Universidade Federal do Rio Grande do Sul
Instituto de Biociências - Centro de Biotecnologia
Curso de Graduação em Biotecnologia

Ana Carolina de Moraes Mello

ANÁLISE DE DADOS DE RNA-SEQ DE CÂNCER DO ENDOMÉTRIO: UMA BUSCA *IN SILICO* POR POTENCIAIS ALVOS TERAPÊUTICOS

Porto Alegre
2018

Ana Carolina de Moraes Mello

ANÁLISE DE DADOS DE RNA-SEQ DE CÂNCER DO ENDOMÉTRIO: UMA BUSCA *IN SILICO* POR POTENCIAIS ALVOS TERAPÊUTICOS

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Biociências da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de Bacharela em Biotecnologia.

Área de habilitação: Bioinformática

Orientadora: Prof[a]. Dr[a]. Ursula Matte
Co-Orientador: Dr. Tiago Falcon

Porto Alegre
2018

Ana Carolina de Moraes Mello

ANÁLISE DE DADOS DE RNA-SEQ DE CÂNCER DO ENDOMÉTRIO: UMA BUSCA *IN SILICO* POR POTENCIAIS ALVOS TERAPÊUTICOS

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Biociências da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de Bacharela em Biotecnologia.

Aprovado em: _____ de _____ de 2018

BANCA EXAMINADORA

Dr. Gabriel de Souza Macedo
Hospital de Clínicas de Porto Alegre

Profa. Dra. Mariana Recamonde Mendoza
Universidade Federal do Rio Grande do Sul

Profª Drª Ursula Matte (Orientadora)
Universidade Federal do Rio Grande do Sul

Dr. Tiago Falcon (Co-orientador)
Hospital de Clínicas de Porto Alegre

Resumo

O câncer do endométrio é o segundo tipo de tumor ginecológico mais frequente e, como outros tipos de câncer, apresenta elevada heterogeneidade molecular. Assim, com o objetivo de detectar assinaturas moleculares de expressão gênica que permitam diferenciar o tumor do tecido normal, foi realizada uma busca *in silico* por potenciais RNAs não codificantes reguladores da homeostase do Câncer de Endométrio, e seus alvos. Para tal, foram analisados dados de expressão do RNA total e de miRNA disponíveis no banco de dados do *The Cancer Genome Atlas*. Foram realizadas análises de expressão diferencial entre amostras do tecido tumoral primário (TP) em relação a amostras do tecido adjacente (NT), utilizando-se o pacote do R, TCGABiolinks, bem como análises de ontologia gênica e *clustering* hierárquico. Vinte e um indivíduos possuíam amostras tanto de TP, quanto de NT, somando um total de 42 amostras, onde foram observados 55.199 transcritos e 1.881 miRNAs. Quando aplicado um corte de logFC maior que 1 e menor que -1 e um FDR maior que 0,01, foram identificados 2.506 transcritos e 680 miRNAs diferencialmente expressos em TP com relação a NT. Aplicando um corte de logFC maior que 2 e menor que -2, e mantendo o valor de corte de FDR, os números reduziram para 1.208 transcritos e 514 miRNAs diferencialmente expressos. As análises de clusterização hierárquica distinguiram significativamente os dois grupos de amostras. A análise de ontologia gênica indicou que os transcritos diferencialmente expressos estão atuando no processo catabólico do DNA e na regulação negativa da organização das organelas. Sete lncRNAs estavam entre os transcritos mais diferencialmente expressos (logFC maior que 5 e menor que -5) e, dentre eles, o lncRNA *CASC22* apresentou potencial significativo para a discriminação dos estados tumoral e normal no Câncer de Endométrio.

Abstract

Endometrial Cancer is the second most common type of gynecological tumor and, just like other types of cancer, it presents elevated molecular heterogeneity. Therefore, with identifying gene expression molecular signatures that distinguish tumor from normal tissue as a goal, an in silico search for potential ncRNAs regulating the Endometrial Cancer homeostasis and its targets was performed. In order to accomplish that, Total RNA and miRNA data available at The Cancer Genome Atlas database were analyzed. Differential expression analyses were performed between primary tumor tissue samples (TP) in contrast to adjacent tissue samples (NT), using the R package TCGABiolinks, as well as gene ontology and hierarchical clustering analyses. Twenty one subjects presented both TP and NT samples, resulting in a total of 42 samples, where 55.199 transcripts and 1.881 miRNAs were observed. When a logFC cut greater than 1 and less than -1 and a FDR cut greater than 0.01 was applied, 2506 transcripts and 680 miRNAs were identified as differentially expressed in TP as compared to NT. Using a logFC cut greater than 2 and less than -2 and the same FDR cut, the numbers reduced to 1208 transcripts and 514 miRNAs differentially expressed. Hierarchical clustering analyses significantly distinguished both sample groups. Gene ontology analyses indicated that the differentially expressed transcripts were acting in the DNA catabolic process and in the negative regulation of the organelle organization. Seven lncRNAs were among the most differentially expressed transcripts (logFC greater than 5 and less than -5) and, among them, lncRNA *CASC22* presented significant potential to discriminate between tumoral and normal states in Endometrial Cancer.

Lista de Abreviações

AUC	Área abaixo da curva (*Area Under the Curve*)
CE	Câncer do Endométrio
FDR	False Discovery Rate
lincRNA	RNA não codificante intergênico de cadeia longa
lncRNA	RNA não codificante de cadeia longa
logFC	log na base 2 do valor de *Fold Change*
MAF	*Mutation Annotation File*
miRNA	microRNA
mRNA	RNA mensageiro
NCI	*National Cancer Institute*
ncRNA	RNA não codificante
NGHRI	*National Human Genome Research*
NGS	Sequenciamento de Nova Geração (*Next Generation Sequencing)*
NT	Tecido Normal Ajacente (*Normal Tissue*)
OMS	Organização Mundial da Saúde
RNA	Ácido Ribonucleico (*Ribonucleic Acid*)
ROC	*Receiver Operating Characteristic*
TCGA	*The Cancer Genome Atlas*
TP	Tecido Tumoral Primário (*Primary Tissue*)
UCEC	*Uterine Corpus Endometrial Cancer*

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Câncer: uma breve introdução

Apesar de conhecido desde 30 séculos antes de Cristo, quando egípcios, persas e indianos já se referiam a tumores malignos (para revisão, veja [1]), e a despeito de tantos avanços na medicina ao longo dos últimos dois milênios, o câncer ainda assusta tanto que é considerado a doença do século XXI. Talvez um dos maiores paradoxos de todos seja que quanto mais a ciência e as tecnologias evoluem, mais preocupações surgem a respeito dessa doença que inspira buscas incessantes por uma cura.

Em uma definição direta, o câncer ocorre através de uma multiplicação descontrolada de células de um tecido do corpo humano, que podem invadir outros tecidos durante a chamada metástase [2]. O acúmulo de mais células do que o normal acaba formando o que chamamos de neoplasia ou tumor, que só tende a crescer enquanto não for controlado. Em sua forma benigna, o tumor é resultado do crescimento de células bem diferenciadas e que não possuem a capacidade de invadir outros tecidos [2]. Neoplasias malignas, resultam da divisão de células imaturas, pouco diferenciadas, que crescem mais rapidamente e podem invadir outros tecidos se não controladas [2].

Em âmbito molecular, essa divisão desenfreada é acarretada por mutações que acabam por influenciar na regulação do processo de divisão celular [2]. Dentre conceitos importantes para entender a série de eventos que podem anteceder e manter o câncer, estão supressores de tumor e oncogenes. Proto-oncogenes são genes que normalmente promovem a replicação de células e, quando mutados ou presentes em grande quantidade, podem tornar-se permanentemente ativados, causando uma divisão descontrolada das células e gerando um tumor, sendo então denomidados oncogenes [3]. Já os supressores de tumor são genes normais que reduzem a taxa de replicação celular, reparam o DNA e regulam a morte celular programada (também chamada de apoptose) [3]. Estes, quando não ativos ou não presentes, podem levar a uma replicação celular descontrolada. Dessa forma, percebe-se que a grande diferença entre oncogenes e supressores de tumor é o fato de que os primeiros causam tumores através de sua ativação, enquanto que os segundos, causam tumores através de sua inativação.

Os mecanismos carcinogênicos não são simples como mecanismos de ligar e desligar, caso contrário, já estaríamos próximos à cura. No entanto, a natureza extremamente única e aleatória do microambiente de um tumor dificulta muito o seu entendimento e a busca por tratamentos definitivos. Em 1960, Klein e colegas [4] descreveram que, em ratos, cânceres induzidos pelo agente cancerígeno metilcolantreno exibiram resposta imune mediada por células linfáticas contra células do tumor primário. No entanto, essa resposta não afetou o tumor primário, pois ele desenvolveu um microambiente próprio, análogo ao de um tecido normal não afetado pela doença. O fato do microambiente de um tumor ser adaptável também influencia na heterogeneidade do câncer, uma propriedade que descreve

a capacidade que diferentes células cancerígenas têm de possuírem perfis morfológicos e epigenéticos distintos, incluindo morfologia celular, expressão dos genes, metabolismo, motilidade, proliferação e potencial metastático [5]. Essa heterogeneidade pode ser tanto intratumoral, ou seja, as propriedades de cada célula variam dentro do tumor primário de um mesmo paciente, quanto intertumoral, quando propriedades celulares distintas são observadas entre diferentes pacientes com o mesmo tumor [5][6].

1.2 Câncer do Endométrio

Dados epidemiológicos da Agência Internacional de Pesquisa em Câncer da Organização Mundial da Saúde (OMS) mostram que o Câncer do Endométrio (CE) abrange 4,8% da incidência mundial de câncer, e 2,1% da taxa de mortalidade relacionada a câncer [7]. O CE, a segunda forma mais comum de tumor ginecológico se considerarmos países desenvolvidos e subdesenvolvidos [8], é uma doença que ocorre quando há uma divisão descontrolada das células do endométrio, mucosa que reveste a face interna do útero, formando um tumor na região [9]. O sintoma mais comum é uma hemorragia vaginal não associada ao período de menstruação [9]. Outros sintomas incluem dor ou dificuldades ao urinar, dor durante relações sexuais, dor na região pélvica e sangramento vaginal pós-menopausa. A maioria dos casos de CE ocorre em mulheres na menopausa [10].

Dentre os fatores de risco endógenos para o desenvolvimento do CE estão obesidade, envelhecimento, menarca precoce e menopausa tardia, nunca ter se reproduzido, histórico de câncer de mama e diabetes mellitus [11]. Já os fatores exógenos incluem terapia com tamoxifeno - medicamento modulador dos receptores de estrógeno, geralmente utilizado para tratamento de câncer de mama [12][13] - fatores dietéticos, exposição à radioterapia, e altos níveis de estrogênio [14]. O CE pode ser classificado de acordo com suas características histopatológicas: carcinoma, quando se origina em células epiteliais glandulares do endométrio; sarcoma, quando as células de origem são não-glandulares do tecido conjuntivo do endométrio (para revisão, veja [15]). Já os carcinomas endometriais podem ser classificados de acordo com suas características clínicas e endócrinas: carcinomas do tipo I são dependentes de estrogênio e associados a hiperplasia endometrial; carcinomas do tipo II são independentes de estrogênio e associados com atrofia endometrial (para revisão, veja [15]).

A classificação de tumores quanto a estágios visa o agrupamento de pacientes através das características do tumor (se está muito avançado ou se recém iniciou), padronizando a escolha do tratamento mais adequado. As classificações mais utilizadas atualmente para o CE são o sistema TNM e a proposta em 2009 pela Federação Internacional de Ginecologia e Obstetria (FIGO) (tabela 1) e ambas baseiam-se no tamanho do tumor, se ele já se espalhou para nódulos linfáticos próximos ou se ele já se espalhou para diferentes tecidos [16] [17]. A cirurgia retirando tanto as trompas de falópio quanto os ovários é o

tratamento padrão para CE quando no estágio 1 e é efetiva na maioria dos casos [18]. Para estágios mais avançados, a cirurgia seguida de tratamentos como radiação, quimioterapia ou uma combinação dos dois pode ser o mais recomendado.

Tabela 1: Estágios do CE de acordo com a classificação da FIGO de 2009.

Estágio	**Descrição**
I	Está crescendo dentro do útero, talvez atingindo glândulas da cervical, mas não o tecido conjuntivo. Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
IA	Atingiu o endométrio e pode estar quase no miométrio. Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
IB	Já invadiu o miométrio, mas ainda não se espalhou para fora do útero. Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
II	Está se espalhando para os tecidos conjuntivos da cervical, mas ainda não saiu do útero. Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
III	Já se espalhou para fora do útero, mas ainda não atingiu o reto ou a bexiga urinária. Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
IIIA	Se espalhou pela serosa do útero, e/ou para as trompas de falópio ou ovários. Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
IIIB	Se espalhou pela vagina ou tecidos ao redor do útero (paramétrio). Ainda não se espalhou para os nódulos linfáticos nem para tecidos distantes.
IIIC1	Está crescendo dentro do útero. Pode ter se espalhado para tecidos próximos, mas ainda não está dentro do reto ou bexiga. Já atingiu nódulos linfáticos da pélvis, mas não atingiu a aorta ou tecidos distantes.
IIIC2	Está crescendo dentro do útero. Pode ter se espalhado para tecidos próximos, mas ainda não está dentro do reto ou bexiga. Se espalhou para nódulos linfáticos ao redor da aorta, mas não atingiu tecidos distantes.
IVA	Já atingiu a superfície do reto ou a mucosa da bexiga. Pode ou não ter atingido nódulos linfáticos próximos, mas não atingiu tecidos distantes.
IVB	Pode estar com qualquer tamanho. Atingiu nódulos linfáticos inguinais e tecidos distantes, e pode ou não ter se espalhado para outros nódulos linfáticos.

1.3 ncRNAs e seu papel no Câncer de Endométrio

Múltiplos estudos mostraram o potencial de diferentes microRNAs (miRNAs) como biomarcadores de prognósticos e diagnósticos em uma variedade de cânceres [15][19][20][21], incluindo CE [22]. miRNAs são RNAs não codificantes (ncRNAs) que regulam negativamente a expressão ao induzir RNAs mensageiros (mRNAs) para degradação ou repressão traducional após ligarem-se à 3'UTR (para revisão, veja [23][24]). Em câncer, seu papel pode ser tanto de supressor tumoral, como promotor do crescimento (oncomiRNAs) [22]. Outra classe de RNAs não codificantes, os ncRNAs de cadeia longa (lncRNAs), possuem mais que 200 pares de base e são moléculas com múltiplas funções que não podem ser inferidas pela sua sequência. Uma das funções, dentre outras, inclui organização da cromatina (abertura e fechamento) afetando a expressão gênica [25][26]. A grande maioria dos lncRNAs é formada por RNAs intergênicos longos não condificantes (lincRNAs) [27], que estão localizados entre genes codificadores de proteína, sem haver sobreposição [28].

Em CE, o lncRNA *HOTAIR* está altamente expresso [29] e contribui para a resistência induzida por cisplatina ao inibir a autofagia [25]. O silenciamento da expressão *in vivo* do *HOTAIR* suprimiu significativamente a tumorigênese endometrial e levou a tumores menores [30][31]. Outro lncRNA, o *MALAT1*, está superexpresso durante a hiperplasia endometrial e o carcinoma endometrial inicial, no entanto, sua expressão é significativamente menor em estágios avançados, bem como no estágio metastático da doença [32].

1.4 RNA-seq e análises *in silico* de transcriptomas

Com o recente avanço das técnicas de sequenciamento de genomas e com o desenvolvimento do Sequenciamento de Nova Geração (NGS), muitas possibilidades e caminhos se abriram, tendo início a era genômica em larga escala. Várias plataformas de NGS estão disponíveis e o RNA-seq (do inglês *RNA sequencing*) é um exemplo de técnica que aplica essas plataformas para análise de transcriptoma. Transcriptoma refere-se ao conjunto total de transcritos (incluindo mRNAs e ncRNAs como miRNAs) de uma célula e suas quantidades em um determinado instante, representando assim uma "foto"do conjunto de transcritos do momento estudado [33]. Esse tipo de análise permite identificar e caracterizar os transcritos quanto a níveis de expressão, inferir modificações pós-transcricionais, como padrão diferente de *splicing* e encontrar mutações e SNPs [34]. Todos os eventos biológicos que ocorrem dentro de uma célula são governados, principalmente, por mudanças de expressão de genes importantes, e essa habilidade de ativar e reprimir a expressão dos genes é um fator importante para regular todas as funções e atividades biológicas [33][35]. Tendo em mãos o perfil de expressão de uma célula e os softwares necessários, é possível realizar a quantificação das modificações dos níveis do transcrito

em diferentes ambientes ou situações em que se encontra (análises de expressão diferencial) [36], como, por exemplo, encontrar transcritos diferencialmente expressos em célula de tecido atingido por tumor em relação ao mesmo tecido saudável. Portanto, análises de transcriptoma tornaram-se uma ferramenta importantíssima na identificação de genes ou grupos de genes que desempenham papel chave no desenvolvimento de doenças e de marcadores de diagnóstico e prognóstico.

1.5 O *The Cancer Genome Atlas*

O NGS permite que grandes quantidades de dados de sequenciamento sejam gerados a cada ano e, boa parte delas, é depositada em bancos de dados públicos. O *The Cancer Genome Atlas* (TCGA), uma colaboração entre o *National Cancer Institute* (NCI) e o *National Human Genome Research Institute* (NGHRI), é um exemplo de banco de dados que já gerou mapas multidimensionais das mudanças genômicas chave de 33 tipos diferentes de tumores, dentre eles, o CE. O TCGA possui mais de 2 petabytes de dados genômicos publicamente disponíveis [37] para pesquisadores que podem utilizar todas essas informações, juntamente com seus conhecimentos, através de análises *in silico*, para melhorar a prevenção, tratamento e diagnóstico do câncer. Todos os dados utilizados neste trabalho foram provenientes deste banco de dados.

2 Hipótese

Dada a disponibilidade de dados referentes ao câncer de endométrio, o impacto e incidência deste tipo de câncer na saúde, a heterogeneidade dos tumores, e a relevância de ncRNAs no âmbito do câncer revelada em estudos recentes [15][19][20][21], nossa hipótese é de que os perfis de expressão de ncRNAs, considerando-se amostras provenientes de pacientes que possuem seu transcriptoma caracterizado para tumor primário (TP) e NT (tecido adjacente ao tumor, considerado normal) podem fornecer assinaturas genômicas (perfis de expressão) que possibilitem a caracterização e detecção de alvos com potencial de diferenciar os dois grupos (TP e NT) e, ao mesmo tempo, sirvam para futuros experimentos que visem afetar a homeostase do tumor.

3 Objetivos

O objetivo geral do trabalho é detectar potenciais alvos que possam servir como reguladores da homeostase do CE, com o foco em ncRNAs, e utilizando os dados de expressão gênica total e dados de expressão de miRNAs. Estes dados estão publicamente disponíveis no banco de dados do TCGA.

3.1 Objetivos Específicos

1. Análise da expressão gênica diferencial total (mRNAs, pre-miRNAs e lncRNAs) de amostras tumorais (TP) em relação à amostras do tecido adjacente ao CE (NT);

2. Análise da expressão diferencial de miRNAs de amostras de TP e NT de CE;

3. Detectar potenciais interações entre os genes diferencialmente expressos (mRNA, lncRNA, pre-miRNA e miRNA) através de redes de co-expressão.

4 Materiais e Métodos

No TCGA estão disponíveis dados de expressão gênica de CE (projeto UCEC) obtidos através de RNA-seq utilizando a plataforma *Illumina HiSeq*, da versão hg38 do genoma humano. O *download* dos dados de expressão das amostras e as análises foram feitos utilizando-se o *software* R (v. 3.4.0) [38], seguindo o pipeline de análise do pacote TCGABiolinks (v 2.7.1) [39], disponível no repositório digital Bioconductor [40]. Foram realizadas análises de expressão diferencial e de co-expressão. Essas análises ainda foram divididas em etapas, que divergiram nos valores de corte estatístico (Fig. 1). Os possíveis alvos encontrados foram utilizados para gerar modelos de classificação que indicam o potencial de acerto da variável (expressão de um determinado gene) através de análises de Característica de Operação do Receptor (do inglês *Receiver Operating Characteristic*, ROC). Abaixo, estão detalhados os métodos utilizados para o cumprimento de cada análise.

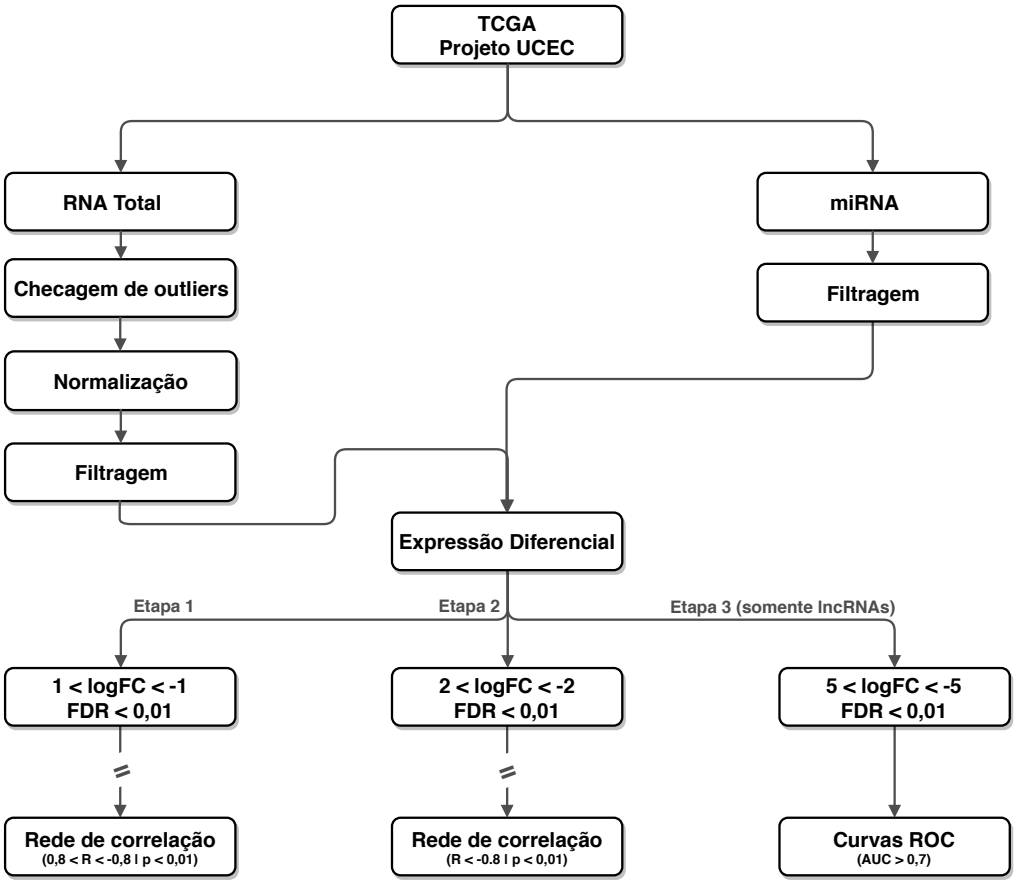


Figura 1: Fluxograma das análises. Os dados de RNA Total passam por três processos anteriores à análise de expressão diferencial, ao contrário dos dados de miRNA, que passam apenas pela filtragem. O trabalho foi dividido em 3 etapas que divergiram nos valores de corte estatístico durante as análises de expressão diferencial.

4.1 Pré-processamento dos dados

Primeiramente, foi definida uma *query* para o *download* das amostras, através da função *GDCquery*. Para os dados de RNA total (mRNA, RNA ribosomal, lncRNAs, pré-miRNAs), os parâmetros utilizados nessa função foram *project = "TCGA-UCEC"*, *data.category = "Gene Expression Quatification"*, *workflow.type = "HTSeq - FPKM-UQ"* e *legacy = FALSE*. Para os dados de miRNA, os parâmetros foram *project = "TCGA-UCEC"*, *data.category = "miRNA Expression Quatification"*, e *legacy = FALSE*. Estes dados estão disponíveis separadamente pelo fato de que miRNAs não são detectados quando se faz uma análise de expressão de RNA Total devido ao tamanho muito pequeno do transcrito, sendo necessário fazer uma filtração diferente das amostras para analisar a expressão de miRNAs. Após, foi feito o *download* dos dados harmonizados de CE (hg38) do TCGA usando a função *GDCdownload* com a *query* definida anteriormente como parâmetro, e com *method = "api"*. Após o *download*, a função *GDCprepare* preparou a *query* de acordo com o tipo de dado a ser analisado. Foi feita uma seleção para que fossem utilizados somente pacientes que possuíssem amostras tanto de TP, quanto de NT, para melhor comparação da diferença de expressão.

As amostras de RNA total selecionadas passaram por checagem de *outliers* com a função *TCGAanalyze_Preprocessing*. Todas as amostras apresentaram correlação de Spearman maior ou igual a 0,8, sendo mantidas nas análises subsequentes. Depois, a função *TCGAanalyze_Normalization* foi utilizada duas vezes para a normalização das amostras a partir do conteúdo GC com o parâmetro *method = "gcContent"* e, depois, a partir do comprimento de genes com *method = "geneLength"*, ambas utilizando também o parâmetro *geneInfo = "geneInfoHT"*. Em seguida, a função *TCGAanalyze_Filtering* realizou a filtragem por *quantile*, como recomendado por Bullard e colaboradores [41], com o parâmetro *method = "quantile"*.

O pré-processamento das amostras de miRNA é mais curto, uma vez que elas passam apenas pelo passo da filtragem por *quantile*. Ainda foi feita uma análise de correlação das amostras após a filtragem e todas apresentaram correlação maior que 0,7.

4.2 Análises de Expressão Diferencial

As análises de expressão diferencial do RNA total e de miRNAs das amostras de TP em relação as amostras de NT foram feitas com a função *TCGAanalyze_DEA*, utilizando como parâmetros os valores de corte de FDR e logFC especificados à seguir, e *method = "glmLRT"*. Os resultados foram visualizados através de *volcano plots* e de *heat maps*. Os *heat maps* de expressão foram plotados utilizando a função *heatmap.2* do pacote gplots (v. 3.0.1) [42]. Foram realizados *clustering* hierárquicos utilizando-se o pacote pvclust (v. 2.0-0) [43] para a visualização e noção estatística do agrupamento das amostras. Os parâmetros utilizados na contrução do *cluster* foram *method.dist = "correlation"*,

method.hclust = "complete" e *nboot = 1000*, assim, o método de distância utilizado foi o de correlação da expressão das amostras, para um *clustering* completo com 1000 *bootstraps*.

4.3 Análises de Correlação de Expressão

Para os transcritos diferencialmente expressos selecionados, foi realizada uma análise de correlação Spearman para detectar quais RNAs regulatórios estão negativamente e positivamente correlacionados com outros RNAs. As correlações selecionadas foram utilizadas como entrada para a construção de redes de correlação no Cytoscape (v. 3.5.1) [44].

4.4 Cortes Estatísticos

Estre trabalho foi subdividido em etapas devido a dificuldade de se encontrar um número de transcritos viável para a construção e análise de uma rede de correlação, como era parte do objetivo. Em cada etapa, diferentes cortes estatísticos foram propostos visando a diminuição de transcritos diferencialmente expressos que apresentavam significância estatística.

4.4.1 Etapa 1

Na primeira etapa, foram considerados significativos estatisticamente transcritos diferencialmente expressos que apresentavam um *log*2 *fold change* (logFC) maior que 1 e menor que -1, e um *false discovery rate* (FDR) menor que 0,01. Do RNA total, apenas os mensageiros (mRNAs), pré-miRNAs e lncRNAs que se encaixavam nesses quesitos seguiram para análises de correlação. Todos os miRNAs diferencialmente expressos significativamente seguiram para as análises. Ainda foram realizadas análises de enriquecimento dos genes diferencialmente expressos com a função *TCGAanalyze_EAcomplete* que foram visualizadas em gráfico de barras com a função *TCGAvisualize_EAbarplot*, respectivamente. Nas análises de correlação, foram aceitas as interações com valor de correlação (r) menor que -0,8 e maior que 0,8, e valor de p menor que 0,05 como significativas estatisticamente para a construção de uma rede.

4.4.2 Etapa 2

Na segunta tentativa, seguiram para as análises de correlação apenas os mRNAs, pré-miRNAs, lncRNAs e miRNAs que apresentaram logFC entre 2 e -2, e FDR menor que 0,01. Foram selecionadas apenas as correlações negativas com valor de r menor que -0,8. A tabela de correlações ainda foi filtrada para que só seguissem para a construção da rede as interações entre lncRNAs e RNAs.

4.4.3 Etapa 3

Nesta etapa, o objetivo de construir uma rede de correlação foi temporariamente abandonado, e o foco voltou-se à validação de possíveis descritores de condição. Para isso, apenas os lncRNAs que obtiveram um logFC maior que 5 e menor que -5 juntamente com um FDR menor que 0,01 na análise de expressão diferencial foram considerados. Para estes, foi feita a construção de sua curva ROC.

4.5 Curva ROC

Análises ROC são utilizadas na área clínica para quantificar a qualidade com que determinado teste ou sistema discriminou entre dois estados: doente e saudável [45]. Neste trabalho, essa análise foi aplicada nos lncRNAs que estavam mais diferencialmente expressos (logFC maior que 5 e menor que -5) entre o tecido tumoral e o tecido normal, ou seja, os que possuíam maiores condições para desempenhar o papel de marcadores do câncer. Os resultados da análise são visualizados em gráficos da curva ROC, gerados com o auxilio do pacote pRoc [46]. O índice de precisão utilizado foi a área abaixo da curva (AUC), sendo o valor aceitável maior que 0,7.

Os gráficos gerados estão baseados na noção de escala de separação, nos quais os resultados da média de expressão do transcrito para o tecido tumoral e normal formam um par de distribuições sobrepostas. A completa separação entre as duas distribuições implica numa discriminação perfeita dos dois estados. Já uma completa sobreposição implica em não discriminação.

A área abaixo da curva (AUC) tem valor máximo de 1 e é uma medida de precisão que combina a sensibilidade e a especificidade, descrevendo a qualidade do marcador predito [47]. Neste trabalho, foram considerados bons marcadores aqueles que obtiveram uma AUC acima de 0,7.

5 Resultados

Com o *download* dos dados, foram obtidas 587 amostras de RNA total e 579 amostras de miRNA. Buscamos apenas as amostras TP e NT vindas do mesmo paciente e presentes tanto nos dados de miRNA quanto nos dados de RNA total. Com isso, ao final dessa filtragem, restaram 21 indivíduos (21 amostras de TP e 21 amostras de NT) que seguiram para as análises. As amostras variaram entre o estágio e o tipo do câncer. Foi encontrado um total de 55.199 transcritos (considerando-se isoformas de mRNA, lncRNA e pré-miRNA) e 1.881 miRNAS.

5.1 Normalização das Amostras de RNA Total

As amostras de RNA total, primeiramente, passaram pelo passo de checagem de *outliers*, observando-se que apresentaram correlação maior ou igual a 0,8 (Apêndice A). Estas, seguiram para o processo de normalização, que demonstrou-se eficaz, uma vez que as amostras não estão apresentando grandes variações em relação à mediana (Fig. 2).

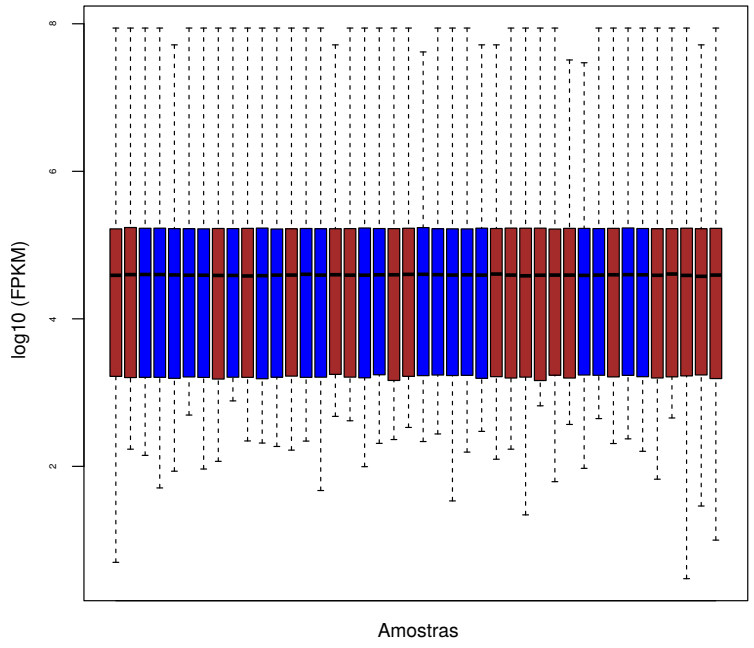


Figura 2: *Boxplot* após a normalização das amostras de RNA Total. Em azul, as amostras de tumor primário (TP), em vermelho, as amostras do tecido adjacente (NT). As linhas pontilhadas indicam os desvios dos valores de expressão.

5.2 Checagem da Correlação das Amostras de miRNA

A checagem da correlação das amostras de miRNA, feita após a filtragem, demonstrou que elas não apresentaram grandes variações em relação à mediana (Fig. 3). Todas as amostras apresentaram correlação maior ou igual a 0,7 (Apêndice B).

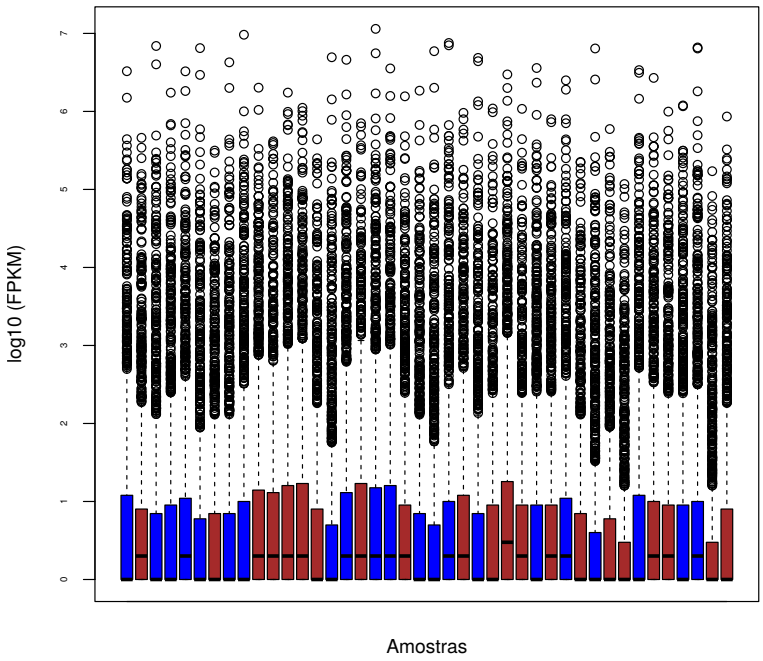


Figura 3: *Boxplot* após a filtragem das amostras de miRNA. Em azul, as amostras de tumor primário (TP), em vermelho, as amostras do tecido adjacente (NT). As linhas pontilhadas indicam os desvios dos valores de expressão.

5.3 Etapa 1

5.3.1 Expressão Diferencial de RNA Total

Dos 55.199 transcritos iniciais, 2.506 foram identificados como diferencialmente expressos em TP em relação a NT, considerando-se um valor de FDR de 0,01, e um logFC menor que -1 para os pouco expressos, e maior que 1 para os mais expressos (Fig. 4). Desses, 1.265 estão mais expressos em TP.

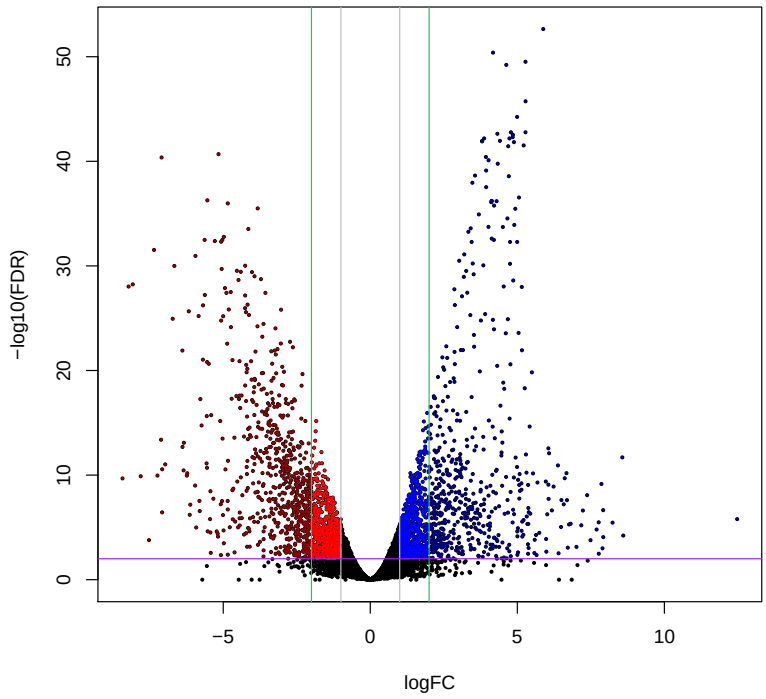


Figura 4: *Volcano plot* da análise de expressão diferencial do RNA Total em tumor primário (TP) com relação ao tecido adjacente (NT). Em azul claro, transcritos mais expressos com logFC maior que 1. Em azul escuro, transcitos mais expressos com $log2$ de *Fold Change* (logFC) maior que 2. Em vermelho claro, transcritos pouco expressos com logFC menor que -1. Em vermelho escuro, transcritos pouco expressos com logFC menor que -2. A linha horizontal roxa representa o corte em 1 no eixo do $log10(FDR)$. As linhas verticais cinzas representam os cortes no eixo do logFC em -1 e 1. As linhas verticias verdes representam os cortes no eixo do logFC em -2 e 2.

O *heat map* de expressão dos 50 transcritos mais diferencialmente expressos (Fig. 5) revelou uma nítida separação dos dois grupos de amostras: TP e NT. Para comprovação estatística da separação dos grupos, foi construído um dendrograma (Fig. 6).

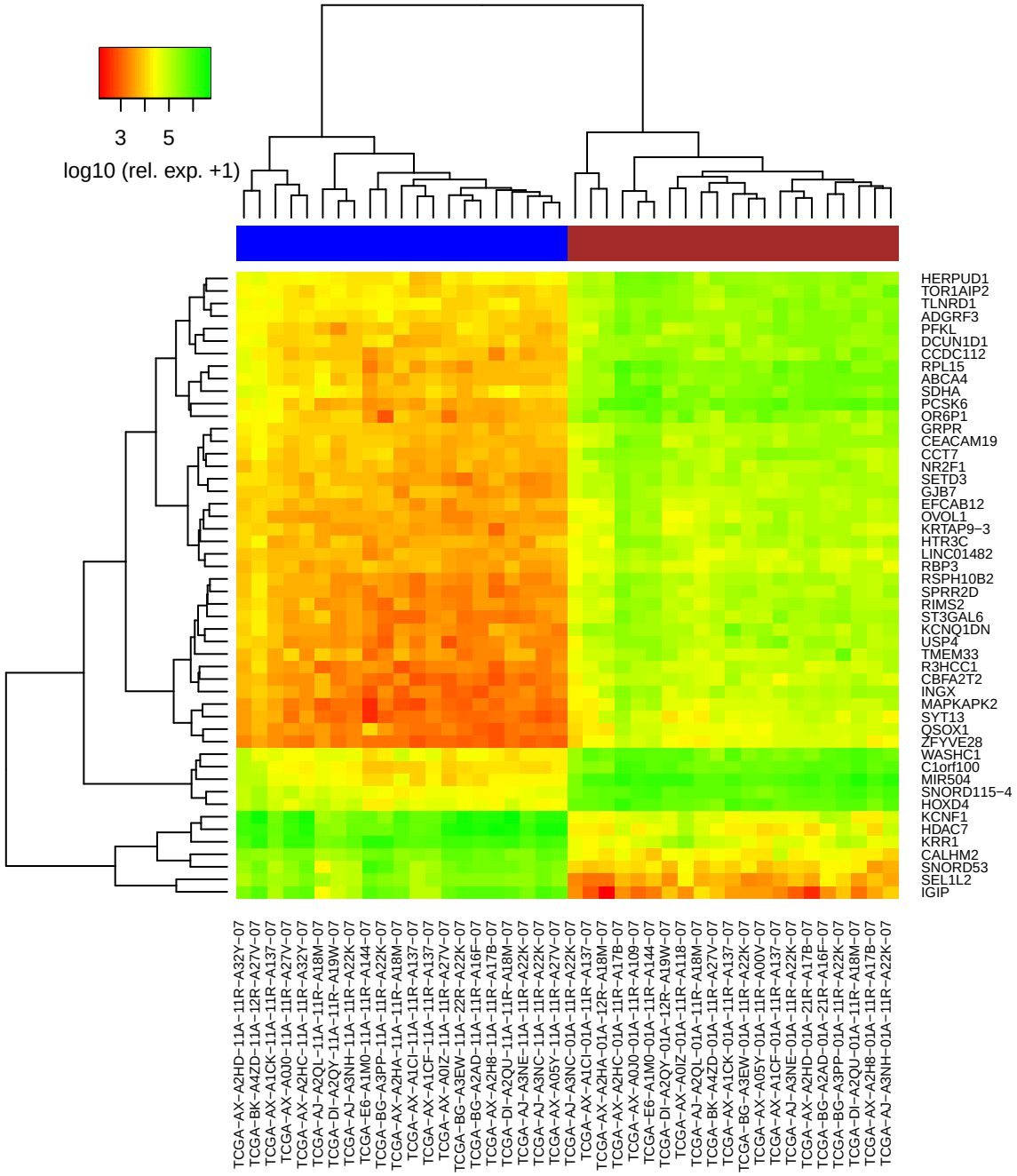


Figura 5: *Heat map* da expressão dos 50 transcritos mais diferencialmente expressos da etapa 1 (RNA Total, logFC maior que 1 e menor que -1, FDR menor que 0,01). Em azul, o agrupamento das amostras de tecido adjacente (NT) e em vermelho, o agrupamento das amostras de tumor primário (TP).

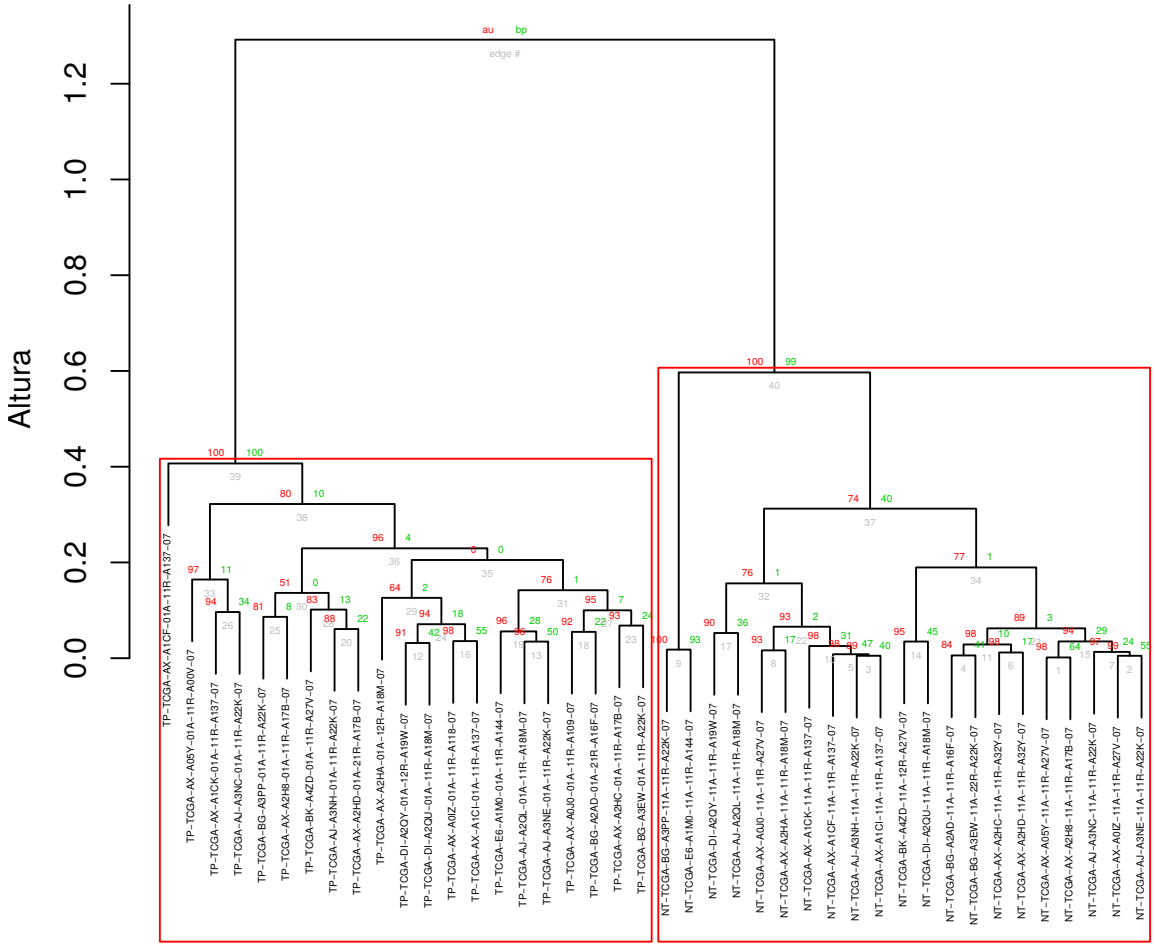


Figura 6: Clusterização hierárquica representando a separação das amostras dos grupos tumor primário (TP) e tecido adjacente (NT), considerando-se os 50 transcritos mais diferencialmente expressos da etapa 1 (RNA Total, logFC maior que 1 e menor que -1, FDR menor que 0,01). Os valores em vermelho (au) representam o suporte de grupo. Valores de au maiores ou iguais a 95 foram considerados estatisticamente significativos. Os valores em verde (bp) representam o suporte de *bootstrap*. Em cinza (edge) os limites dos ramos. Os quadrados vermelhos destacam os maiores grupos estatisticamente significativos.

O resultado da análise de ontologia gênica pode ser visualizado na figura 7. Esta análise indicou que os transcritos diferencialmente expressos estão atuando em processos biológicos como o processo catabólico do DNA e na regulação negativa da organização das organelas.

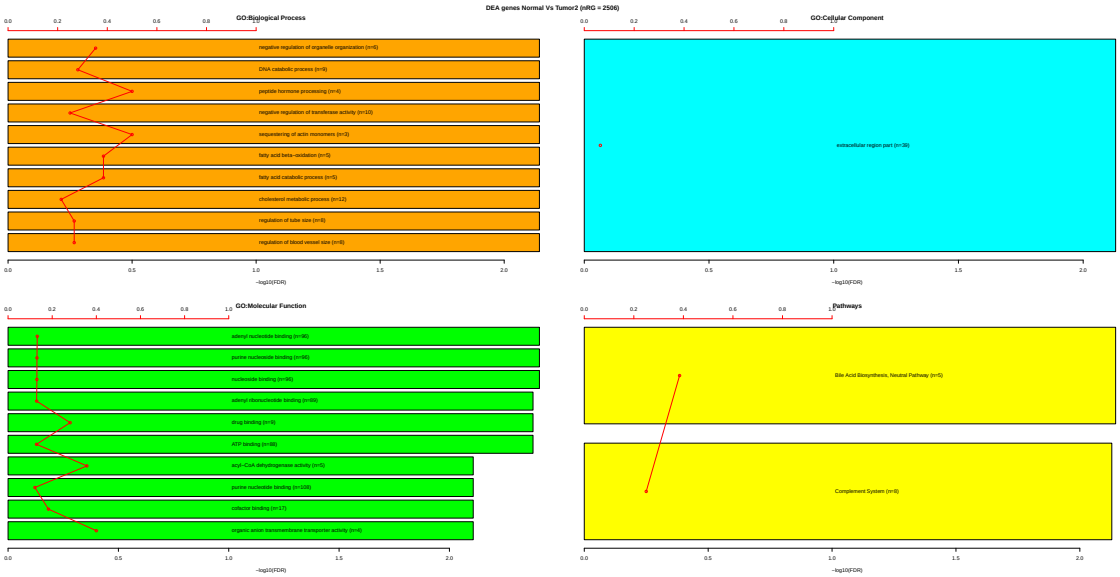


Figura 7: Resultado da análise de ontologia gênica dos transcritos diferencialmente expressos da etapa 1. A linha vermelha representa a taxa de transcritos diferencialmente expressos encontrados para via em relação ao número total de genes para aquela via específica. Dentro de cada barra, n é o número de transcritos. Os tamanhos das barras estão de acordo com o logarítimo na base 10 do valor de FDR de cada via/ontologia enriquecida.

5.3.2 Expressão Diferencial de miRNA

Dos 1.881 miRNAs iniciais, 680 estavam diferencialmente expressos nas amostras de tecido tumoral em relação às amostras de tecido normal adjacente, considerando um FDR de 0,01 e um logFC menor que -1 para os pouco expressos, e maior que 1 para os mais expressos (Fig. 8). Desses, 620 estavam mais expressos em TP.

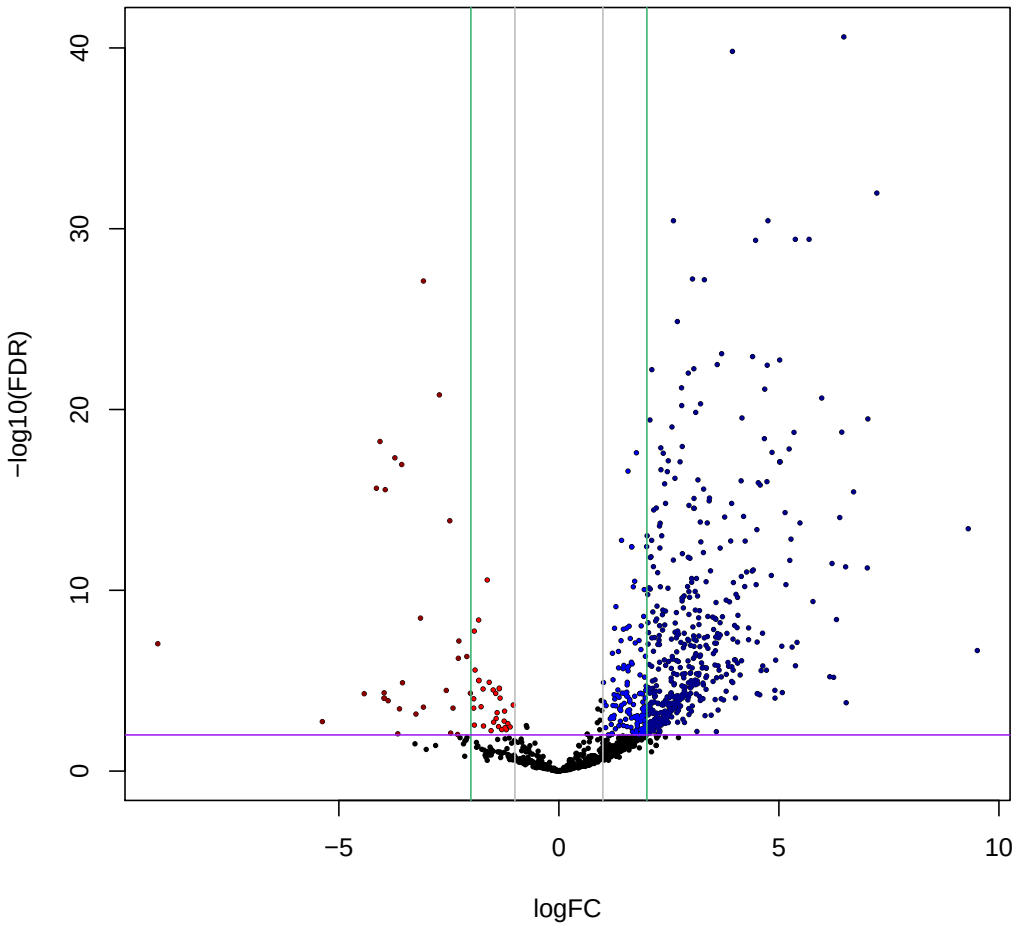


Figura 8: *Volcano plot* da análise de expressão diferencial dos microRNAs em tumor primário (TP) com relação ao tecido adjacente (NT). Em azul claro, os miRNAs mais expressos com $log2$ *Fold Change* (logFC) maior que 1. Em azul escurto, os miRNAs mais expressos com logFC maior que 2. Em vermelho claro, os miRNAs pouco expressos com logFC menor que -1. Em vermelho escuro, os miRNAs pouco expressos com logFC menor que -2. A linha horizontal roxa representa o corte em 1 no eixo de $-log10(FDR)$ (FDR menor que 0,01). As linhas verticais cinzas representam os cortes no eixo do logFC em -1 e 1. As linhas verticais verdes representam os cortes no eixo do logFC em -2 e 2.

O *heat map* de expressão dos 50 transcritos mais diferencialmente expressos (Fig. 9) revelou uma nítida separação dos dois grupos de amostras: TP e NT. Para comprovação estatística da separação dos grupos, foi construído um dendrograma. Nesta análise, uma das amostras do grupo NT agrupou com amostras TP (Fig. 10), pois, diferentemente do dendrograma gerado no *heatmap*, foi utilizada uma distância baseada em correlação. Foram gerados três grupos estatisticamente significativos (suporte de grupo maior que 95%) sendo um deles composto apenas por amostras NT.

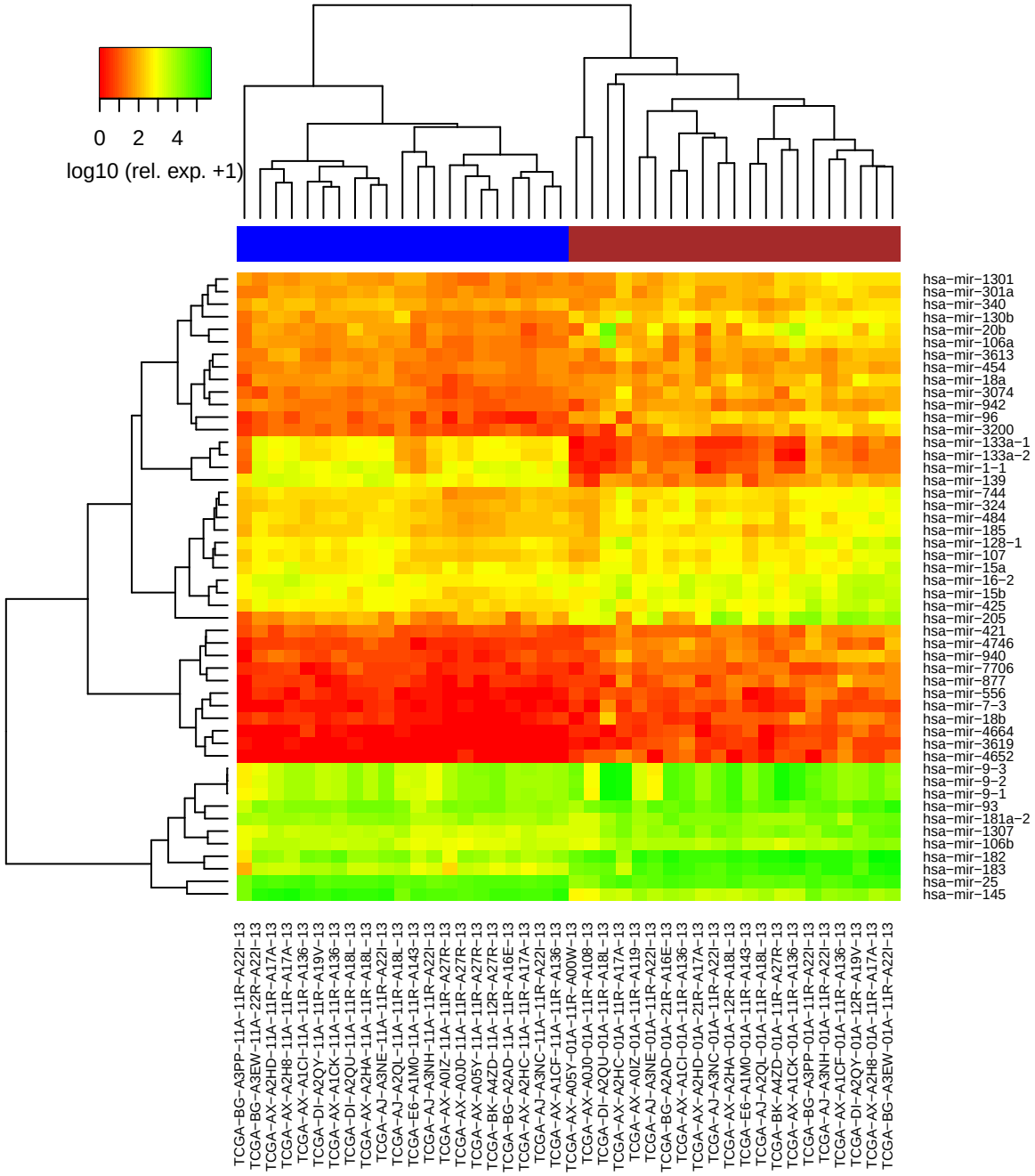


Figura 9: *Heat map* da expressão dos 50 miRNAs mais diferencialmente expressos da etapa 1 (miRNA, logFC maior que 1 e menor que -1, FDR menor que 0,01). Em azul, o agrupamento das amostras de tecido adjacente (NT) e em vermelho, o agrupamento das amostras de tumor primário (TP).

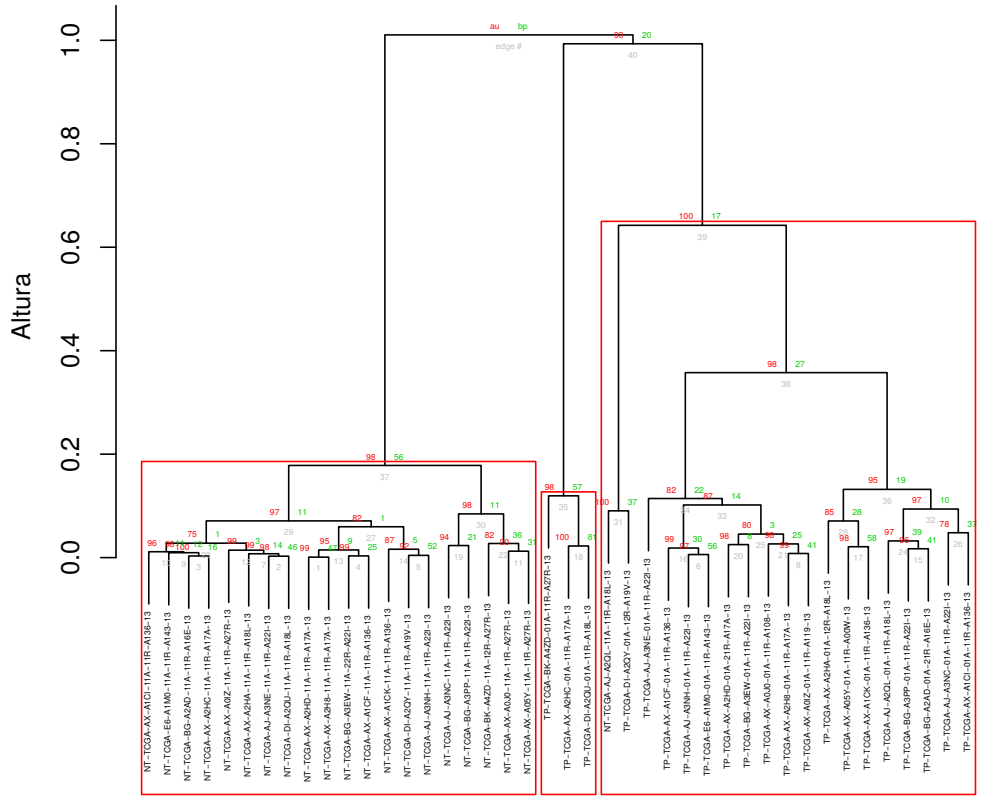


Figura 10: Clusterização hierárquica representando a separação das amostras dos grupos tumor primário (TP) e tecido adjacente (NT), considerando-se os 50 miRNAs mais diferencialmente expressos da etapa 1 (miRNA, logFC maior que 1 e menor que -1, FDR menor que 0,01). Os valores em vermelho (au) representam o suporte de grupo. Valores de au maiores ou iguais a 95 foram considerados estatisticamente significativos. Os valores em verde (bp) representam o suporte de *bootstrap*. Em cinza (edge) os limites dos ramos. Os quadrados vermelhos destacam os maiores grupos estatisticamente significativos.

5.3.3 Análise de Correlação entre os transcritos

A análise de correlação de Spearman apresentou 11.581 correlações com valor de r acima de 0,8 e abaixo de -0,8 e valor de p menor que 0,01. A construção de uma rede para melhor visualização e identificação de transcritos reguladores tornou-se inviável. Foi feito um corte estatístico ainda mais rigoroso numa tentativa de reduzir o número de transcritos que entrariam para a rede de correlações durante a Etapa 2, descrita à seguir.

5.4 Etapa 2

5.4.1 Expressão Diferencial de RNA Total

Nesta etapa, foi considerado um FDR de 0,01 e um logFC menor que -2 para os menos expressos em TP, e maior que 2 para os mais expressos em TP (Fig. 4). Foi um total de 1.208 diferencialmente expressos em tecido tumoral em relação ao tecido normal adjacente. Desses, 567 estão mais expressos em TP.

O *heat map* de expressão dos 50 transcritos mais diferencialmente expressos (Fig. 11) revelou uma nítida separação dos dois grupos de amostras: TP e NT. Para comprovação estatística da separação dos grupos, foi construído um dendrograma (Fig. 12).

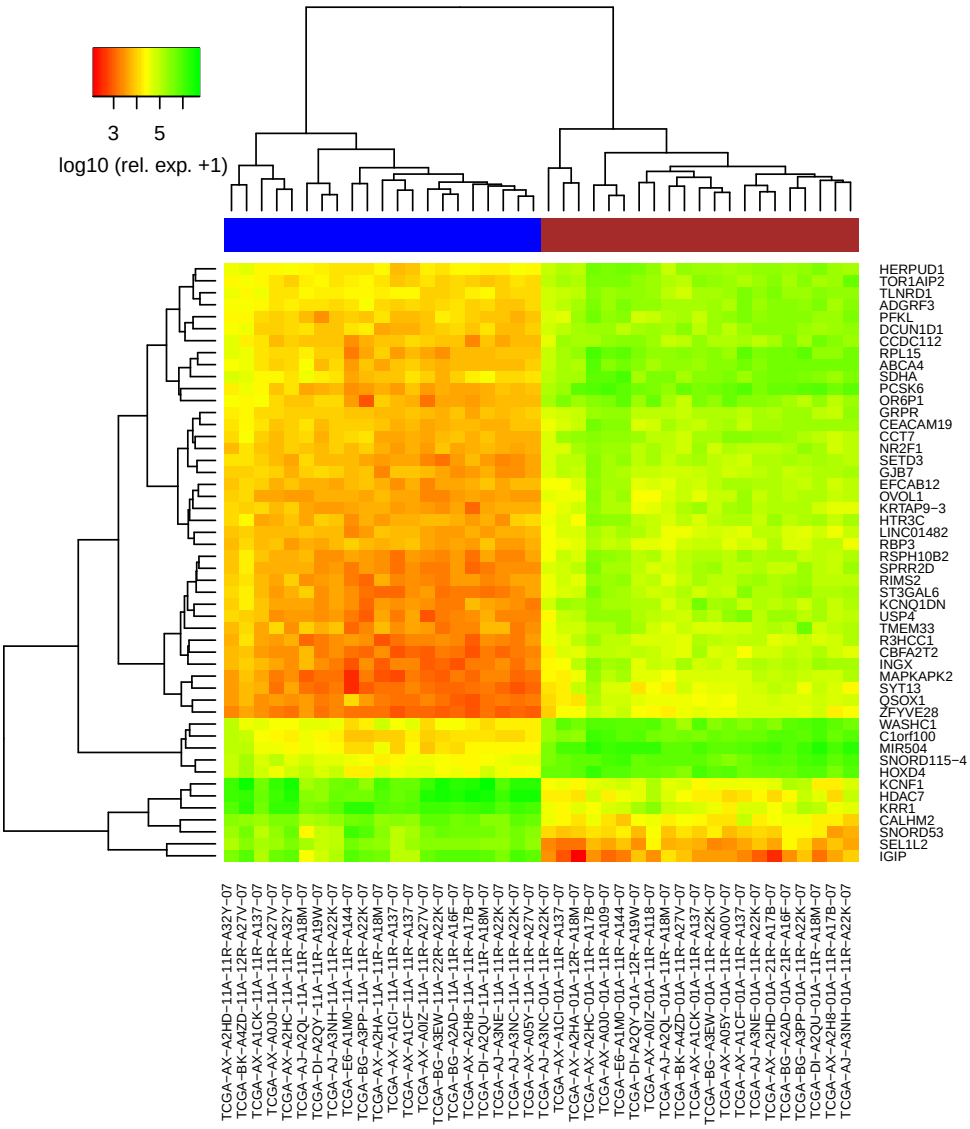


Figura 11: *Heat map* da expressão dos 50 transcritos mais diferencialmente expressos da etapa 2 (RNA Total, logFC maior que 2 e menor que -2, FDR menor que 0,01). Em azul, o agrupamento das amostras de tecido adjacente (NT) e em vermelho, o agrupamento das amostras de tumor primário (TP).

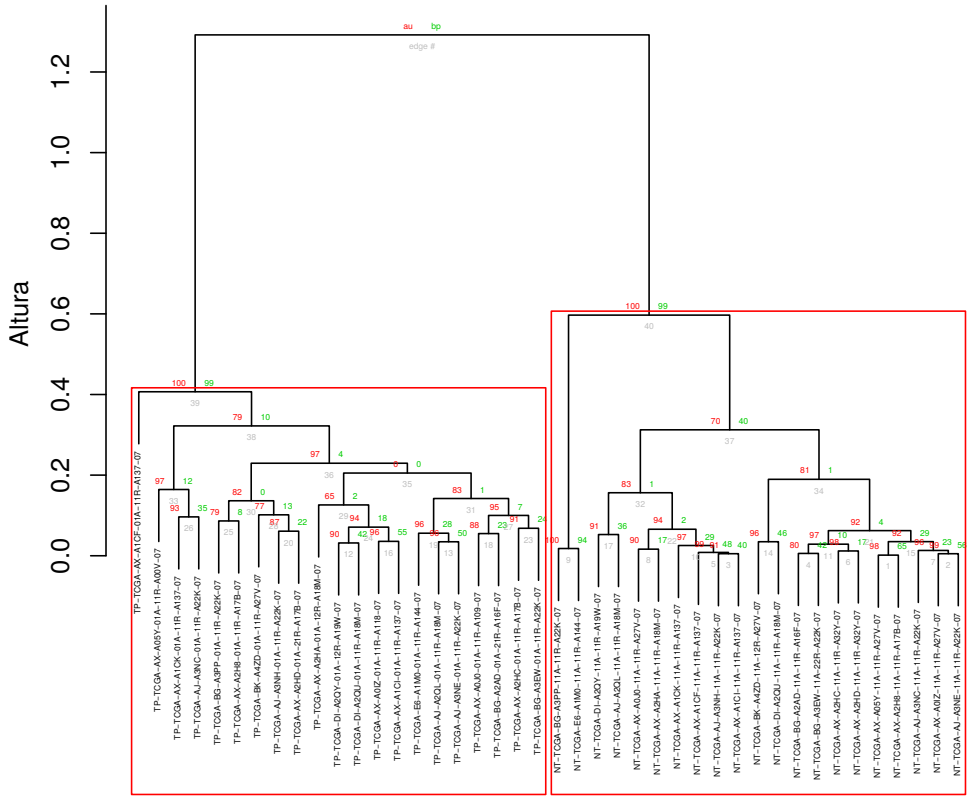


Figura 12: Clusterização hierárquica representando a separação das amostras dos grupos tumor primário (TP) e tecido adjacente (NT), considerando-se os 50 transcritos mais diferencialmente expressos da etapa 2 (RNA Total, logFC maior que 2 e menor que -2, FDR menor que 0,01). Os valores em vermelho (au) representam o suporte de grupo. Valores de au maiores ou iguais a 95 foram considerados estatisticamente significativos. Os valores em verde (bp) representam o suporte de *bootstrap*. Em cinza (edge) os limites dos ramos. Os quadrados vermelhos destacam os maiores grupos estatisticamente significativos.

5.4.2 Expressão Diferencial de miRNA

Dos 1.881 miRNAs iniciais, 514 estavam diferencialmente expressos nas amostras de tecido tumoral em relação às amostras de tecido normal adjacente, considerando um FDR de 0,01 e um logFC menor que -2 para os pouco expressos, e maior que 2 para os mais expressos (Fig. 8). Desses, 485 estavam mais expressos em TP.

A expressão dos 50 miRNAs mais diferencialmente expressos (Fig. 13) revelou uma nítida separação dos dois grupos de amostras: TP e NT. Para comprovação estatística, foi construída uma árvore de clusterização (Fig. 14). Apesar de haver dois grandes

grupos, cada um composto por amostras TP ou NT, estes fazem parte de um grupo maior, estatisticamente aproximadas, uma vez que o suporte de seus respectivos grupos é menor que 95%. Um terceiro grupo, composto por três amostras TP destacou-se do grupo maior.

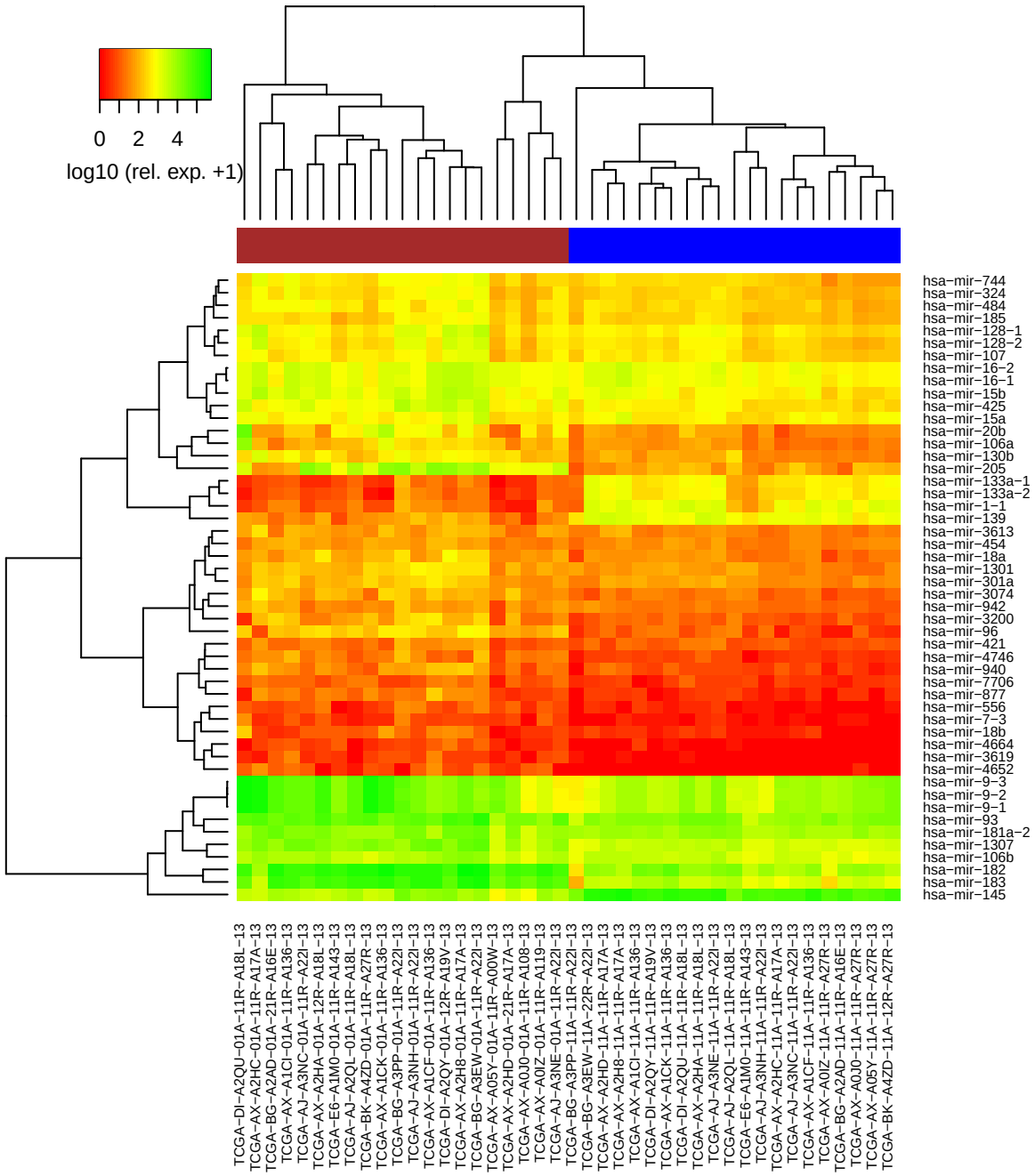


Figura 13: *Heat map* da expressão dos 50 miRNAs mais diferencialmente expressos da etapa 2 (miRNAs, logFC maior que 2 e menor que -2, FDR menor que 0,01). Em azul, o agrupamento das amostras de tecido adjacente (NT) e em vermelho, o agrupamento das amostras de tumor primário (TP).

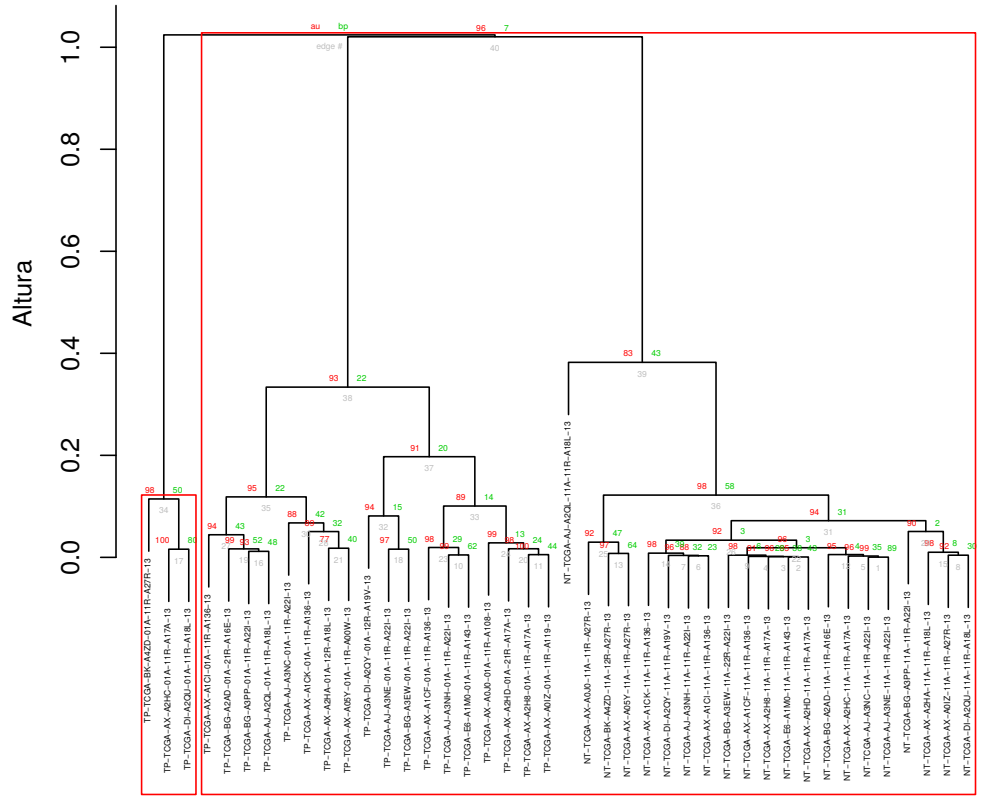


Figura 14: Clusterização hierárquica representando a separação das amostras dos grupos tumor primário (TP) e tecido adjacente (NT), considerando-se os 50 miRNAs mais diferencialmente expressos da etapa 2 (miRNAs, logFC maior que 2 e menor que -2, FDR menor que 0,01). Os valores em vermelho (au) representam o suporte de grupo. Valores de au maiores ou iguais a 95 foram considerados estatisticamente significativos. Os valores em verde (bp) representam o suporte de *bootstrap*. Em cinza (edge) os limites dos ramos. Os quadrados vermelhos destacam os maiores grupos estatisticamente significativos.

5.4.3 Análise de Correlação entre os transcritos

Nesta nova tentativa de visualização das correlações entre os transcritos diferencialmente expressos, foram selecionadas somente as interações com valor de r abaixo de -0,8, uma vez que correlações positivas, neste caso, não são de interesse, já que os miRNAs reguladores, em sua maioria, são repressores. Ou seja, as correlações positivas não apresentariam potencial de representar a regulação da expressão gênica. No entanto, o número de correlações continuou alto, 971, tornando, novamente, muito complicada a análise e visualização da rede (Apêndice C). Assim, o objetivo de construção de uma rede de cor-

relação para a identificação de reguladores em vias metabólicas bem determinadas foi impossibilitado, como está descrito à seguir na etapa 3.

5.5 Etapa 3

Nesta etapa, o foco foi a identificação dos lncRNAs que teriam maior possibilidade de servir como assinatura molecular diferenciando o CE da condição normal. Para isso, era necessário analisar os transcritos mais diferencialmente expressos. Dos sete lncRNAs selecionados (logFC maior que 5 e menor que menos 5), dois estavam menos expressos em TP: *LINC00870* e *LINC01269*. Os cinco restantes estavam mais expressos em TP: *BSN-AS2*, *CASC22*, *LINC01185*, *MIAT* e *RAD21-AS1*. Para estes transcritos, foi gerada uma curva ROC. O *CASC22* foi o único lncRNA que satisfez nosso critério de significância estatística, com uma AUC de 0,728 e um intervalo de confiança de 95% entre 0,563 e 0,893 (Fig. 15). Todos os demais obtiveram uma AUC abaixo de 0,7 (Apêndice D).

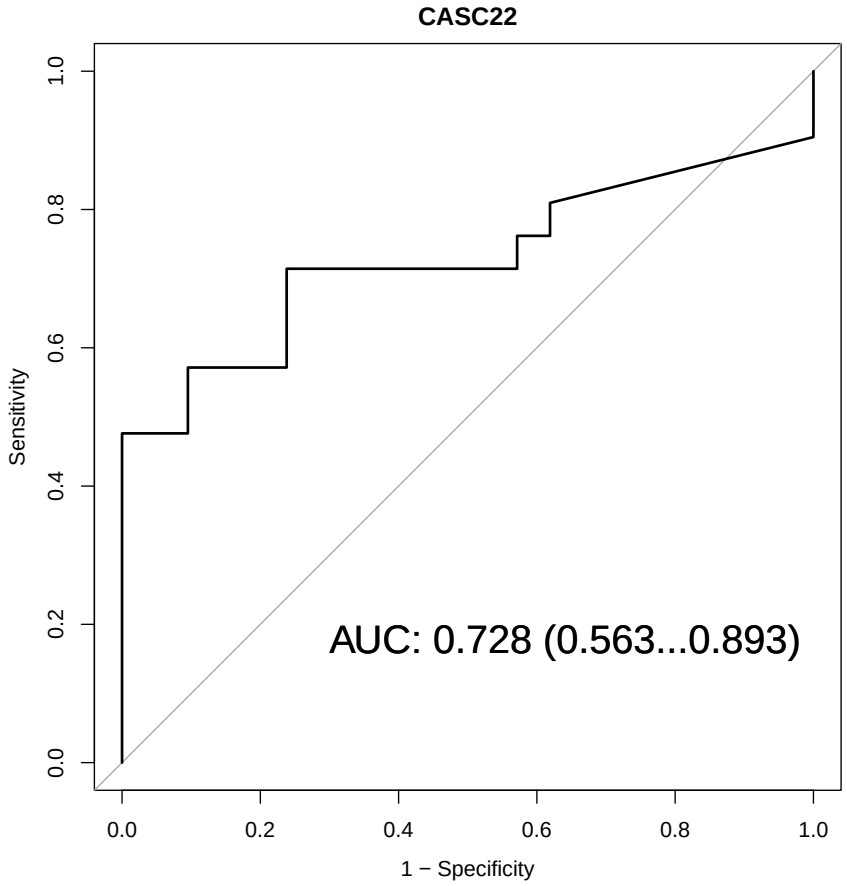


Figura 15: Curva ROC do lncRNA *CASC22* destacando-se o valor de AUC. No eixo y, a sensitividade (fração dos verdadeiros positivos). No eixo x, a 1-especificidade (fração dos falsos negativos).

6 Discussão

Uma forte correlação e uma boa normalização reduzem consideravelmente a possibilidade de que os transcritos apontados como diferencialmente expressos entre TP e NT estejam nessa condição devido a artefatos experimentais. Adicionalmente, a utilização de amostras TP e NT do mesmo paciente proporciona equilíbrio na heterogeineidade das amostras, evitando que sejam analisadas amostras com expressão excessivamente elevadas ou reduzidas sem seu respectivo controle de comparação. Além disso, reduz a possibilidade de influência de fatores externos às variações encontradas, de forma que para cada amostra normal existe uma tumoral sujeita às mesmas condições. Assim, a variação na expressão gênica pode ser confiavelmente relacionada às influências do ambiente tumoral. No entanto, o alto número de transcritos diferencialmente expressos, mesmo quando foi aplicado um corte mais rígido (logFC maior que 2 e menor que -2), resultou em um volume de dados que impossibilitou uma análise mais restrita sobre as funções desses genes e como eles interagem entre si, o que dificultou as visualizações das redes de co-expressão e a detecção de vias metabólicas diretamente relacionadas às ações dos RNAs reguladores.

Grande parte da literatura que diz respeito ao perfil de expressão no CE faz a comparação entre os diferentes subtipos do câncer, ao invés de dar enfoque à diferença de expressão no tumor em relação ao tecido normal (como exemplo, veja [48][49][50]). Nos trabalhos que analisaram TP e NT, foi observado um elevado número de transcritos diferencialmente expressos [51][52]. Risinger e colaboradores [51] observaram 6.168 transcritos diferencialmente expressos, considerando-se os mRNAs em amostras de tumor em estágio inicial, porém consideraram apenas análises de teste T utilizando um corte de p menor que 0,05. Realizando um corte de FDR como neste trabalho de conclusão (FDR menor que 0,01), esse número cai para 5.293 transcritos. Quando comparados com os 2.506 transcritos de RNA Total encontrados na etapa 1 do presente trabalho, foram observados 291 transcritos em comum. A comparação com os 1.208 transcritos diferencialmente expressos observados na etapa 2 resultou em 134 em comum. Saghir e colaboradores [52], com dados gerados por análise microarranjo, observaram um total de 621 genes diferencialmente expressos com um logFC maior que 2 e menor que -2 entre tecido tumoral e normal. Desses, 146 estavam mais expressos e 476 menos expressos no tumor em relação ao tecido normal. Em nossos dados, a minoria, 567, também estava mais expressa, enquanto que 641 estavam menos expressos. A soma dessas informações indica que as amostras individuais apresentam grande heterogeneidade e pouca sobreposição entre os trabalhos relacionados a CE comparando-se os tecidos tumorais e normais.

Os miRNAs miR-93, miR-205, miR-944 e miR-145, que foram observados altamente expressos em CE em diversos trabalhos [53][54][55][56], também estavam na mesma condição na presente análise. Em contraste ao observado por Gong e colaboradores [57], o miRNA-194 cuja baixa expressão está associada a um prognóstico ruim, encontrou-se altamente

expresso em nossas análises, apesar de que não foram realizadas correlações com os prognósticos dos pacientes. Os miRNAs mir-183, mir-182, mir-429, mir-135b e mir-200a estão mais expressos tanto nas nossas análises, como no trabalho realizado por Jurcevic e colaboradores [58]. Estes autores também apontaram que os miRNAs miR-1247, miR-424-3p, miR-376c, miR-542-5p e miR-377 estão menos expressos em CE, bem como foi observado em nosso trabalho.

Curiosamente, os lncRNAs mais citados na literatura como mais expressos durante o CE, *MALAT1*, *HOTAIR*, *OVAL*, *H19* e *SRA* (para revisão, veja [29][32][59]) não foram observados neste trabalho. Quando foram selecionados diretamente os lncRNAs que apresentavam expressão diferencial com logFC maior que 5 para os mais expressos e menor que -5 para os pouco expressos, foi possível construir a curva ROC para os 7 lncRNAs mais diferencialmente expressos encontrados. Aplicando um corte de AUC maior que 0,7 apara garantir uma boa confiabilidade do resultado, apenas o *CASC22* foi considerado um bom regulador da homeostase do CE, com uma AUC de 0,728. Também conhecido como "Gene de Suscetibilidade a Câncer 22", o *CASC22* é um lincRNA localizado no cromossomo 16, que está mais expresso nas amostras de CE, com um logFC de 5,21. Em 2014, ele foi associado ao aumento significativo do risco de Câncer de Mama quando há a troca da base C pela base T na posição rs12325489C>T, uma vez que essa mudança destrói o sítio de ligação do miRNA-370, influenciando na atividade transcricional do *CASC22*, como revelado por análises bioquímicas *in vitro* e *in vivo* [60]. No arquivo MAF (do inglês *Mutation Anotation File*) obtido a partir do TCGA, utilizando o hg38 como referência, o *CASC22* não apresentou mutação no CE. Já o miRNA-370 não apresentou variação estatisticamente significativa em sua expressão no CE.

7 Conclusões e Perspectivas

Pode-se concluir que as amostras individuais de CE apresentam grande heterogeneidade e pouca sobreposição entre os trabalhos presentes na literatura. Grande parte das pesquisas relacionadas ao perfil de expressão do CE tem enfoque na diferenciação dos subtipos do tumor, sendo assim, nosso trabalho é inovador por focar em ncRNAs e sua expressão diferenciada em TP com relação a NT. Um desafio para o seguimento deste trabalho seria analisar a imensa rede de correlações (Apêndice C) para que os potenciais lncRNAs alvos mostrados nela sejam utilizados para predizer sua interação utilizando o software disponível online TargetScan (release 7.1) [61]. Ainda deveriam ser realizadas análises de ontologia de processos biológicos das proteínas associadas à rede utilizando os *plugins* BiNGO e ClueGO do Cytoscape [62].

Adicionalmente, foi identificado um potencial marcador ainda não descrito na literatura para o Câncer de Endométrio, o *CASC22*. Não é de nosso conhecimento trabalhos que descrevem este lincRNA em CE. Diferentemente do que ocorre no Câncer de Mama, em CE não há mutação no *CASC22* e nem variação da expressão do miRNA-370. As interações do *CASC22* quando em ambiente do Câncer de Endométrio ainda devem ser analisadas. Por este trabalho tratar-se de avaliações *in silico*, testes *in vitro* e *in vivo* devem ser realizados para validação dos resultados.

Referências

[1] TEIXEIRA, L. A.; FONSECA, C. O. *De doença desconhecida a problema de saúde pública: o INCA e o controle do câncer no brasil*. 1. ed. Praça Cruz Vermelha, 23 – Centro, 20231-130 – Rio de Janeiro – RJ: Instituto Nacional de Câncer (INCA), 2007. v. 1.

[2] NATIONAL CANCER INSTITUTE. What is cancer? Disponível em: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Acessado em: 29/05/2018.

[3] AMERICAN CANCER SOCIETY. Oncogenes and tumor suppressor genes. Disponível em: https://www.cancer.org/cancer/cancer-causes/genetics/genes-and-cancer/oncogenes-tumor-suppressor-genes.html. Acessado em: 16/04/2018.

[4] KLEIN, G.; SJOGREN, H. O.; KLEIN, E.; HELLSTROM, K. E. Demonstration of resistance against methylcholanthreneinduced sarcomas in the primary autochthonous host. *Cancer Research*, v. 20, n. 11, p. 1561–1572, 1960.

[5] MARUSYK, A.; POLYAK, K. Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, v. 1805, n. 1, p. 105 – 117, 2010.

[6] VOGELSTEIN, B.; PAPADOPOULOS, N.; VELCULESCU, V. E.; ZHOU, S.; DIAZ, L. A.; KINZLER, K. W. Cancer genome landscapes. *Science*, Washington, v. 339, n. 6127, p. 1546–1558, 2013.

[7] FERLAY, J.; SOERJOMATARAM, I.; DIKSHIT, R.; ESER, S.; MATHERS, C.; REBELO, M.; PARKIN, D. M.; FORMAN, D.; BRAY, F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*, v. 136, n. 5, p. E359–386, Mar 2015.

[8] JEMAL, A.; BRAY, F.; CENTER, M. M.; FERLAY, J.; WARD, E.; FORMAN, D. Global cancer statistics. *CA Cancer J Clin*, v. 61, n. 2, p. 69–90, 2011.

[9] NATIONAL CANCER INSTITUTE. Endometrial cancer treatment. Disponível em: https://www.cancer.gov/types/uterine/patient/endometrial-treatment-pdq#section/all. Acessado em: 06/04/2018.

[10] GALAAL, K.; AL MOUNDHRI, M.; BRYANT, A.; LOPES, A. D.; LAWRIE, T. A. Adjuvant chemotherapy for advanced endometrial cancer. *Cochrane Database Syst Rev*, , n. 5, p. CD010681, May 2014.

[11] SASO, S.; CHATTERJEE, J.; GEORGIOU, E.; DITRI, A. M.; SMITH, J. R.; GHAEM-MAGHAMI, S. Endometrial cancer. *BMJ*, v. 343, p. d3954, Jul 2011.

[12] NATIONAL CENTER FOR BIOTECHNOLOGGY INFORMATION. Pubchem compound database. Disponível em: https://pubchem.ncbi.nlm.nih.gov/compound/2733526. Acessado em: 29/05/2018.

[13] JORDAN, V. C. The role of tamoxifen in the treatment and prevention of breast cancer. *Curr Probl Cancer*, v. 16, n. 3, p. 129–176, 1992.

[14] MURALI, R.; SOSLOW, R. A.; WEIGELT, B. Classification of endometrial carcinoma: more than two types. *Lancet Oncol.*, v. 15, n. 7, p. e268–278, Jun 2014.

[15] PICHLER, M.; CALIN, G. A. MicroRNAs in cancer: from developmental genes in worms to their clinical application in patients. *Br. J. Cancer*, v. 113, n. 4, p. 569–573, Aug 2015.

[16] FREEMAN, S. J.; ALY, A. M.; KATAOKA, M. Y.; ADDLEY, H. C.; REINHOLD, C.; SALA, E. The revised FIGO staging system for uterine malignancies: implications for MR imaging. *Radiographics*, v. 32, n. 6, p. 1805–1827, Oct 2012.

[17] ON CANCER, A. J. C. *Corpus uteri-carcinoma and carcinosarcoma: Ajcc cancer staging manual*. 8. ed. New York, NY: Springer, 2017. v. 1.

[18] MORICE, P.; LEARY, A.; CREUTZBERG, C.; ABU-RUSTUM, N.; DARAI, E. Endometrial cancer. *The Lancet*, v. 387, n. 10023, p. 1094–1108, 2016.

[19] DIAZ, G.; MELIS, M.; TICE, A.; KLEINER, D. E.; MISHRA, L.; ZAMBONI, F.; FARCI, P. Identification of microRNAs specifically expressed in hepatitis C virus-associated hepatocellular carcinoma. *Int. J. Cancer*, v. 133, n. 4, p. 816–824, Aug 2013.

[20] DING, W.; YANG, H.; GONG, S.; SHI, W.; XIAO, J.; GU, J.; WANG, Y.; HE, B. Candidate miRNAs and pathogenesis investigation for hepatocellular carcinoma based on bioinformatics analysis. *Oncol Lett*, v. 13, n. 5, p. 3409–3414, May 2017.

[21] FALCON, T.; FREITAS, M.; MELLO, A. C.; DA SILVA, M. R. A.; MATTE, U.; COUTINHO, L. Analysis of The Cancer Genome Atlas (TCGA) Data Reveals Novel Putative ncRNAs Targets in Hepatocellular Carcinoma. *BioMed Research International*, 2018. Aceito para publicação.

[22] YANOKURA, M.; BANNO, K.; IIDA, M.; IRIE, H.; UMENE, K.; MASUDA, K.; KOBAYASHI, Y.; TOMINAGA, E.; AOKI, D. MicroRNAS in endometrial cancer: recent advances and potential clinical applications. *EXCLI J*, v. 14, p. 190–198, 2015.

[23] CHEN, C. Z. MicroRNAs as oncogenes and tumor suppressors. *N. Engl. J. Med.*, v. 353, n. 17, p. 1768–1771, Oct 2005.

[24] KUERSTEN, S.; GOODWIN, E. B. The power of the 3' UTR: translational control and development. *Nat. Rev. Genet.*, v. 4, n. 8, p. 626–637, Aug 2003.

[25] SUN, M. Y.; ZHU, J. Y.; ZHANG, C. Y.; ZHANG, M.; SONG, Y. N.; RAHMAN, K.; ZHANG, L. J.; ZHANG, H. Autophagy regulated by lncRNA HOTAIR contributes to the cisplatin-induced resistance in endometrial cancer cells. *Biotechnol. Lett.*, v. 39, n. 10, p. 1477–1484, Oct 2017.

[26] BERNSTEIN, E.; ALLIS, C. D. RNA meets chromatin. *Genes Dev.*, v. 19, n. 14, p. 1635–1655, Jul 2005.

[27] TAKENAKA, K.; CHEN, B. J.; MODESITT, S. C.; BYRNE, F. L.; HOEHN, K. L.; JANITZ, M. The emerging role of long non-coding RNAs in endometrial cancer. *Cancer Genet*, v. 209, n. 10, p. 445–455, Oct 2016.

[28] HRDLICKOVA, B.; DE ALMEIDA, R. C.; BOREK, Z.; WITHOFF, S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta*, v. 1842, n. 10, p. 1910–1922, Oct 2014.

[29] SMOLLE, M. A.; BULLOCK, M. D.; LING, H.; PICHLER, M.; HAYBAECK, J. Long Non-Coding RNAs in Endometrial Carcinoma. *Int J Mol Sci*, v. 16, n. 11, p. 26463–26472, Nov 2015.

[30] HUANG, L.; LIAO, L. M.; LIU, A. W.; WU, J. B.; CHENG, X. L.; LIN, J. X.; ZHENG, M. Overexpression of long noncoding RNA HOTAIR predicts a poor prognosis in patients with cervical cancer. *Arch. Gynecol. Obstet.*, v. 290, n. 4, p. 717–723, Oct 2014.

[31] HUANG, J.; KE, P.; GUO, L.; WANG, W.; TAN, H.; LIANG, Y.; YAO, S. Lentivirus-mediated RNA interference targeting the long noncoding RNA HOTAIR inhibits proliferation and invasion of endometrial carcinoma cells in vitro and in vivo. *Int. J. Gynecol. Cancer*, v. 24, n. 4, p. 635–642, May 2014.

[32] ZHAO, Y.; YANG, Y.; TROVIK, J.; SUN, K.; ZHOU, L.; JIANG, P.; LAU, T. S.; HOIVIK, E. A.; SALVESEN, H. B.; SUN, H.; WANG, H. A novel wnt regulatory axis in endometrioid endometrial cancer. *Cancer Res.*, v. 74, n. 18, p. 5103–5117, Sep 2014.

[33] WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, v. 10, n. 1, p. 57–63, Jan 2009.

[34] CHU, Y.; COREY, D. R. RNA sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, v. 22, n. 4, p. 271–274, 2012.

[35] MURRAY, D.; DORAN, P.; MACMATHUNA, P.; MOSS, A. C. In silico gene expression analysis–an overview. *Mol. Cancer*, v. 6, p. 50, Aug 2007.

[36] MAHER, C. A.; KUMAR-SINHA, C.; CAO, X.; KALYANA-SUNDARAM, S.; HAN, B.; JING, X.; SAM, L.; BARRETTE, T.; PALANISAMY, N.; CHINNAIYAN, A. M. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, London, v. 458, n. 7234, p. 97–101, Mar 2009.

[37] WEINSTEIN, J. N.; COLLISSON, E. A.; MILLS, G. B.; SHAW, K. M.; OZENBERGER, B. A.; ELLROTT, K.; SHMULEVICH, I.; SANDER, C.; STUART, J. M.; CANCER GENOME ATLAS RESEARCH NETWORK. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, v. 45, n. 10, p. 1113–1120, Oct 2013.

[38] R FOUNDATION FOR STATISTICAL COMPUTING. R: A language and environment for statistical computing. Disponível em: https://www.R-project.org/. Acessado em: 12/01/2018.

[39] COLAPRICO, A.; SILVA, T. C.; OLSEN, C.; GAROFANO, L.; CAVA, C.; GAROLINI, D.; SABEDOT, T. S.; MALTA, T. M.; PAGNOTTA, S. M.; CASTIGLIONI, I.; CECCARELLI, M.; BONTEMPI, G.; NOUSHMEHR, H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, v. 44, n. 8, p. e71, 05 2016.

[40] GENTLEMAN, R. C.; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J.; HORNIK, K.; HOTHORN, T.; HUBER, W.; IACUS, S.; IRIZARRY, R.; LEISCH, F.; LI, C.; MAECHLER, M.; ROSSINI, A. J.; SAWITZKI, G.; SMITH, C.; SMYTH, G.; TIERNEY, L.; YANG, J. Y.; ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, v. 5, n. 10, p. R80, 2004.

[41] BULLARD, J. H.; PURDOM, E.; HANSEN, K. D.; DUDOIT, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, v. 11, p. 94, Feb 2010.

[42] WARNES, G. R.; BOLKER, B.; BONEBAKKER, L.; GENTLEMAN, R.; LIAW, W. H. A.; LUMLEY, T.; MAECHLER, M.; MAGNUSSON, A.; MOELLER, S.; SCHWARTZ, M.; VENABLES, B. Various r programming tools for plotting data. R package version 3.0.1. Disponível em: https://CRAN.R-project.org/package=gplots. Acessado em: 22/08/2017.

[43] SUZUKI, R.; SHIMODAIRA, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, v. 22, n. 12, p. 1540–1542, Jun 2006.

[44] SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N. S.; WANG, J. T.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, v. 13, n. 11, p. 2498–2504, Nov 2003.

[45] SWETS, J. A. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychol Bull*, v. 99, n. 1, p. 100–117, Jan 1986.

[46] ROBIN, X.; TURCK, N.; HAINARD, A.; TIBERTI, N.; LISACEK, F.; SANCHEZ, J. C.; MULLER, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, v. 12, p. 77, Mar 2011.

[47] KUMAR, R.; INDRAYAN, A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatr*, v. 48, n. 4, p. 277–287, Apr 2011.

[48] SALVESEN, H. B.; CARTER, S. L.; MANNELQVIST, M.; DUTT, A.; GETZ, G.; STEFANSSON, I. M.; RAEDER, M. B.; SOS, M. L.; ENGELSEN, I. B.; TROVIK, J.; WIK, E.; GREULICH, H.; B?, T. H.; JONASSEN, I.; THOMAS, R. K.; ZANDER, T.; GARRAWAY, L. A.; OYAN, A. M.; SELLERS, W. R.; KALLAND, K. H.; MEYERSON, M.; AKSLEN, L. A.; BEROUKHIM, R. Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation. *Proc. Natl. Acad. Sci. U.S.A.*, v. 106, n. 12, p. 4834–4839, Mar 2009.

[49] SMID-KOOPMAN, E.; BLOK, L. J.; HELMERHORST, T. J.; CHADHA-AJWANI, S.; BURGER, C. W.; BRINKMANN, A. O.; HUIKESHOVEN, F. J. Gene expression profiling in human endometrial cancer tissue samples: utility and diagnostic value. *Gynecol. Oncol.*, v. 93, n. 2, p. 292–300, May 2004.

[50] LEVINE, D. A.; THE CANCER GENOME ATLAS RESEARCH NETWORK. Integrated genomic characterization of endometrial carcinoma. *Nature*, London, v. 497, n. 7447, p. 67–73, May 2013.

[51] RISINGER, J. I.; ALLARD, J.; CHANDRAN, U.; DAY, R.; CHANDRAMOULI, G. V.; MILLER, C.; ZAHN, C.; OLIVER, J.; LITZI, T.; MARCUS, C.; DUBIL, E.; BYRD, K.; CASSABLANCA, Y.; BECICH, M.; BERCHUCK, A.; DARCY, K. M.; HAMILTON, C. A.; CONRADS, T. P.; MAXWELL, G. L. Gene expression analysis of early stage endometrial cancers reveals unique transcripts associated with grade and histology but not depth of invasion. *Front Oncol*, v. 3, p. 139, 2013.

[52] SAGHIR, F. S.; ROSE, I. M.; DALI, A. Z.; SHAMSUDDIN, Z.; JAMAL, A. R.; MOKHTAR, N. M. Gene expression profiling and cancer-related pathways in type I endometrial carcinoma. *Int. J. Gynecol. Cancer*, v. 20, n. 5, p. 724–731, Jul 2010.

[53] CHEN, S.; CHEN, X.; SUN, K. X.; XIU, Y. L.; LIU, B. L.; FENG, M. X.; SANG, X. B.; ZHAO, Y. MicroRNA-93 Promotes Epithelial-Mesenchymal Transition of Endometrial Carcinoma Cells. *PLoS ONE*, v. 11, n. 11, p. e0165776, 2016.

[54] ZHOU, X.; GAO, Q.; WANG, J.; ZHANG, X.; LIU, K.; DUAN, Z. Linc-RNA-RoR acts as a "sponge"against mediation of the differentiation of endometrial cancer stem cells by microRNA-145. *Gynecol. Oncol.*, v. 133, n. 2, p. 333–339, May 2014.

[55] WILCZYNSKI, M.; DANIELSKA, J.; DZIENIECKA, M.; SZYMANSKA, B.; WOJCIECHOWSKI, M.; MALINOWSKI, A. Prognostic and Clinical Significance of miRNA-205 in Endometrioid Endometrial Cancer. *PLoS ONE*, v. 11, n. 10, p. e0164687, 2016.

[56] JAYARAMAN, M.; RADHAKRISHNAN, R.; MATHEWS, C. A.; YAN, M.; HUSAIN, S.; MOXLEY, K. M.; SONG, Y. S.; DHANASEKARAN, D. N. Identification of novel diagnostic and prognostic miRNA signatures in endometrial cancer. *Genes Cancer*, v. 8, n. 5-6, p. 566–576, May 2017.

[57] GONG, B.; YUE, Y.; WANG, R.; ZHANG, Y.; JIN, Q.; ZHOU, X. Overexpression of microRNA-194 suppresses the epithelial-mesenchymal transition in targeting stem cell transcription factor Sox3 in endometrial carcinoma stem cells. *Tumour Biol.*, v. 39, n. 6, p. 1010428317706217, Jun 2017.

[58] JURCEVIC, S.; OLSSON, B.; KLINGA-LEVAN, K. MicroRNA expression in human endometrial adenocarcinoma. *Cancer Cell Int.*, v. 14, n. 1, p. 88, 2014.

[59] VALLONE, C.; RIGON, G.; GULIA, C.; BAFFA, A.; VOTINO, R.; MOROSETTI, G.; ZAAMI, S.; BRIGANTI, V.; CATANIA, F.; GAFFI, M.; NUCCIOTTI, R.; COSTANTINI, F. M.; PIERGENTILI, R.; PUTIGNANI, L.; SIGNORE, F. Non-Coding RNAs and Endometrial Cancer. *Genes (Basel)*, v. 9, n. 4, Mar 2018.

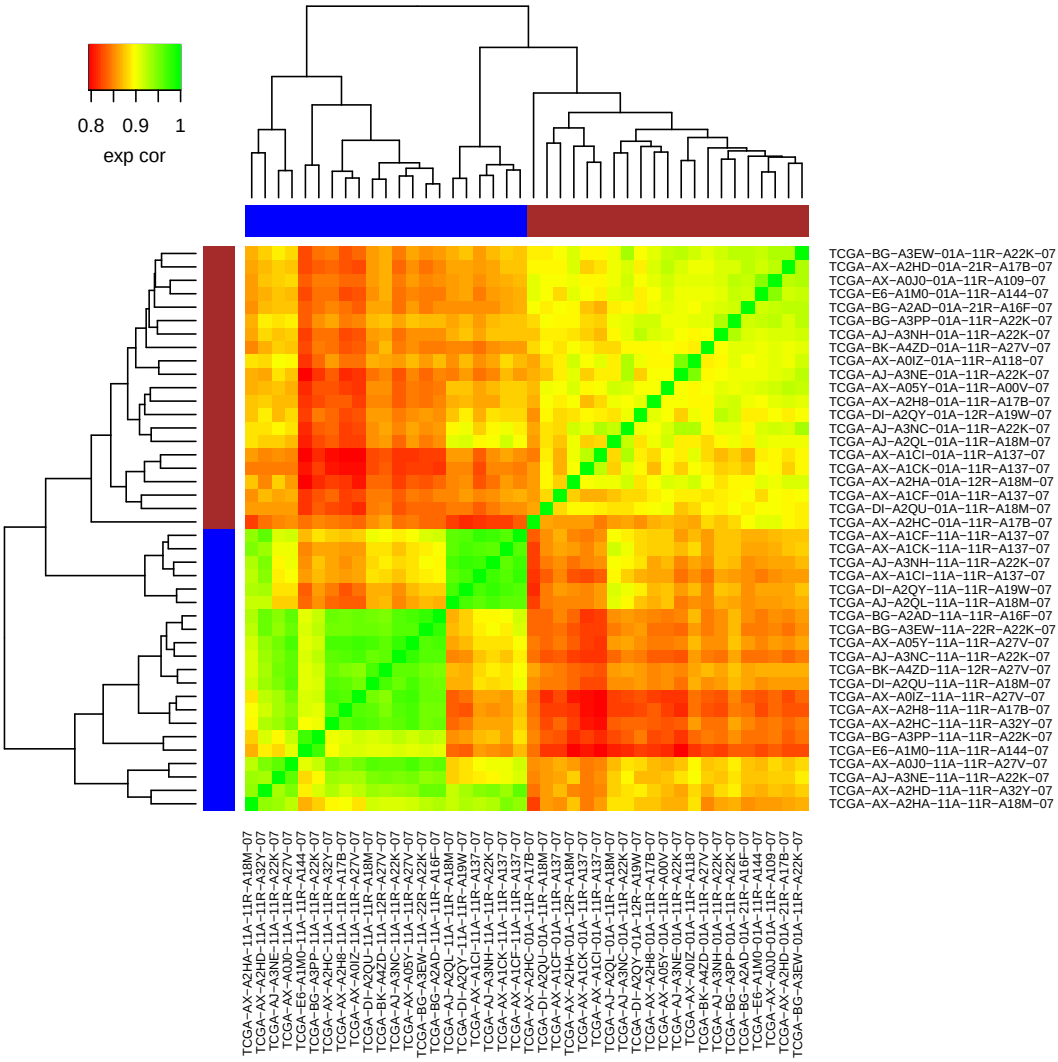
[60] LI, N.; ZHOU, P.; ZHENG, J.; DENG, J.; WU, H.; LI, W.; LI, F.; LI, H.; LU, J.; ZHOU, Y.; ZHANG, C. A polymorphism rs12325489C¿T in the lincRNA-ENST00000515084 exon was found to modulate breast cancer risk via GWAS-based association analyses. *PLoS ONE*, v. 9, n. 5, p. e98251, 2014.

[61] AGARWAL, V.; BELL, G. W.; NAM, J. W.; BARTEL, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, v. 4, Aug 2015.

[62] MAERE, S.; HEYMANS, K.; KUIPER, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, v. 21, n. 16, p. 3448–3449, Aug 2005.

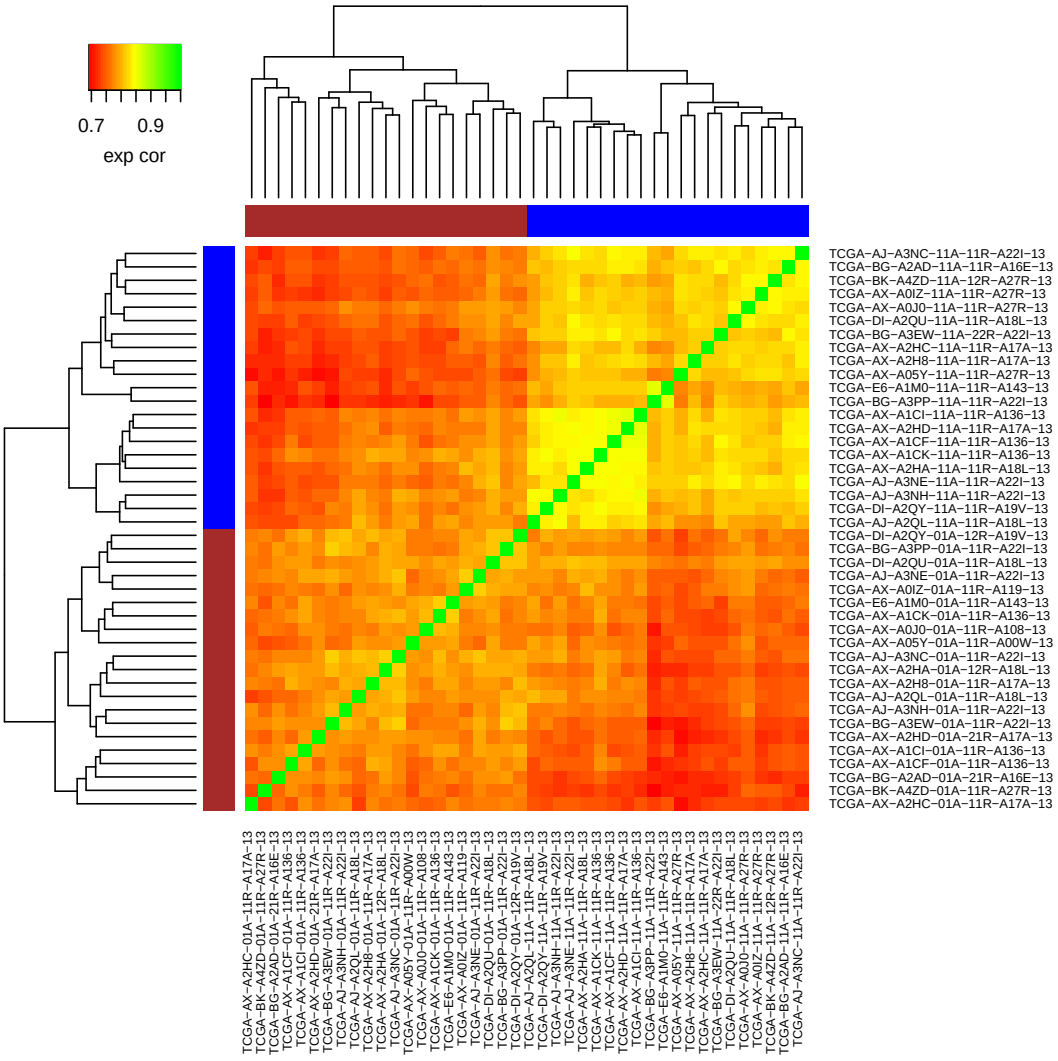
Apêndice A

O *heatmap* da correlação das amostras após o passo de normalização revela que todas as amostras analisadas neste trabalho possuíam correlação maior que 0,8.



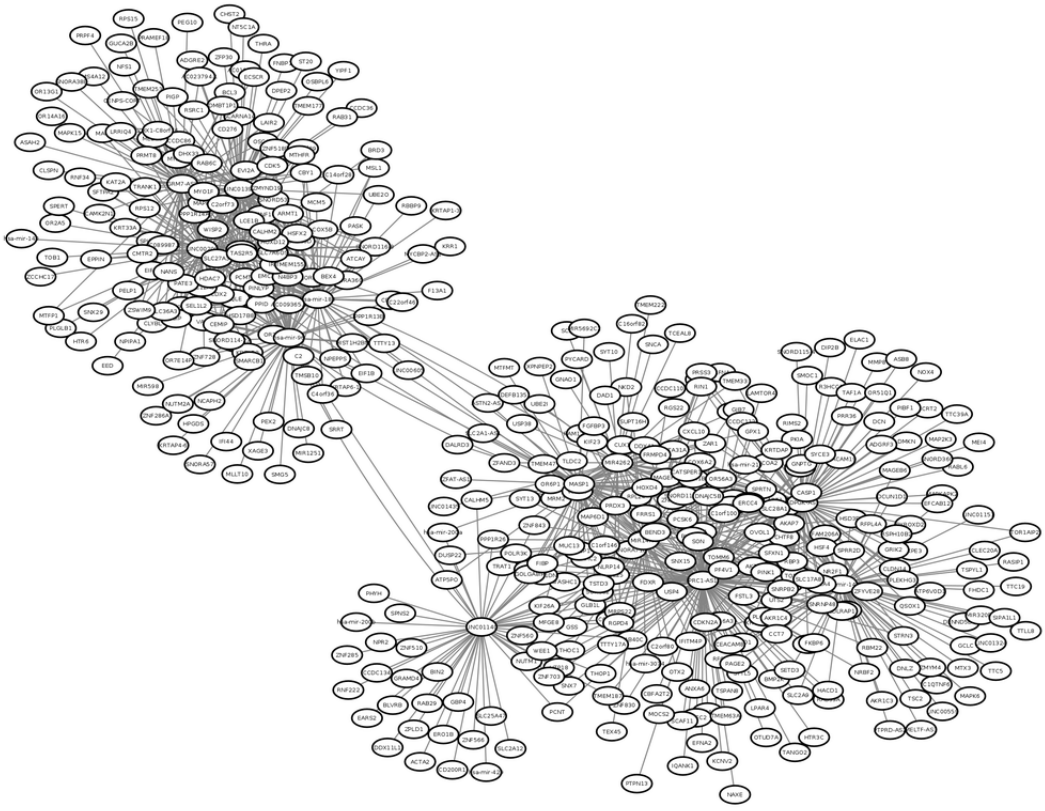
Apêndice B

O *heatmap* da correlação das amostras de miRNA após o passo de filtragem revela que todas as amostras analisadas neste trabalho possuíam correlação maior ou igual a 0,7.



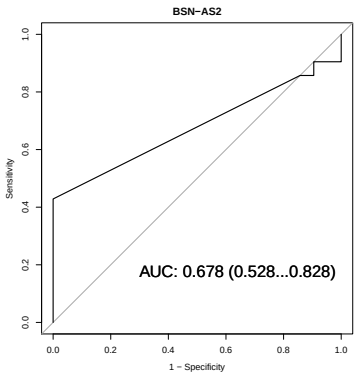
Apêndice C

A rede de correlações entre transcritos diferencialmente expressos com logFC acima de 2 e abaixo de -2, apresentou 971 correlações com valor de r abaixo de -0,8, o que dificulta a análise e visualização de possíveis reguladores e seus alvos.

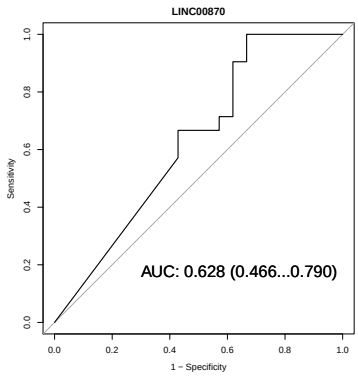


Apêndice D

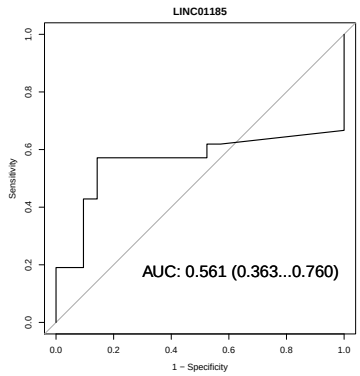
Curvas ROC dos lncRNAs mais diferencialmente expressos que não obtiveram AUC acima de 0,7. (a) O lncRNA antisenso *BSN-AS2* obteve uma AUC de 0,678. (b) O *LINC00870* obteve uma AUC de 0,628. (c) O *LINC01185* obteve uma AUC de 0,561. (d) O *LINC01269* obteve uma AUC de 0,535. (e) O lncRNA antisenso *RAD21-AS1* obteve uma AUC de 0,562. (f) O lncRNA *MIAT* obteve uma AUC de 0,437.



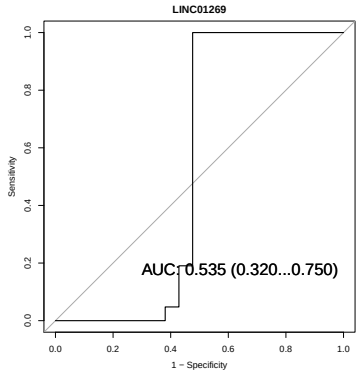
(a)



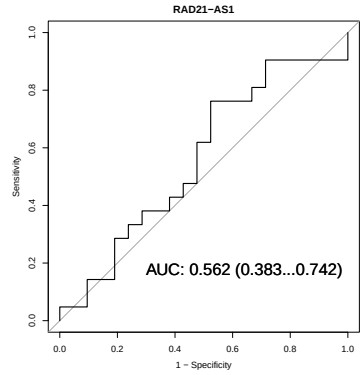
(b)



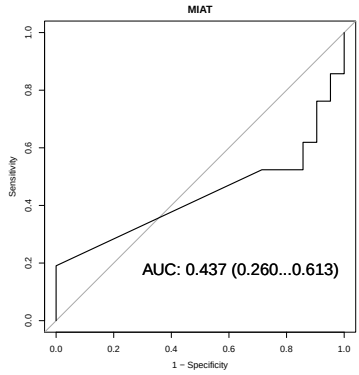
(c)



(d)



(e)



(f)