificados 667 genes diferencialmente expressos que se encontram enriquecidos em diversas vias moleculares identificadas em nosso estudo, como '*Cell Cycle*', '*DNA Repair*', '*WNT signalling*', '*Cell Growth/Proliferation*' (Sivakumar et al., 2017). A associação entre mutações em KRAS e alterações na expressão da via de *Hedgehog* detectada em nosso estudo não foi observada no trabalho de Sivakumar et al. Contudo, a partir de uma análise computacional subsequente que integrou a assinatura transcricional de KRAS oncogênico com uma rede de co-expressão gerada a partir de dados de centenas de casos de PDAC, foram inferidos fatores de transcrição responsáveis pela regulação destes genes e que apontou para o envolvimento da via de *Hedgehog* na reprogramação transcricional induzida por KRAS oncogênico (Sivakumar et al., 2017). É importante notar que algumas das vias correlacionadas com a presença de mutação em KRAS só foram detectadas utilizando dados da frequência alélica KRAS mutado/selvagem nas amostras estimada por sequenciamento com alta-profundidade dos códons 12 e 13 do gene, o que sugere que nem todas as associações relevantes entre genes mutados e módulos de expressão podem ter sido capturadas em nossa análise do exoma. Uma perspectiva futura de nosso trabalho é uma análise mais detalhada de quais vias e/ou genes são mais afetados dentro destes módulos da rede de co-expressão com alta correlação com a presença/ausência de mutação no KRAS e nos outros genes selecionados.

Alterações na via do fator de crescimento epidérmico (EGF), incluindo superexpressão da família ErbB de tirosina quinases receptoras estão presentes na grande maioria dos tumores sólidos incluindo PDAC (Sussman et al., 2010). Neste

trabalho observamos que presença de mutações em TP53 em amostras e PDAC apresentou alta correlação ($p = 7 \times 10^{-6}$) com um módulo da rede de co-expressão contendo genes da via de sinalização de ErbB. Já foi observado no câncer de mama que a presença de mutações em TP53 está associada com alterações na expressão de diversas vias de sinalização incluindo ErbB (Joshi et al., 2012), o que suporta a observação feita em nosso estudo no contexto do PDAC.

Outros genes frequentemente mutados nas amostras de PDAC e significativamente correlacionados com módulos específicos da rede de co-expressão foram menos caracterizados como “drivers” no contexto do câncer. Abaixo segue uma breve discussão do que já foi reportado para estes genes na literatura.

Mutações em ARHGEF4 (*Rho guanine nucleotide exchange factor 4*) já foram observadas em tumores de mama (Sjöblom et al., 2006). Já foi demonstrado que ARHGEF4 é diferencialmente expresso e promove invasão celular em PDAC (Taniuchi et al., 2018), interagindo na via MAPK/ERK *signaling pathway*. Corroborando este resultado, em nosso trabalho a mutação neste gene se mostrou correlacionada (p -valor = 0,03) com a diferença de expressão no módulo vermelho, no qual o gene MAPK está contido, embora a via de MAPK não tenha se mostrado enriquecida no módulo.

Uma análise de “hotspots” de mutação em domínios proteicos com dados de 23 diferentes tipos de tumores humanos (não incluiu PDAC) anotou alterações somáticas frequentes no gene ZFH4 (*Zinc finger homeobox protein 4*) (Miller et al., 2015). Em nosso trabalho a presença de mutações em ZFH4 se mostrou correlacionada ($p =$

3×10^{-5}) com o módulo que contém a via '*Tissue development*'. Ainda não existem relatos de mutações ou função de ZFHX4 no câncer de pâncreas.

Neste trabalho observamos uma correlação significativa ($p = 7 \times 10^{-6}$) entre a presença de mutações em TIAM1 e o módulo contendo genes da via '*protein processing in the endoplasmic reticulum*'. Foi mostrado que mutações no gene TIAM1 (*T lymphoma invasion and metastasis protein 1*) ativam a via de Rac1 (He et al., 2018), e que a expressão conjunta de TIAM1 e Rac1 em PDAC leva à proliferação celular e invasão de tecidos (Guo et al., 2013).

Há um trabalho mostrando que o silenciamento do gene TLN2 (talin 2) reduz a invasão e a migração de células de carcinoma hepatocelular além de reduzir o tempo na fase G0/G1 do ciclo celular (Fang et al., 2016), esse mesmo trabalho ainda levanta a hipótese de que TLN2 possa estimular a metástase. Em nosso estudo a mutação de TLN2 se mostrou significativamente correlacionada com módulos enriquecidos em genes envolvidos em '*Regulation of IFNA signalling*' ($p = 9 \times 10^{-4}$) e '*protein processing in the endoplasmic reticulum*' ($p = 3 \times 10^{-5}$). Não foram reportadas mutações de TLN2 em tumores e seu papel na oncogênese.

Os genes CSMD3 e OBSCN já foram observados frequentemente mutados em tumores de mama e colorretal (Sjöblom et al., 2006), sendo CSMD3 diferencialmente expresso em nosso transcrito do PDAC (p -ajustado = 8×10^{-7}). Em nosso estudo amutações nestes genes se mostraram significativamente correlacionadas com o módulo enriquecido em genes envolvidos em '*ubiquitin mediated proteolysis*' (CSMD3 , $p = 1 \times 10^{-4}$; OBSCN, $p = 1 \times 10^{-6}$), entre outras vias.

Em um pequeno artigo foi mostrado que CELSR1 estimula a migração e a angiogênese em culturas de células endoteliais de aorta e de cordão umbilical (Zhan et al., 2016), este gene também foi observado com mutações em amostras de câncer de pulmão em (Miller et al., 2015) em nosso transcrito este gene possui um transcrito diferencialmente expresso (p -ajustado = 8×10^{-11}) representado na rede de co-expressão no módulo vermelho e faz parte da via '*Cell adhesion*'.

TBC1D1 foi observado como um gene essencial para a função e sobrevivência de células beta pancreáticas (Rütti et al., 2017), mas nenhuma implicação em câncer foi observada, em nossas análises as alterações somáticas em TBC1D1 estão associadas à diferenças de expressão em módulos contendo vias de metabolismo e expressão gênica em células beta, é possível que as mutações neste gene estejam associadas à perda de função pancreática. Sobre SIPA1L1 apenas foi encontrado na literatura a observação de que codifica para uma proteína alvo de papilomavírus (Lee et al., 2007).

Além de calcular a correlação entre expressão gênica nos módulos da rede e a existência ou não de mutações somáticas detectadas nas amostras de PDAC, também foram calculadas correlações com outras características moleculares das amostras ou variáveis clínicas dos pacientes.

Para 6 módulos diferentes da rede de co-expressão foram observadas correlações significativas ($p < 0,01$) entre a expressão gênica e a intensidade relativa das assinaturas mutacionais S1 ou S2 das amostras (Figura 19, razão signeR). Este

resultado pode indicar que a expressão gênica de algumas vias moleculares possa ser afetada de forma distinta por diferentes processos mutagênicos que atuam no PDAC.

A comparação da expressão dos *eigengenes* dos módulos com a sobrevida dos pacientes indicou dois módulos relevantes, o módulo azul celeste (enriquecido por genes envolvidos com '*nonsense mediated decay*' e '*rRNA processing*') em que a sobrevida é correlacionada com maior expressão no módulo e o módulo marrom (enriquecido por genes envolvidos com '*ubiquitin mediated proteolysis*') em que a expressão diminuída dos genes está associada à maior sobrevida dos pacientes avaliados em nosso estudo.

Não foi possível encontrar correlação entre a expressão de qualquer um dos módulos com a instabilidade tumoral medida pelo número de mutações observadas em cada amostra.

Análise do transcritoma utilizando redes de co-expressão

A maior parte dos RNAs não codificadores longos anotados anda não foi estudada funcionalmente e a investigação de suas funções biológicas é um campo intenso de estudo. Neste trabalho foi utilizada uma abordagem computacional que detecta lncRNAs, ainda sem função conhecida, com expressão aberrante no PDAC e com uma correlação muito maior que a esperada entre a sua expressão gênica e a de genes conhecidos, inferindo portanto que estes lncRNAs possam ter alguma função relacionada à destes genes. (Cogill e Wang, 2014)

Inicialmente foi gerada uma rede de co-expressão com medidas globais de

expressão gênica de uma reconstrução do transcrito do pâncreas exócrino previamente gerada no laboratório utilizando dados de RNAseq de 14 pares de amostras de tecido pancreático tumoral e não tumoral (Tese de Doutorado de Julio Sosa, 2019). A rede de co-expressão foi construída com 23.900 transcritos (10.380 genes), sendo que 10.133 (3.383 genes) transcritos foram excluídos por possuírem baixa conectividade total na rede, resultando em 22 módulos de genes agrupados em função de seu padrão de conectividade. Os módulos gerados apresentaram um número variável de transcritos, entre 42 até 3015 transcritos (655,6 em média), com uma porcentagem de lncRNAs também variável (13% até 57% do total de transcritos do módulo, 36% em média).

Em seguida, para cada um dos 22 módulos presentes na rede de co-expressão gênica gerada pelo WGCNA foram identificadas as vias moleculares e processos biológicos significativamente enriquecidos entre os genes do módulo utilizando anotações funcionais padronizadas (Gene Ontology, KEGG, Reactome, Wiki Pathways). As vias moleculares significativamente enriquecidas foram examinadas e priorizadas aquelas que contivessem lncRNAs diferencialmente expressos nas amostras de PDAC (p -valor $< 0,01$) e que apresentassem conectividade local alta (isto é, acima da conectividade média dos demais transcritos do módulo). Em alguns módulos foram encontradas vias moleculares significativamente enriquecidas e com transcritos apresentando alta conectividade com lncRNAs com expressão aberrante em tumores, como '*Cell Adhesion*' (16 lncRNAs), '*O-linked glycosylation of mucins*' 8 lncRNAs), e '*DNA Repair*' (9 lncRNAs). Para estes lncRNAs a alta conectividade local

com genes associados a uma via molecular específica permite fazer a hipótese de que estes transcritos estejam regulando, direta ou indiretamente, os genes envolvidos nesta via.

Para verificar a existência de causalidade das relações funcionais entre alguns lncRNAs candidatos e genes de vias moleculares específicas inferidas através da rede de co-expressão foram realizados experimentos de validação experimental, conduzidos em colaboração com membros do laboratório. Foi usada a estratégia de silenciamento de alguns lncRNAs candidatos (UCA1, CCAT1 e LINC01559) em culturas de células imortalizadas de PDAC (linhagem AsPC-1) seguido da observação do efeito sobre a expressão de genes coexpressos e fenótipos celulares associados às vias moleculares inferidas. Como controle da especificidade da associação entre o lncRNA candidato e o gene alvo, foi avaliado o efeito sobre a expressão do silenciamento transiente do lncRNA sobre genes não relacionados, selecionados de outros módulos da rede. Foram escolhidos como controle negativo genes com alta conectividade em uma via molecular enriquecida em um módulo distinto, para garantir que a alta conectividade em um módulo não tem influência alguma em membros importantes de outro módulo. A expressão gênica de lncRNAs e genes coexpressos em células tratadas com siRNAs e células controle foi avaliada por RT-qPCR.

Ao silenciar o lncRNA UCA1, localizado no módulo verde-limão e que apresentou alta conectividade local com genes da via de '*DNA repair*', foi observada a diminuição significativa (13 e 14 vezes, respectivamente) na expressão dos genes PPP4C e SPIDR ambos associados à via '*DNA repair*' e coexpressos com UCA1. Por

outro lado não foi observada alteração de expressão de MUC13, um gene *hub* do módulo violeta. (Paixão, 2020). Os genes PPP4C e SPIDR são ambos parte de uma sub-categoria de reparo de DNA, a de reparo de quebra de dupla fita de DNA. Para validar a ação do lncRNA UCA1 no reparo de dupla fita foi feito um experimento em que culturas de AsPC-1 foram expostas à radiação gama para que o DNA fosse danificado, e posteriormente foi avaliada a capacidade de reparo ao longo do tempo, pela medida da integridade do DNA utilizando um ensaio cometa alcalino (Singh et al. 1988). Nas células em que UCA1 foi silenciado observou-se uma menor capacidade de reparo das lesões no DNA causadas pela radiação em comparação a células controle irradiadas com a mesma intensidade de radiação gama e submetidas ao mesmo tempo de recuperação (Figura 21, retirado de Paixão, 2020). Um trabalho da literatura já mostrou que UCA1 está superexpresso em tumores de próstata resistentes à radioterapia, e que o silenciamento de UCA1 induz a um pequeno aumento na mortalidade de células expostas à radiação (Ghiam et al., 2017), o que corrobora o envolvimento do lncRNA UCA1 com a via de reparo de DNA apontado em nosso estudo.

Figura 21: Efeito do silenciamento do lncRNA sobre o reparo de DNA em células de PDAC. Linhagem controle de AsPC-1, em verde, e a linhagem com silenciamento de UCA1, em roxo, receberam dosagem 10 Gy de radiação gama e após 15 minutos (menor intervalo experimentalmente possível) foi avaliado o efeito de dano no DNA pelo aumento do arraste de migração no ensaio cometa (medido em unidades arbitrárias de OTM “Olive Tail Movement”). Devido aos mecanismos de reparo de DNA presentes na célula, com o passar do tempo observa-se menos DNA fragmentado em ambas as culturas celulares, mas na linhagem em que UCA1 foi silenciado observa-se uma recuperação mais lenta. A diferença mais notável em 30 minutos e em uma hora, sendo o p-valor destas diferenças notado por: * $P \leq 0,05$; **** p-valor $\leq 0,001$. (Paixão, 2020).

No módulo violeta, que contém KRAS, dois lncRNAs com alta conectividade local na via ‘*O-linked glycosylation of mucins*’ foram selecionados para validação, o LINC01559, lncRNA sem função descrita na literatura, e o CCAT1, um lncRNA conhecido envolvido em câncer (Nissan et al., 2012) inclusive PDAC (Yu et al.,

2016) mas cujos mecanismos de ação ainda não foram caracterizados em detalhe. Ao silenciar individualmente estes dois lncRNAs observou-se uma redução da expressão de GALNT3 e B3GNT3, ambos membros da via '*O-linked glycosylation of mucins*', sendo mais notável o efeito após o silenciamento de LINC01559 do que de CCAT1. Também foi notado que enquanto o silenciamento de CCAT1 diminui a expressão de LINC01559, o silenciamento de LINC01559 não altera a expressão de CCAT1, sugerindo que LINC01559 esteja em uma posição intermediária entre CCAT1 e os outros genes da via. Para verificar se os efeitos na expressão gênica observados após o silenciamento de LINC01559 ou CCAT1 refletem efeitos regulatórios específicos no módulo violeta que contém genes da via de glicosilação de mucinas, foi também medida a expressão de PPP4C, um gene *hub* do módulo verde-limão, e observou-se que sua expressão não foi alterada pelo silenciamento de LINC01559 ou CCAT1. (Paixão, 2020)

Conclusões

- A análise de alterações somáticas simples de 14 amostras clínicas de PDAC através do sequenciamento com alta cobertura (40x em média) do exoma identificou 28.683 mutações simples (SNVs e *indels*) em 12.438 genes. Apenas 12% (3.314) das alterações somáticas detectadas já foram anotadas em tumores humanos (COSMIC). Do total de genes mutados, 8% (976 genes) estão anotados

como exercendo algum grau de impacto oncogênico ou supressor tumoral no PDAC (CCGD).

- Entre 1.603 genes “drivers” candidatos, 18 apresentam-se mutados em ao menos 25% das amostras e também se encontram nas listas COSMIC e CCGD. Entre estes se encontra o oncogene KRAS e genes reportados como mutados e/ou funcionalmente relevantes em outros tipos de tumores mas ainda não descritos no PDAC, incluindo componentes do complexo Ubiquitina Ligase E3 (UBR5, UBR4, HECTD4), fatores de transcrição (INTS8, ZFHX3, DIP2C), modificador de histonas (KMT2A), remodeladores de cromatina (ARID1B, ARID2), transdução de sinal e apoptose (PTPRK, CDC42BPB, LRBA, BIRC6, SASH1, SYNE2, FRAS1, MON2).
- Uma análise de enriquecimento de vias revelou 696 vias moleculares enriquecidas entre os 1.603 genes com mutações somáticas recorrentes no PDAC. Entre essas destacam-se vias com poucos componentes e que contém o oncogene KRAS (*‘Positive regulation of MAP kinase activity’*, *‘G protein signaling pathways’*, *‘Neurotrophin signaling pathway’*, *‘FLT3 Signaling’*, *‘Focal Adhesion-PI3K-Akt-mTOR-signaling pathway’*, *‘Rap1 signaling pathway’*). Em todas essas vias observou-se que todos os casos avaliados possuem ao menos um gene da via mutado e a existência de casos sem mutação em KRAS, sugerindo a possibilidade de que mutações em outros componentes destas vias possam fenocopiar o efeito oncogênico de mutações em KRAS.
- A sobreposição de vias moleculares enriquecidas em mutações somáticas e alterações transcricionais identificou 83 vias moleculares em comum, incluindo *‘Ras*

protein signal transduction, *apoptotic process*, e *cell adhesion*, indicando que alterações somáticas e transcricionais em genes em vias de sinalização específicas são importantes para a tumorigênese e manutenção do PDAC.

- Para 11 genes (KRAS, ARHGEF4, SIPA1L1, ZFHX4, CSMD3, CELSR1, OBSCN, TIAM1, TP53, TLN2 e TBC1D1) foi detectada uma associação estatisticamente significativa entre a presença de mutação no gene e a abundância relativa de transcritos localizados em módulos específicos da rede de coexpressão gênica gerada com dados de RNAseq de amostras pareadas de tumor e tecido não tumoral dos mesmos 14 casos. Entre estes destacam-se a associação de mutações em KRAS com a expressão de módulos enriquecidos nas vias *'Hedgehog signalling'*, *'WNT signalling'*, *'DNA repair'* e *'Cell Cycle'*, e mutações em TP53 e a expressão alterada em módulos contendo genes das vias de sinalização de ErbB.
- Foram também identificadas associações estatisticamente significativas entre a expressão gênica de módulos da rede e características moleculares das amostras ou variáveis clínicas dos pacientes, como intensidade relativa de assinaturas mutacionais (S1/S2) das amostras e sobrevida dos pacientes.
- Foi possível observar na rede de co-expressão lncRNAs diferencialmente expressos no PDAC e com alta conectividade a genes associados a vias moleculares específicas como *'Cell Adhesion'* (16 lncRNAs), *'O-linked glycosylation of mucins'* (8 lncRNAs), e *'DNA Repair'* (9 lncRNAs). Para alguns candidatos foram feitas validações experimentais que suportam um papel funcional exercido por estes lncRNAs nestas vias.

- Este trabalho faz parte de um esforço contínuo e conjunto dos membros do laboratório em que está inserido, e assim como auxiliou na escolha dos lncRNAs LINC01614, LINC01559, e UCA1 para estudo molecular detalhado (Paixão, 2020), é esperado que sustente outros trabalhos computacionais ou experimentais envolvendo outros genes e vias moleculares citados.

Referências

Abbott,K., Nyre,E., Abrahante,J., Ho,Y., Vogel,R., Starr,T. (2015) The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.*, D1, D844-D848.

Aguirre, Andrew J., et al. "Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine." *Cancer discovery* 8.9 (2018): 1096-1111.

Alexandrov, Ludmil B., et al. "The repertoire of mutational signatures in human cancer." *Nature* 578.7793 (2020): 94-101.

Alkhairy, Omar K., et al. "Spectrum of phenotypes associated with mutations in LRBA." *Journal of clinical immunology* 36.1 (2016): 33-45.

Anders S and Huber W (2010). "Differential expression analysis for sequence count data." *Genome Biology*, 11, pp. R106.

Appel, Betsy L., et al. "Current staging systems for pancreatic cancer." *The Cancer Journal* 18.6 (2012): 539-549.

Armstrong, Samantha A., et al. "ATM Dysfunction in Pancreatic Adenocarcinoma and Associated Therapeutic Implications." *Molecular cancer therapeutics* 18.11 (2019): 1899-1908.

Bailey, Peter et al. "Genomic analyses identify molecular subtypes of pancreatic cancer." *Nature* (2016).

Bao, Riyue et al. "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing." *Cancer*

Informatics 13.Suppl 2 (2014): 67–82. PMC. Web. 20 Jan. 2016.

Baraniskin, A., S. Nopel-Dunnebacke, M. Ahrens, S. G. Jensen, H. Zollner, et al. (2012). "Circulating U2 small nuclear RNA fragments as a novel diagnostic biomarker for pancreatic and colorectal adenocarcinoma." *International journal of cancer. Journal international du cancer*.

Benjamin, David, et al. "Calling Somatic SNVs and Indels with Mutect2." *BioRxiv* (2019): 861054.

Biankin, A. V., N. Waddell, K. S. Kassahn, M. C. Gingras, L. B. Muthuswamy, et al. (2012). "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes." *Nature* 491(7424): 399-405.

Bosman, Carneiro, Hruban, Theise "WHO classification of tumours of the digestive system." 2010

Chen, En-guo, et al. "Effects of SASH1 on lung cancer cell proliferation, apoptosis, and invasion in vitro." *Tumor Biology* 33.5 (2012): 1393-1401.

Chen, Mengjie, Murat Gunel, and Hongyu Zhao. "Somatica: Identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data." *PLOS One* 8.11 (2013): e78143.

Cheng, Yating, et al. "The long non-coding RNA HOTTIP enhances pancreatic cancer cell proliferation, survival and migration." *Oncotarget* 6.13 (2015): 10840.

Chudnovsky, Yakov, et al. "ZFHX4 interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state." *Cell reports* 6.2 (2014): 313-324.

Cicenas, Jonas, et al. "KRAS, TP53, CDKN2A, SMAD4, BRCA1, and BRCA2 mutations in pancreatic cancer." *Cancers* 9.5 (2017): 42.

Cogill, Steven B., and Liangjiang Wang. "Co-expression network analysis of human lncRNAs and cancer genes." *Cancer informatics* 13 (2014): CIN-S14070.

Connor, Ashton A., et al. "Integration of genomic and transcriptional features in pancreatic cancer reveals increased cell cycle progression in metastases." *Cancer cell* 35.2 (2019): 267-282.

Collisson, Eric A et al. "Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy." *Nature medicine* 17.4 (2011): 500-503.

Craig, David W et al. "Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities." *Molecular cancer therapeutics* 12.1 (2013): 104-116.

Cros, Jerome, et al. "Tumor heterogeneity in pancreatic adenocarcinoma." *Pathobiology* 85.1-2 (2018): 64-71.

Eser, Stefan, et al. "Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer." *Cancer cell* 23.3 (2013): 406-420.

Esteller, M. (2011). "Non-coding RNAs in human disease." *Nature reviews. Genetics* 12(12): 861-874.

Fabregat, Antonio, et al. "The reactome pathway knowledgebase." *Nucleic acids research* 46.D1 (2018): D649-D655.

Fang, Kun-peng, et al. "Both Talin-1 and Talin-2 correlate with malignancy potential of the human hepatocellular carcinoma MHCC-97 L cell." *BMC cancer* 16.1 (2016): 45.

Friedman, Eileen. "The kinase Mirk/dyrk1B: a possible therapeutic target in pancreatic cancer." *Cancers* 2.3 (2010): 1492-1512.

Forbes, Simon A., et al. "COSMIC: somatic cancer genetics at high-resolution." *Nucleic acids research* 45.D1 (2016): D777-D783. <http://cancer.sanger.ac.uk>

Gene Ontology Consortium. "The gene ontology resource: 20 years and still GOing strong." *Nucleic acids research* 47.D1 (2019): D330-D338.

Ghiam, Alireza Fotouhi, et al. "Long non-coding RNA urothelial carcinoma associated 1 (UCA1) mediates radiation response in prostate cancer." *Oncotarget* 8.3 (2017): 4668.

Gillingham, Alison K., et al. "Mon2, a relative of large Arf exchange factors, recruits Dop1 to the Golgi apparatus." *Journal of Biological Chemistry* 281.4 (2006): 2273-2280.

Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, et al. (2011). "lincRNAs act in the circuitry controlling pluripotency and differentiation." *Nature* advance online publication.

Guo, Xingjun, et al. "Balanced Tiam1-rac1 and RhoA drives proliferation and invasion of pancreatic cancer cells." *Molecular Cancer Research* 11.3 (2013): 230-239.

Han, Chuangye, et al. "EGFR and SYNE2 are associated with p21 expression and SYNE2 variants predict post-operative clinical outcomes in HBV-related hepatocellular carcinoma." *Scientific reports* 6 (2016): 31237.

Hariharan, D., A. Saied and H. M. Kocher (2008). "Analysis of mortality rates for pancreatic cancer across the world." *HPB : the official journal of the International Hepato Pancreato Biliary Association* 10(1): 58-62.

Hidalgo, M. (2010). "Pancreatic cancer." *N Engl J Med* 362(17): 1605-1617.

Hingorani, Sunil R., et al. "Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse." *Cancer cell* 4.6 (2003): 437-450.

Hoskins, Jason W., et al. "Transcriptome analysis of pancreatic cancer reveals a tumor suppressor function for HNF1A." *Carcinogenesis* 35.12 (2014): 2670-2678.

Hruban, R. H., A. D. van Mansfeld, G. J. Offerhaus, D. H. van Weering, D. C. Allison, et al. (1993). "K-ras oncogene activation in adenocarcinoma of the human pancreas. A study of 82 carcinomas using a combination of mutant-enriched polymerase chain reaction analysis and allele-specific oligonucleotide hybridization." *The American journal of pathology* 143(2): 545-554

Hu, Tingting, et al. "Overexpression of BIRC6 is a predictor of prognosis for colorectal cancer." *PLoS One* 10.5 (2015).

Huang, Xiaoyi, et al. "LncRNAs in pancreatic cancer." *Oncotarget* 7.35 (2016): 57379.

Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." *Nature protocols* 4(1): 44-57.

Hung, T., Y. Wang, M. F. Lin, A. K. Koegel, Y. Kotake, et al. (2011). "Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters." *Nature genetics* 43(7): 621-629.

IARC - International Agency for Research on Cancer (2008). *GLOBOCAN 2008 -Estimated cancer Incidence, Mortality, Prevalence and Disability-adjusted life years (DALYs) Worldwide in 2008*.

ICGC - International Cancer Genome Consortium. "International network of cancer genome projects." *Nature* 464.7291 (2010): 993.

Illumina, "Bcl2fastq." bcl2fastq Conversion Software, 1 Sept. 2017, support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.

Jiang, Li, et al. "Co-expression network analysis of the lncRNAs and mRNAs associated with cervical cancer progression." Archives of medical science: AMS 15.3 (2019): 754.

Jones, Siân et al. "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses." science 321.5897 (2008): 1801-1806.

Joshi, H., et al. "Potential tumorigenic programs associated with TP53 mutation status reveal role of VEGF pathway." British journal of cancer 107.10 (2012): 1722-1728.

Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000)

Karczewski, Konrad J., et al. "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes." BioRxiv (2019): 531210.

Kaur, Sukhwinder, et al. "Mucins in pancreatic cancer and its microenvironment." Nature Reviews Gastroenterology and Hepatology 10.10 (2013): 607.

Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR. (2009) Mining Biological Pathways Using WikiPathways Web Services. PLoS ONE 4(7)

Kim, K., I. Jutooru, G. Chadalapaka, G. Johnson, J. Frank, et al. (2013). "HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer." Oncogene32(13): 1616-1625.

Kim, Daehwan, et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." *Genome biology* 14.4 (2013): R36.

Kitts, Adrienne, and Stephen Sherry. "The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation." *The NCBI Handbook*. McEntyre J, Ostell J, eds. Bethesda, MD: US National Center for Biotechnology Information (2002).

Kleeff, Jorg, et al. "Pancreatic cancer." *Nature reviews Disease primers* 2.1 (2016): 1-22.

Kristensen, Vessela N., et al. "Principles and methods of integrative genomic analyses in cancer." *Nature Reviews Cancer* 14.5 (2014): 299.

Lamers, Fieke, et al. "Identification of BIRC6 as a novel intervention target for neuroblastoma therapy." *BMC cancer* 12.1 (2012): 285.

Leary, Rebecca J et al. "Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing." *Science translational medicine* 4.162 (2012): 162ra154-162ra154.

Leboeuf, Dominique, et al. "Ubiquitin ligases: a new target for RNAi therapy of hepatocellular carcinoma." (2017): 3131-3131.

Lee, Choongho, Tonia R. Wooldridge, and Laimonis A. Laimins. "Analysis of the roles of E6 binding to E6TP1 and nuclear localization in the human papillomavirus type 31 life cycle." *Virology* 358.1 (2007): 201-210.

Li, Bo, and Colin N. Dewey. "RSEM: Accurate Transcript Quantification from RNA-SeqData with or without a Reference Genome." *BMC Bioinformatics* 12 (1): 323. (2011)

Li, Jing, et al. "DIP2C expression in breast cancer and its clinical significance." *Pathology-Research and Practice* 213.11 (2017): 1394-1399.

Li, Xianran, Michael J. Scanlon, and Jianming Yu. "Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems." *Nucleic acids research* 43.7 (2015): 3614-3625.

Liao, Qi, et al. "Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network." *Nucleic acids research* 39.9 (2011): 3864-3878.

Lin, Wenchu, et al. "Kmt2a cooperates with menin to suppress tumorigenesis in mouse pancreatic islets." *Cancer biology & therapy* 17.12 (2016): 1274-1281.

Liu, Jiang-Hua, et al. "Expression and prognostic significance of lncRNA MALAT1 in pancreatic cancer tissues." *Asian Pac J Cancer Prev* 15.7 (2014): 2971-2977.

Lochhead, P. A., et al. "Activating ROCK1 somatic mutations in human cancer." *Oncogene* 29.17 (2010): 2591-2598.

Logsdon, C. D., D. M. Simeone, C. Binkley, T. Arumugam, J. K. Greenon, et al. (2003). "Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer." *Cancer research* 63(10): 2649-2657.

Lohr, M., G. Kloppel, P. Maisonneuve, A. B. Lowenfels and J. Luttges (2005). "Frequency of K-ras mutations in pancreatic intraductal neoplasias associated with pancreatic ductal adenocarcinoma and chronic pancreatitis: a meta-analysis." *Neoplasia* 7(1): 17-23.

Lorenz, Richard, et al. "The cAMP/Epac1/Rap1 pathway in pancreatic carcinoma." *Pancreas* 37.1 (2008): 102-103.

Love, Michael, Simon Anders, and Wolfgang Huber. "Differential analysis of count data—the DESeq2 package." *Genome Biol* 15.550 (2014): 10-1186.

Luk, Iris Sze Ue, et al. "BIRC6 targeting as potential therapy for advanced, enzalutamide-resistant prostate cancer." *Clinical Cancer Research* 23.6 (2017): 1542-1551.

Martin, Lance, and Howard Y. Chang. "Uncovering the role of genomic “dark matter” in human disease." *The Journal of clinical investigation* 122.5 (2012): 1589.

McGranahan, Nicholas, and Charles Swanton. "Biological and therapeutic impact of intratumor heterogeneity in cancer evolution." *Cancer cell* 27.1 (2015): 15-26.

Meucci, Stefano, et al. "Mutational load and mutational patterns in relation to age in head and neck cancer." *Oncotarget* 7.43 (2016): 69188.

Miller, Martin L., et al. "Pan-cancer analysis of mutation hotspots in protein domains." *Cell systems* 1.3 (2015): 197-209.

Moffitt, Richard A et al. "Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma." *Nature genetics* 47.10 (2015): 1168-1178.

Mularoni, Loris, et al. "OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations." *Genome biology* 17.1 (2016): 128.

Murthy, Divya, Kuldeep S. Attri, and Pankaj K. Singh. "Phosphoinositide 3-kinase signaling pathway in pancreatic ductal adenocarcinoma progression, pathogenesis, and therapeutics." *Frontiers in physiology* 9 (2018): 335.

Nagaraj, Shivashankar H., and Antonio Reverter. "A Boolean-based systems biology approach to predict novel genes associated with cancer: Application to colorectal cancer." *BMC systems biology* 5.1 (2011): 35.

Naslavsky, Michel Satya, et al. "Exomic variants of an elderly cohort of Brazilians in the ABraOM database." *Human mutation* 38.7 (2017): 751-763.

Nissan, Aviram, et al. "Colon cancer associated transcript-1: A novel RNA expressed in malignant and pre-malignant human tissues." *International journal of cancer* 130.7 (2012): 1598-1606.

Paixão, Vinícius Ferreira da. "Anotação e caracterização de novos transcritos expressos no adenocarcinoma de pâncreas: RNAs não-codificadores longos e formas alternativas de splicing associados a fenótipos tumorais e clínicos." Diss. Universidade de São Paulo, 2020.

Pelleg, Dan, and Andrew W. Moore. "X-means: Extending k-means with efficient estimation of the number of clusters." *Icml*. Vol. 1. 2000.

Pellegrina, Diogo Vieira da Silva, et al. "Insights into the function of long noncoding RNAs in sepsis revealed by gene co-expression network analysis." *Non-coding RNA* 3.1 (2017): 5.

Pertea, Mihaela, et al. "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." *Nature biotechnology* 33.3 (2015): 290.

Prior, Ian A., Paul D. Lewis, and Carla Mattos. "A comprehensive survey of Ras mutations in cancer." *Cancer research* 72.10 (2012): 2457-2467.

Raphael, Benjamin J., et al. "Integrated genomic characterization of pancreatic ductal adenocarcinoma." *Cancer Cell* 32.2 (2017): 185-203.

Reimand, Arak, Adler, Kolberg, Reisberg, Peterson, Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2016 update) *Nucleic Acids Research* 2016; doi: 10.1093/nar/gkw199

Reis, E. M. and S. Verjovski-Almeida (2012). "Perspectives of Long Non-Coding

RNAs in Cancer Diagnostics." *Frontiers in genetics* 3: 32.

Renz, Bernhard W., et al. " β 2 adrenergic-neurotrophin feedforward loop promotes pancreatic cancer." *Cancer cell* 33.1 (2018): 75-90.

Rimkus, C., et al. "Prognostic significance of downregulated expression of the candidate tumour suppressor gene SASH1 in colon cancer." *British journal of cancer* 95.10 (2006): 1419-1423.

Rosales, Rafael A., et al. "signeR: an empirical Bayesian approach to mutational signature discovery." *Bioinformatics* 33.1 (2016): 8-16.

Rütti, Sabine, et al. "Expression, phosphorylation and function of the Rab-GTPase activating protein TBC1D1 in pancreatic beta-cells." *FEBS letters* 588.1 (2014): 15-20.

Samuel, Nardin et al. "Integrated genomic, transcriptomic, and RNA-Interference analysis of genes in somatic copy number gains in pancreatic ductal adenocarcinoma." *Pancreas* 42.6 (2013): 1016-1026.

Shearer, Robert F., et al. "Functional roles of the E3 ubiquitin ligase UBR5 in cancer." *Molecular Cancer Research* 13.12 (2015): 1523-1532.

Siegel, R., Ma, J., Zou, Z. and Jemal, A. (2014), Cancer statistics, 2014. *CA A Cancer Journal for Clinicians*, 64: 9–29. doi:10.3322/caac.21208

Sin, Mandy LY, et al. "Deep sequencing of urinary RNAs for bladder cancer molecular diagnostics." *Clinical Cancer Research* 23.14 (2017): 3700-3710.

Singh, Narendra P., et al. "A simple technique for quantitation of low levels of DNA damage in individual cells." *Experimental cell research* 175.1 (1988): 184-191.

Sivakumar, Shivan, et al. "Master regulators of oncogenic KRAS response in pancreatic cancer: an integrative network biology analysis." *PLoS medicine* 14.1 (2017).

Sjöblom, Tobias, et al. "The consensus coding sequences of human breast and colorectal cancers." *science* 314.5797 (2006): 268-274.

Sorlie, Therese, et al. "Repeated observation of breast tumor subtypes in independent gene expression data sets." *Proceedings of the National Academy of Sciences of the United States of America* 100.14 (2003): 8418-8423.

Sosa, Omar Julio. Identificação e anotação funcional de novos transcritos com expressão alterada no câncer pancreático. Diss. Universidade de São Paulo, 2019.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15545-15550.

Sussman, Caroline R., Gwen Lomberk, and Raul Urrutia. "EGFR Signaling Pathways in Pancreatic Cancer Pathogenesis." *Pancreatic Cancer*. 2010

Sun, Ping-Hui, et al. "Receptor-like protein tyrosine phosphatase κ negatively regulates the apoptosis of prostate cancer cells via the JNK pathway." *International journal of oncology* 43.5 (2013): 1560-1568.

Tahira, A. C., M. S. Kubrusly, M. F. Faria, B. Dazzani, R. S. Fonseca, et al. (2011). "Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer." *Molecular cancer* 10: 141.

Taniuchi, Keisuke, et al. "ARHGEF4 predicts poor prognosis and promotes cell invasion by influencing ERK1/2 and GSK-3 α/β signaling in pancreatic cancer." *International journal of oncology* 53.5 (2018): 2224-2240.

Tong, Hui, et al. "inTs8 accelerates the epithelial-to-mesenchymal transition in hepatocellular carcinoma by upregulating the TgF- β signaling pathway." *Cancer management and research* 11 (2019): 1869.

Tuborgh, A., et al. "Complex three-way translocation involving MLL, ELL, RREB1, and CMAHP genes in an infant with acute myeloid leukemia and t (6; 19; 11)(p22. 2; p13. 1; q23. 3)." *Cytogenetic and genome research* 141.1 (2013): 7-15.

Van der Auwera, Geraldine A., et al. "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline." *Current protocols in bioinformatics* (2013): 11-10.

Waddell, Nicola et al. "Whole genomes redefine the mutational landscape of pancreatic cancer". *Nature* 518, 495–501 (2015) doi:10.1038/nature14169

Walker, Christopher J., et al. "Patterns of CTCF and ZFX3 mutation and associated outcomes in endometrial cancer." *Journal of the National Cancer Institute* 107.11 (2015): djv249.

Wan, Tao, et al. "LncRNA HEIH promotes cell proliferation, migration and invasion in cholangiocarcinoma by modulating miR-98-5p/HECTD4." *Biomedicine & Pharmacotherapy* 125 (2020): 109916.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." *Nature reviews. Genetics* 10(1): 57-63.

Wang, Zhen-Qiang, et al. "Long noncoding RNA UCA1 induced by SP1 promotes cell proliferation via recruiting EZH2 and activating AKT pathway in gastric cancer." *Cell death & disease* 8.6 (2017): e2839-e2839.

Yachida, S., S. Jones, I. Bozic, T. Antal, R. Leary, et al. (2010). "Distant metastasis occurs late during the genetic evolution of pancreatic cancer." *Nature*

467(7319): 1114-1117.

Ying, Haoqiang, et al. "Genetics and biology of pancreatic ductal adenocarcinoma." *Genes & development* 30.4 (2016): 355-385.

Yu, Qiuyun, et al. "Long non-coding RNA CCAT1 that can be activated by c-Myc promotes pancreatic cancer cell proliferation and migration." *American journal of translational research* 8.12 (2016): 5444.

Zhan, Qiong, et al. "FRAS1 knockdown reduces A549 cells migration and invasion through downregulation of FAK signaling." *International journal of clinical and experimental medicine* 7.7 (2014): 1692.

Zhan, Yi-Hong, et al. "CELSR1 is a positive regulator of endothelial cell migration and angiogenesis." *Biochemistry (Moscow)* 81.6 (2016): 591-599.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 1128.

Zhao, Lan, Hongya Zhao, and Hong Yan. "Gene expression profiling of 1200 pancreatic ductal adenocarcinoma reveals novel subtypes." *BMC cancer* 18.1 (2018): 603.

Zhao, Qi et al. "Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing." *Genome Biol* 11.11 (2010): R114.

Zong, Wei, et al. "Overexpression of SASH1 inhibits TGF- β 1-induced EMT in gastric cancer cells." *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* 24.1 (2016): 17-23.

Anexos

1. Dados clínicos dos pacientes
2. Cobertura de sequenciamento em cada amostra para cada gene
3. Todas as alterações somáticas detectadas
4. Genes com alterações somáticas
5. Vias moleculares enriquecidas por mutações somáticas
6. Rede de co-expressão gênica
7. Vias moleculares enriquecidas por mutações somáticas e transcricionais
8. Artigo em preparação que inclui a inferência funcional de lncRNAs a partir de rede de co-expressão gênica.

Os anexos (14.5MB) podem ser encontrados a partir do link <http://tiny.cc/TesePell>

ou ainda: drive.google.com/drive/folders/1fJUVLJbnQJ6MCsgaQC5APuCv9oFMmtoQ