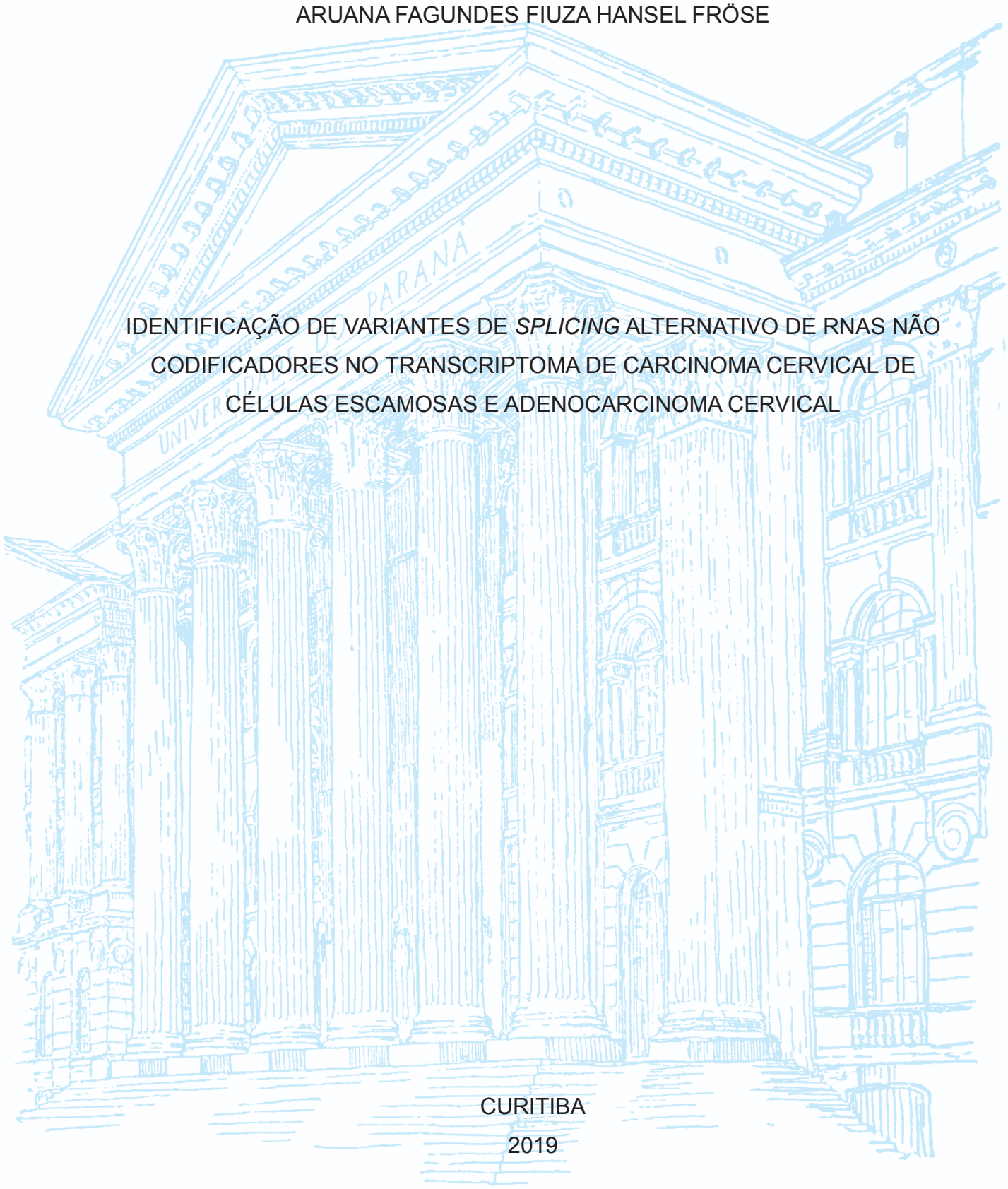
UNIVERSIDADE FEDERAL DO PARANÁ

ARUANA FAGUNDES FIUZA HANSEL FRÖSE



IDENTIFICAÇÃO DE VARIANTES DE *SPLICING* ALTERNATIVO DE RNAS NÃO CODIFICADORES NO TRANSCRIPTOMA DE CARCINOMA CERVICAL DE CÉLULAS ESCAMOSAS E ADENOCARCINOMA CERVICAL

CURITIBA

2019

ARUANA FAGUNDES FIUZA HANSEL FRÖSE

IDENTIFICAÇÃO DE VARIANTES DE *SPLICING* ALTERNATIVO DE RNAS NÃO CODIFICADORES NO TRANSCRIPTOMA DE CARCINOMA CERVICAL DE CÉLULAS ESCAMOSAS E ADENOCARCINOMA CERVICAL

Dissertação apresentada ao Curso de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica (SEPT) da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Bioinformática.

Orientador: Dr. Fabio Passetti

CURITIBA

2019

Catalogação na publicação
Sistema de Bibliotecas UFPR
Biblioteca de Educação Profissional e Tecnológica

Fröse, Aruana Fagundes Fiuza Hansel
Identificação de variantes de splicing alternativo de RNAs não codificadores no transcriptoma de carcinoma cervical de células escamosas e adenocarcinoma cervical – Curitiba, 2019.
116 p.: il.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor Educação Profissional e Tecnológica, Programa de Pós-Graduação em Bioinformática, 2019.
Orientador: Fabio Passetti

1. Coluna cervical - Câncer. 2. Análise de sequência de RNA. 3. Bioinformática. I. Passetti, Fábio. II. Título. III. Universidade Federal do Paraná.

Elaboração: Angela Pereira de Farias Mengatto - CRB 9/1002

MINISTÉRIO DA EDUCAÇÃO
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO BIOINFORMÁTICA - 40001016066P

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **ARUANA FAGUNDES FIUZA HANSEL FROSE** intitulada: **"Identificação de variantes de *splicing* alternativo de RNAs não codificadores no transcriptoma de carcinoma cervical de células escamosas e adenocarcinoma cervical"**, sob orientação do Prof. Dr. FABIO PASSETTI, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 18 de fevereiro de 2020

DR FÁBIO PASSETTI
Presidente/Instituto Carlos Chagas-FIOCRUZ
Programa de Pós-graduação em Bioinformática-UFPR

DR BRUNO DALLAGIOVANNA MUNIZ
Avaliador Externo/Instituto Carlos Chagas-FIOCRUZ

DR MAURO ANTONIO ALVES CASTRO
Avaliador Interno/Programa de Pós-graduação em Bioinformática-UFPR

Dedico esse trabalho a Deus, à minha amada família e ao meu amor.

AGRADECIMENTOS

Agradeço ao Programa de Pós-Graduação em Bioinformática da Universidade Federal do Paraná, e ao Instituto Carlos Chagas/FIOCRUZ Paraná, pela oportunidade concedida, receptividade, estrutura e apoio.

Agradeço aos órgãos de fomento CAPES, CNPQ, FIOCRUZ e FAPESP pelo suporte financeiro em todas as frentes necessárias para a realização desse trabalho.

Agradeço ao orientador Fabio pelos ensinamentos e orientação durante o desenvolvimento desse trabalho.

Agradeço à todas as colaboradoras e colaboradores dessa pesquisa, em especial a Drª Luísa Lina Villa por autorizar o uso dos dados inéditos em minha dissertação, e também ao Dr Helisson Faoro pelo suporte.

Agradeço à banca examinadora pelas sugestões e críticas construtivas para aprimorar esse trabalho.

Agradeço às amigas e amigos do laboratório de Bioinformática, por todas as boas conversas e o grande suporte nessa jornada científica, tornando a rotina mais alegre.

Agradeço à minha família, pela confiança, amor e apoio insubstituível, que me alegrou e motivou por todos os momentos da minha vida.

Agradeço ao meu melhor amigo e esposo, pelo companheirismo e pela motivação para seguir meus sonhos e a vencer os desafios da minha trajetória acadêmica.

E mais importante de todos, agradeço a Deus pela Sua presença em minha vida, por toda a força e coragem que me foi dada, e por todas as inestimáveis bençãos concedidas, principalmente nessa mais recente aventura. Agradeço pelo Seu filho Jesus Cristo, por ser meu Senhor e Salvador, melhor amigo e confidente, meu maior exemplo de vida, amor e compaixão.

“I did not want to just know *names* of things. I remember really wanting to know *how* it all worked.”

(Elizabeth Blackburn)

RESUMO

A porção inicial do útero que realiza a comunicação com o canal vaginal é chamada de cérvice uterina. Nessa região pode-se desenvolver o câncer cervical, também conhecido popularmente como câncer de colo de útero. Atualmente, esse tipo de câncer é o terceiro mais incidente em mulheres brasileiras, inclusive até mais que câncer de mama dependendo da região do país. Dos tipos histológicos de câncer cervical, o mais comum é o carcinoma cervical de células escamosas (SCC), que acomete cerca de 70% dos casos. Há ainda o adenocarcinoma (ADC), tipo histológico mais raro que acomete cerca de 20% dos casos, e outros tumores mistos compõem o restante dos diagnósticos. Apesar da distinção de tipos histológicos tumorais através do diagnóstico imuno-histoquímico, o tratamento administrado para ambos SCC e ADC é o mesmo. Isso compromete a resposta e remissão dos tumores ADC, uma vez que esse tipo apresenta o pior prognóstico. Para solucionar esse problema, é importante buscar por novos candidatos a biomarcadores dos subtipos de câncer cervical. Dessa forma será possível aprimorar a caracterização e distinção molecular dos subtipos para complementar seus diagnósticos e propor novos alvos farmacológicos. Transcritos de RNA derivados do processamento de *splicing* alternativo são bons candidatos a biomarcadores, pois a maioria dos genes humanos sofre esse processamento pós-transcricional. Além disso, já foram encontradas variantes alternativas diferencialmente expressas em neoplasias cervicais. Visto que a maior parte do transcriptoma humano é composto por genes não-codificadores, nós propomos que transcritos variantes de *splicing* alternativo oriundos de genes de RNAs não codificadores (ncRNA) podem contribuir na construção do perfil de expressão dos subtipos de câncer cervical. Por isso, buscamos novas variantes de *splicing* alternativo de ncRNAs diferencialmente expressas em dados oriundos do sequenciamento de alta-vazão do RNA de amostras de ADC e SCC. Utilizamos o software CLASS2 para identificar computacionalmente as variantes de *splicing* alternativo de ncRNAs em múltiplas amostras de câncer cervical, comparando duas abordagens de identificação de novos transcritos e três parâmetros diferentes. Através dos nossos métodos propostos, encontramos mais de 30 transcritos potencialmente novos não anotados nas referências do Ensembl e RefSeq, diferencialmente expressos entre ADC e SCC. Além disso, mostramos que identificar variantes de *splicing* alternativo nas amostras individualmente e então concatenar as anotações finais permite encontrar mais transcritos potencialmente novos. Após conferir manualmente as anotações personalizadas, encontramos três transcritos novos que são candidatos a biomarcadores positivos de SCC e negativos de ADC, além de dois transcritos novos bons candidatos a biomarcadores positivos de ADC e negativos de SCC. Ademais, encontramos genes já anotados que também são bons candidatos a biomarcadores de ADC e SCC. Como perspectivas futuras, propomos repetir nossa metodologia com um maior número amostral de tumores SCC e ADC principalmente, derivados tanto de bancos de dados globais quanto amostras de pacientes brasileiras.

Palavras-chave: Tipos de Câncer cervical. RNA-seq. *Splicing* Alternativo. CLASS2.

ABSTRACT

The uterine cervix is the initial portion of the uterus that communicates this organ with the vaginal canal. In this region, cervical cancer can develop. Currently, this type of cancer is the third most incident in Brazilian woman, even more than breast cancer in some regions of the country. Out of the histological types of cervical cancer, the most common is the squamous cell cervical carcinoma (SCC) that affects about 70% of the cases. There is also cervical adenocarcinoma (ADC), the rarer histological type that affects about 20% of the cases, and the rest of the diagnostics are composed of mixed tumours. Even though there is subtype distinction through immunohistochemical diagnostic, the treatment for both ADC and SCC is the same. This compromises response and remission of ADC tumours, since this histological type shows the worse prognosis. To solve this issue, it is important to seek for new biomarker candidates for ADC and SCC. In this way it will be possible to improve the characterization and molecular distinction of the subtypes, in order to complement their diagnostics and propose new pharmacological targets. RNA transcripts derived from alternative splicing are good biomarker candidates, because the majority of human genes suffer this post-transcriptional process. In addition, alternative splice variants were found differentially expressed in cervical neoplasias. Since the majority of the human transcriptome is composed of non-coding genes, we propose that alternative splice variants derived from non-coding RNA (ncRNA) genes can contribute to the expression profile of cervical cancer histological subtypes. Therefore, we seek new alternative splice variants of ncRNA differentially expressed in data from RNA sequencing (RNA-seq) derived from ADC and SCC samples. We used the software CLASS2 to computationally identify the alternative splice variants of ncRNAs in many cervical cancer samples, and compared two approaches and three parameters to identify new splice variants. Through our proposed methods, we have found more than 30 potentially new transcripts, unannotated in Ensembl and RefSeq references, differentially expressed between ADC and SCC. Furthermore, we showed that identifying alternative splice variants in the individual samples prior to annotation merging is best to find more potentially new transcripts. After thorough analysis of our personalized annotations, we found three new transcripts that are potentially positive biomarkers candidates for SCC and negative for ADC, and also two new transcripts that are good positive biomarker candidates for ADC and negative for SCC. In addition, we also found annotated genes that are good biomarker candidates for ADC and SCC. As future prospects, we propose to repeat our methods with a higher number of samples of SCC and most important ADC, of which data can derive from global databases or more samples from Brazilian women.

Keywords: Cervical Cancer subtypes. RNA-seq. Alternative Splicing. CLASS2.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

ADC	Adenocarcinoma cervical
asRNA	RNAs anti-senso
circRNA	RNAs circulares
circRNAs	RNAs circulares
HPV	Papiloma vírus humano
HSIL	Lesão intraepitelial escamosa de alto grau
INCA	Instituto Nacional do Câncer
lincRNA	lncRNA intergênicos
lncRNA	RNA longo não-codificador
logFC	Log_2 Fold Change
LSIL	Lesão intraepitelial escamosa de baixo grau
miRNA	microRNAs
mRNA	Ácido ribonucleico mensageiro
ncRNA	RNAs não-codificadores
PCR	Reação em cadeia da polimerase
RNA-seq	Sequenciamento de alta vazão do RNA
SCC	Carcinoma cervical de células escamosas

SUMÁRIO

1 INTRODUÇÃO

A porção inicial do útero que realiza a comunicação uterina com o canal vaginal é chamada de colo uterino ou cérvice. Nessa região, a proliferação incessante de células anormais cervicais dá origem ao câncer cervical, sendo este um dos que mais acometem as mulheres brasileiras atualmente, somando mais de 16 mil novos casos em 2018 (Instituto Nacional do Câncer - INCA, 2018). Há diversos tipos histológicos de câncer cervical, mas os principais são o carcinoma cervical de células escamosas (SCC) e o adenocarcinoma cervical (ADC) (Cancer Research UK, 2017). A maioria dos tumores cervicais são diagnosticados como SCC, compondo cerca de 75% a 90% dos novos casos (GIEN; BEAUCHEMIN; THOMAS, 2010; WILLIAMS et al., 2015). Apesar da menor incidência de ADC, em torno de 20 a 25% dos casos, esse subtipo apresenta um pior prognóstico que o anterior, além da possibilidade de recorrência e/ou metástase (GADDUCCI; GUERRIERI; COSIO, 2019; WILLIAMS et al., 2015; ZHOU et al., 2018). Entretanto, apesar das importantes diferenças histopatológicas e clínicas entre os subtipos, os tratamentos administrados são os mesmos, o que compromete a resposta e remissão dos tumores ADC principalmente (WILLIAMS et al., 2015). O diagnóstico determinante para diferenciar os subtipos é majoritariamente imuno-histoquímico (BUZA; HUI, 2017; KASPAR; CRUM, 2015), carecendo de elementos moleculares que complementem a identificação do subtipo histológico do câncer de colo de útero.

Uma maneira de abordar essa questão é buscar por novos candidatos a biomarcadores que aprimorem a caracterização e diferenciação dos subtipos de câncer cervical, complementando o diagnóstico. Além disso, bons biomarcadores poderiam se tornar alvos farmacológicos para o tratamento específico de ADC. Transcritos de RNA são bons candidatos para auxiliar nessa diferenciação de subtipos uma vez que diferentes RNAs já foram identificados com expressão diferenciada em tumores e neoplasias cervicais. Exemplos de transcritos são RNAs longos não-codificadores (lncRNA), RNAs circulares (circRNA), microRNAs (miRNAs), receptores diversos e fatores de transcrição (KORI; YALCIN ARGA, 2018; XIAOGUANG et al., 2017), além de RNAs variantes de *splicing* alternativo (SONG et al., 2007; WOERNER et al., 1995; XIAOGUANG et al., 2017).

Nesse contexto, o programa CLASS2 (SONG; SABUNCIYAN; FLOREA, 2016) identifica variantes de *splicing* alternativo em dados de transcriptoma oriundos de sequenciamento de alta vazão de RNA (RNA-seq). Portanto, nesse trabalho buscamos identificar variantes de *splicing* alternativo em dados inéditos de RNA-seq de ADC e SCC sequenciados pelo nosso grupo de pesquisa, com potencial de aprimorar a caracterização e classificação desses subtipos histológicos de câncer cervical.

1.1 JUSTIFICATIVA

Apesar das diferenças histopatológicas e clínicas, tumores cervicais classificados como ADC e SCC recebem o mesmo tratamento, dificultando a remissão dos tumores ADC que apresentam pior prognóstico e possibilidade de recorrência e/ou metástase. A descoberta de novos candidatos a biomarcadores que diferenciem e caracterizem esses dois subtipos complementa o diagnóstico imuno-histoquímico, além de permitir identificar novos alvos especialmente para o tratamento de ADC. Considerando isso, bons candidatos a biomarcadores são variantes de *splicing* alternativo diferencialmente expressas no transcriptoma de ADC e SCC, que podem ser investigadas na análise do sequenciamento de RNA tumoral, através de uma abordagem de bioinformática.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Identificar transcritos variantes de *splicing* alternativo, de genes não codificadores de proteínas, diferencialmente expressos no transcriptoma de câncer cervical dos tipos ADC e SCC.

1.2.2 Objetivos específicos

- Comparar a identificação de novos transcritos variantes de *splicing* de genes não codificadores de proteínas pelo programa CLASS2 utilizando duas abordagens e três parâmetros de limiar distintos.
- Identificar a expressão diferencial de variantes de *splicing* de transcritos derivados de genes não codificadores de proteína entre os tipos histológicos SCC e ADC de câncer cervical.

1.3 REVISÃO DE LITERATURA

1.3.1 Câncer cervical

1.3.1.1 Incidência e mortalidade no país

Câncer compreende uma variedade de doenças, todas caracterizadas pela proliferação incontrolável de células anormais que formam massas celulares chamadas de tumores. O Instituto Nacional do Câncer (INCA, 2018) estimou mais de 580 mil novos casos de câncer no Brasil para cada ano do biênio 2018-2019, totalizando um risco estimado de 15 casos a cada 100 mil mulheres. Desses novos casos, o terceiro mais recorrente é o câncer cervical, com cerca de 16 mil novos casos de tumores primários localizados no colo do útero das mulheres brasileiras. Esse tipo tumoral está atrás apenas do câncer de cólon e reto, com cerca de 18 mil novos casos, e do câncer de mama, liderando com mais de 59 mil novos casos em 2018 (INCA, 2018). De acordo com o mesmo instituto, a mortalidade proveniente do câncer de colo de útero é a quarta mais comum nas mulheres do país, chegando a mais de 6 mil óbitos em 2017 (INCA, 2018).

Conforme a estimativa de 2018-2019, no Brasil há variação da incidência do câncer cervical entre as brasileiras de diferentes regiões do país, com o maior número de casos ocorrendo nas regiões centro-oeste, nordeste e norte (Figura 1, Quadro 1) (INCA, 2018). De acordo com os dados do INCA de 2018, nas mulheres da região

centro-oeste o câncer cervical é o segundo tipo tumoral mais incidente em 10,3% dos casos, atrás do câncer de mama com 29% dos casos. A mesma ordem de incidência se mantém nas mulheres nordestinas, em que o câncer cervical é o segundo tipo mais incidente em 13,4% casos, seguido pelo câncer de mama com 26,3% dos casos. Em contrapartida, esse padrão se inverte nas mulheres da região norte, onde o câncer de colo de útero lidera como tipo de câncer mais incidente na região e no Brasil, correspondendo a 24,8% dos casos, enquanto que o câncer de mama segue em segundo lugar com 18,6% dos casos. Nas mulheres da região sudeste, a incidência de câncer de colo de útero ocupa a quarta posição dos mais incidentes, com 4,7%, atrás de câncer de traqueia, brônquio e pulmão com 6%, cólon e reto com 11,2% e liderado por câncer de mama com 32,6%. Nas mulheres da região sul, a ordem de mais incidência segue semelhante às da região sudeste, com incidências de câncer cervical de 5,5%, atrás de câncer de traqueia, brônquio e pulmão com 8,1%, câncer de cólon e reto com 9% e liderado por câncer de mama com 28,7% (Quadro 1).

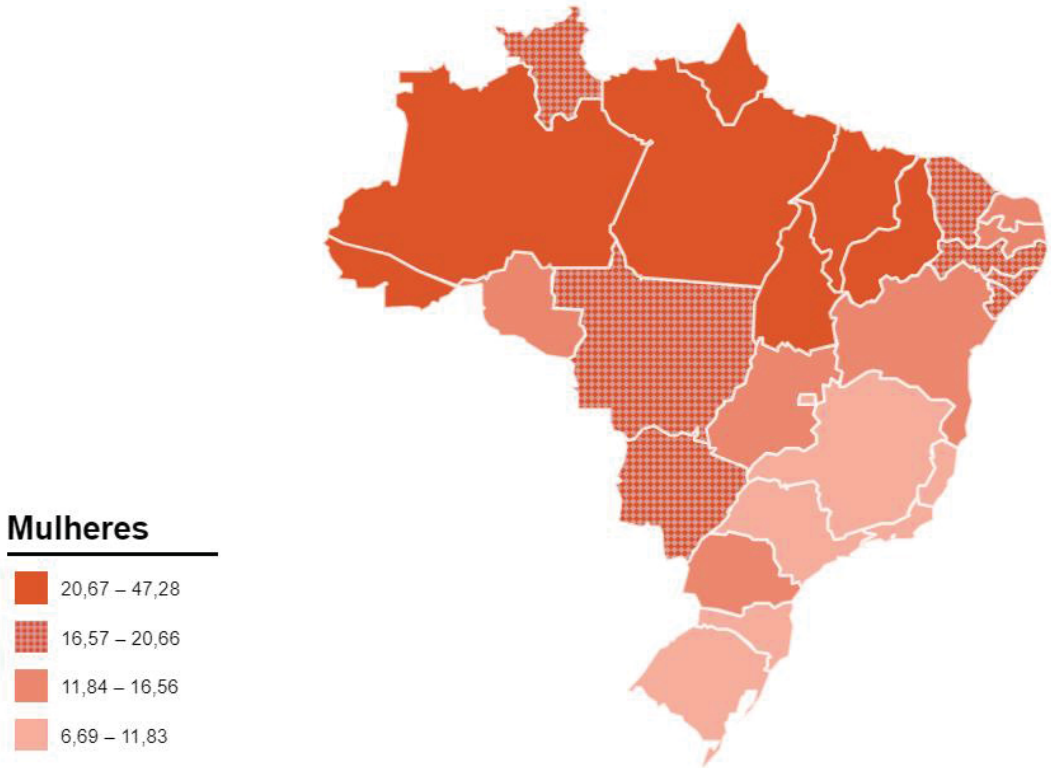


Figura 1: Estimativa de incidência de câncer cervical a cada 100 mil mulheres, para o ano de 2018, conforme as regiões do Brasil.

Fonte: INCA, 2018.

Quadro 1: Incidência de câncer cervical proporcional aos tipos de câncer mais incidentes em mulheres no Brasil em 2018, segundo o INCA.

Região	Incidência de câncer cervical	Ranking na região
Norte	24,8%	1
Nordeste	13,4%	2
Centro-oeste	10,3%	2
Sul	5,5%	4
Sudeste	4,7%	4

Fonte: INCA, 2018.

1.3.1.2 Tipos histológicos de câncer cervical

A porção inicial do útero que se estende para dentro do canal vaginal é denominada de colo uterino, ou cérvice. Essa região faz a comunicação entre o útero e o canal vaginal e é composta por diferentes tipos celulares: as células epiteliais ou epitélio escamoso, que cobrem externamente o colo uterino, demarcando a região escamosa ectocérvice; e as células glandulares produtoras de muco, ou epitélio colunar, que revestem internamente o colo na endocérvice (Figura 2). Esses tipos celulares denominam os principais subtipos de câncer cervical de acordo com a identidade da célula anormal que compõe a massa tumoral. Assim, originado da ectocérvice há o carcinoma de células escamosas (SCC) e originado da endocérvice há o adenocarcinoma cervical (ADC). Há ainda tumores mistos compostos por mais de um tipo celular, denominados como carcinoma adenoescamoso cervical (Cancer Research UK, 2017).

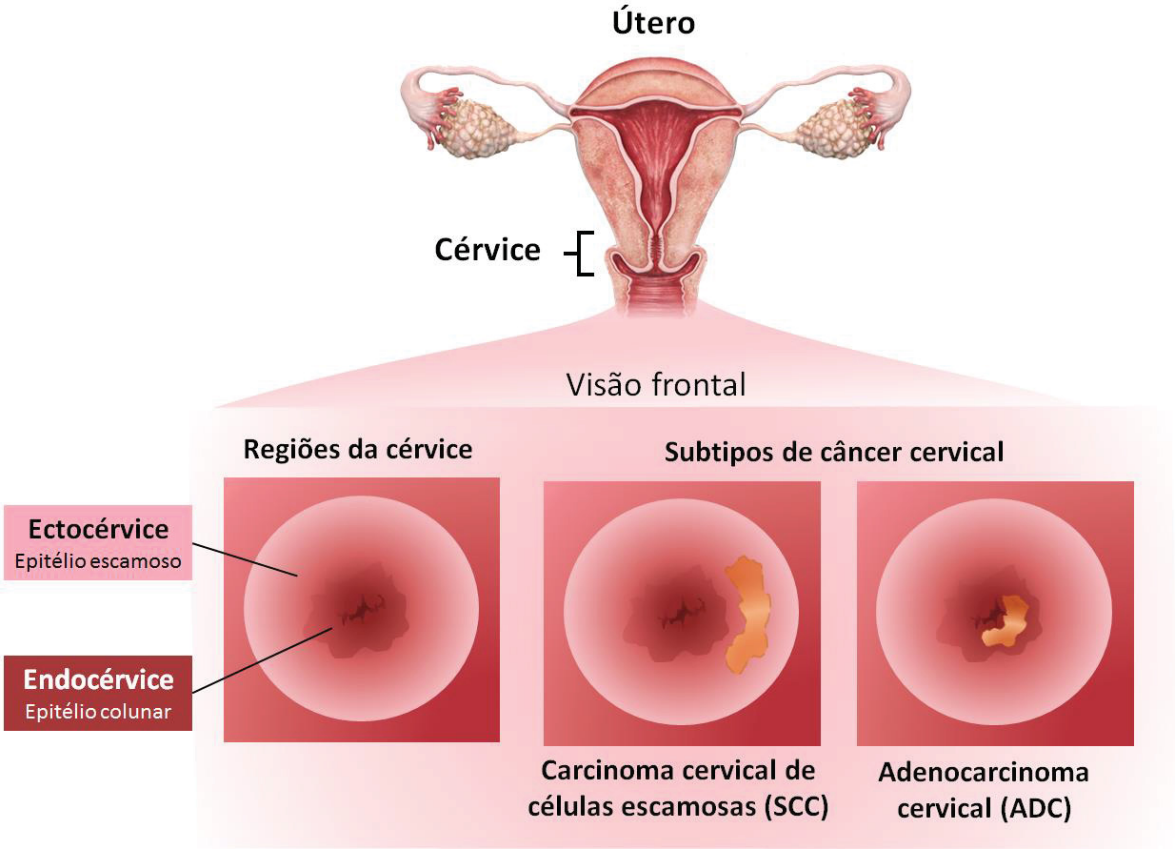


Figura 2: Representação da região cervical uterina e os tipos histológicos ADC e SCC de câncer cervical abordados nesse trabalho.

Fonte: Ilustração de útero, ovários e canal vaginal por Oleg e Andrew, disponível em everviz.artstation.com. Tipos histológicos de câncer cervical ilustrados pela autora (2020).

A maior parte dos tumores cervicais são identificados como SCC, compondo cerca de 75% até 90% dos casos (GIEN; BEAUCHEMIN; THOMAS, 2010; WILLIAMS et al., 2015). A incidência de ADC está em torno de 20 a 25% dos casos, mas apresenta um pior prognóstico e com possibilidade de recorrência ou metástase (GADDUCCI; GUERRIERI; COSIO, 2019; WILLIAMS et al., 2015; ZHOU et al., 2018). Entretanto, apesar das diferenças significativas entre ADC e SCC, ambos os tipos histológicos recebem o mesmo tratamento, em que os tumores ADC apresentam pior resposta às estratégias (SHIMADA et al., 2013; WILLIAMS et al., 2015). Outra característica importante é o fato de que, enquanto os diagnósticos são feitos com base principalmente em técnicas imuno-histoquímicas (BUZA; HUI, 2017; KASPAR; CRUM, 2015), a existência de diferenças moleculares entre os tumores SCC e ADC dificulta ainda mais o diagnóstico preciso e tratamento eficaz (GADDUCCI; GUERRIERI; COSIO, 2019; RONNETT, 2016; WILLIAMS et al., 2015).

1.3.1.3 HPV como fator de risco

O fator de risco mais evidente para o desenvolvimento de câncer de colo de útero em geral é a infecção pelo papilomavirus humano (HPV, do inglês “human papiloma virus”), apontando-o como agente etiológico dessa doença. Outras características comportamentais da mulher também pode predispor o colo uterino para o desenvolvimento de lesões e neoplasias cervicais (CASTELLSAGUÉ; MUÑOZ, 2003). Ao ser sexualmente transmissível, a infecção pelo HPV também está associada ao comportamento sexual da mulher, como a falta do uso de preservativos, o início precoce de relações sexuais, a maior quantidade de parceiros e a infecção por outros patógenos (CASTELLSAGUÉ; MUÑOZ, 2003; KIM et al., 2012). Durante a gestação e parto, mudanças hormonais e exposição do colo uterino ao vírus podem ocorrer, o que também aponta a alta paridade para maior risco de desenvolver a doença (CASTELLSAGUÉ; MUÑOZ, 2003; KIM et al., 2012). Além disso, o tabagismo também é um cofator significativo para o desenvolvimento de neoplasias em mulheres infectadas com HPV, além de maior risco de neoplasias cervicais com a longa administração de contracepção hormonal (CASTELLSAGUÉ; MUÑOZ, 2003; KIM et al., 2012; KJELLBERG et al., 2000).

A infecção viral pelo HPV pode levar à lesões de baixo risco (LSIL, do inglês “Low-grade Squamous Intraepithelial Lesion”) e de alto risco (HSIL, do inglês “High-grade Squamous Intraepithelial Lesion”) na região do colo uterino, e a subsequente progressão para neoplasias cervicais (National Cancer Institute, 2019). As lesões cervicais causadas pela infecção viral possivelmente podem regredir espontaneamente (SCHLECHT et al., 2003). Entretanto, as lesões de alto risco possuem grandes chances de progredirem para um tumor maligno cervical (MCCREDIE et al., 2008).

Dentre os mais de 200 tipos diferentes de HPV já documentados, cerca de 14 tipos de HPV são considerados de alto risco oncogênico para formação de lesões na cérvice segundo a Organização Mundial da Saúde (“World Health Organization”, 2019). Os tipos de HPV de baixo risco, como os tipos 66 e 11, usualmente ocorrem em lesões benignas como verrugas. Já os tipos de alto risco, os tipos 16 e 18 de HPV,

são os causadores mais evidentes da maioria dos tumores e lesões cervicais, somando cerca de 70% dos casos (GHITTONI et al., 2015; “World Health Organization”, 2019). Somente o HPV16 é responsável por 50% dos tumores cervicais mundiais (GHITTONI et al., 2015; “World Health Organization”, 2019).

No intuito de identificar marcadores moleculares para identificação de neoplasias cervicais e associação com infecção por HPV, alguns biomarcadores foram propostos para complementar o diagnóstico histopatológico, como a proteína humana endógena p16 (símbolo gênico CDKN2A) (CUSCHIERI; WENTZENSEN, 2008) e as proteínas virais E6 e E7. Todos esses marcadores são encontrados em HSIL ao invés de LSIL, auxiliando na diferenciação e diagnóstico das neoplasias cervicais (CUSCHIERI; WENTZENSEN, 2008; GHITTONI et al., 2015; IVANOVA et al., 2007; WANG et al., 2005; WENTZENSEN; VON KNEBEL DOEBERITZ, 2007).

Os genes virais E6 e E7 produzem proteínas necessárias para a replicação do HPV, sendo expressos durante o ciclo de vida viral. A alta expressão dessas proteínas pode influenciar no controle e regulação do ciclo celular humano e levar ao desenvolvimento de lesões e neoplasias cervicais (CUSCHIERI; WENTZENSEN, 2008). A proteína E6 pode-se ligar à proteína humana supressora de tumor p53. Ao degradá-la, há prejuízos no reparo do DNA e manutenção do genoma humano, visto que ela é um fator de transcrição essencial para regular a resposta aos danos no DNA (GHITTONI et al., 2015).

Já a proteína viral E7 interfere no controle correto do ciclo celular da célula infectada, ao mimetizar a sua regulação inata. Normalmente, o ciclo celular das células humanas é regulado pela fosforilação de pRb (proteína retinoblastoma), que por sua vez é regulada pela proteína p16, um fator que bloqueia a fosforilação de pRb (CUSCHIERI; WENTZENSEN, 2008). Durante uma infecção por HPV, a proteína viral E7 se liga à pRb humana, promove sua degradação e consequentemente a célula progride para a fase S do ciclo celular, de duplicação do DNA, independente da regulação usual (GHITTONI et al., 2015). Assim, a constante ativação do ciclo celular durante a infecção viral é causada pela ausência de fosforilação de pRb pela p16, a qual é consequentemente superexpressa e acumulada nas células em processos neoplásicos (GHITTONI et al., 2015). Dessa forma, a identificação da alta expressão diferencial da proteína humana p16 está associada ao diagnóstico de neoplasia

cervical (CUSCHIERI; WENTZENSEN, 2008; IVANOVA et al., 2007), considerada um biomarcador pela correlação entre a sua expressão com a de mRNA de proteínas virais E6 e E7 (CUSCHIERI; WENTZENSEN, 2008).

Apesar da associação significativa com câncer cervical, os biomarcadores p16 e as proteínas virais E6 e E7 podem estar atribuídos tanto a SCC quando ADC, o que não são parâmetros interessantes para diferenciar esses tipos histológicos de câncer cervical (CUSCHIERI; WENTZENSEN, 2008; GHITTONI et al., 2015).

1.3.2 Ciências ômicas

A área das ciências ômicas é multidisciplinar que tem como objeto de estudo os sistemas biológicos, integrando tais conhecimentos através da computação científica. O conjunto de dados biológicos e suas interações compõem os agrupamentos conhecidos como: genoma, o conjunto haploide de cromossomos da espécie; transcriptoma, o conjunto de transcritos expressos no objeto de estudo; proteoma, o conjunto de proteínas expressas no objeto de estudo; metaboloma, o conjunto de metabólitos produzidos ou modificados pelo organismo; interatoma, o conjunto de interações de moléculas biológicas; metaômas, o conjunto das comunidades de organismos encontrados em uma determinada amostra, para subsequente análise de biomoléculas específicas, como metagenômica, metaproteômica e metatranscriptômica.

Dependendo do conjunto biológico a ser estudado, as ciências ômicas são divididas em subdisciplinas ômicas, dentre as quais as principais são: genômica, analisando o genoma; proteômica analisando o proteoma; transcriptômica analisando o transcriptoma; metabolômica analisando o metaboloma; interatômica analisando o interatoma; e metaômicas analisando os metaômas (ESPINDOLA et al., 2010). Também é possível a integração entre áreas, compondo a Biologia Sistêmica ou Biologia de Sistemas, que trata das interações entre componentes de um sistema biológico (ESPINDOLA et al., 2010).

O estudo das ciências ômicas também é interdisciplinar, utilizando da capacidade computacional e do método estatístico para investigar esses dados.

Através da Bioestatística, Computação Científica e da Bioinformática, a pesquisa sobre cada subárea ômica utiliza ferramentas e experimentos apropriados para os objetos de estudo, com possibilidade de integrar as técnicas e obter a visão mais sistêmica (ESPINDOLA et al., 2010). Com o avanço tecnológico, as áreas da Computação Científica e da Bioinformática também cresceram, aprimorando recursos ferramentais para poder lidar com a grande quantidade de dados gerados pelas ciências ômicas (ESPINDOLA et al., 2010).

Os dados gerados podem ser armazenados em bancos de dados biológicos, permitindo realizar análises em larga escala com centenas e até milhares de amostras. Em alguns casos, os mesmos dados podem ser usados por diferentes grupos de pesquisa, aumentando o número amostral e consequentemente enriquecendo o trabalho (EBI EMBL, 2020; ZOU et al., 2015).

1.3.2.1 Transcriptômica para a pesquisa biomédica oncológica

A transcriptômica permite catalogar os diferentes RNAs presentes no objeto de estudo, determinar sua estrutura e padrão de processamento, e também quantificar o nível de expressão dos transcritos conforme a condição (WANG; GERSTEIN; SNYDER, 2009). A aplicação da área da transcriptômica e de outras ciências ômicas através da bioinformática têm ganhado popularidade em várias facetas biomédicas, como na pesquisa oncológica e de outras diversas doenças (MANZONI et al., 2018).

O estudo do RNA tumoral através da transcriptômica é uma maneira de caracterizar e diferenciar os perfis de expressão de determinados tipos de câncer, sendo possível até identificar os transcritos virais que evidenciam a infecção por HPV, no caso de câncer cervical (GOEDERT et al., 2016; SCHMITT et al., 2010). Os transcritos que podem ser identificados nessa pesquisa são tanto RNAs codificadores como RNA mensageiro (mRNA), quanto RNAs não-codificadores (ncRNA) e enzimáticos, como microRNAs (miRNA), RNA ribossomal (rRNA), entre outros.

Além disso, podem-se descobrir novos candidatos a biomarcadores, genes ou proteínas diferencialmente expressos no câncer e específicos para a condição em que foram encontrados, com perspectiva de se tornarem alvos farmacológicos para futuras

terapias (ARONSON; FERNER, 2017; WENTZENSEN; VON KNEBEL DOEBERITZ, 2007).

1.3.2.2 Sequenciamento de RNA

A busca por esses biomarcadores no transcriptoma tumoral pode ser realizada através do sequenciamento de alta vazão do RNA, conhecido popularmente como RNA-seq.

Nessa técnica, inicia-se com um conjunto de RNAs, como RNA total, RNA parcial, RNA positivo para cauda de poliadeninas (poli-A), provenientes de um tecido, organismo ou célula (Figura 3). Esses RNAs são convertidos a fragmentos de DNA complementar (cDNA), através de adaptadores associados às suas extremidades. Esses cDNA podem ser posteriormente amplificados ou não.

Então esses fragmentos são sequenciados em equipamentos sequenciadores de alta vazão, podendo ocorrer de duas maneiras pré-determinadas: a partir de uma única extremidade do fragmento, conhecido como sequenciamento "*single-end*"; ou a partir das duas extremidades do fragmento, no sequenciamento conhecido como pareado ou "*paired-end*". O sequenciamento dos fragmentos gera um conjunto de dados denominados leituras, ou "*reads"*, sequências de tamanhos variáveis de cerca de 30 a 400 pares de base, dependendo da tecnologia utilizada para sequenciamento (WANG; GERSTEIN; SNYDER, 2009). Para tanto, muitas vezes o termo RNA-seq também é cunhado de sequenciamento de RNA de pequenas leituras, do inglês "*short reads RNA sequencing*" (WANG; GERSTEIN; SNYDER, 2009).

Agora em formato de dados computacionais, a fim de identificação das sequências, essas leituras podem ser alinhadas contra a sequência de um genoma de referência (WANG; GERSTEIN; SNYDER, 2009). O alinhamento é uma maneira de organizar as leituras, buscando similaridade de sequência conforme critério estabelecido e utilizando um genoma como referência de organização (WANG; GERSTEIN; SNYDER, 2009). Há dois tipos de alinhamento que podem ser feitos: alinhamento local e alinhamento global. O alinhamento local irá buscar pequenas regiões de alta similaridade entre as duas sequências alinhadas (WANG; GERSTEIN;

SNYDER, 2009). Já o alinhamento global considera toda a extensão das sequências para procurar o melhor alinhamento entre elas (WANG; GERSTEIN; SNYDER, 2009). Em ambos os casos, caso ocorra pequenas regiões sem similaridade entre regiões de alta similaridade, podem ser incorporados espaços em branco conhecidos como "*gap"* (WANG; GERSTEIN; SNYDER, 2009).

No caso de novas espécies ou organismos ainda pouco sequenciados em que não há genoma de referência para poder comparar no alinhamento, há a possibilidade de fazer uma organização das leituras sem utilizar uma referência para comparação, método conhecido como montagem *de novo* (WANG; GERSTEIN; SNYDER, 2009).

Dessa forma, são feitas anotações precisas dos transcritos encontrados na amostra sequenciada, com resolução por bases, e até identificando variações exclusivas da amostra ou da condição em comparação com a referência. Ademais, pode-se realizar cálculos estatísticos da quantidade de vezes em que o transcrito foi sequenciado, como na expressão diferencial, métrica utilizada para avaliar o nível de expressão gênica de maneira quantitativa (WANG; GERSTEIN; SNYDER, 2009).

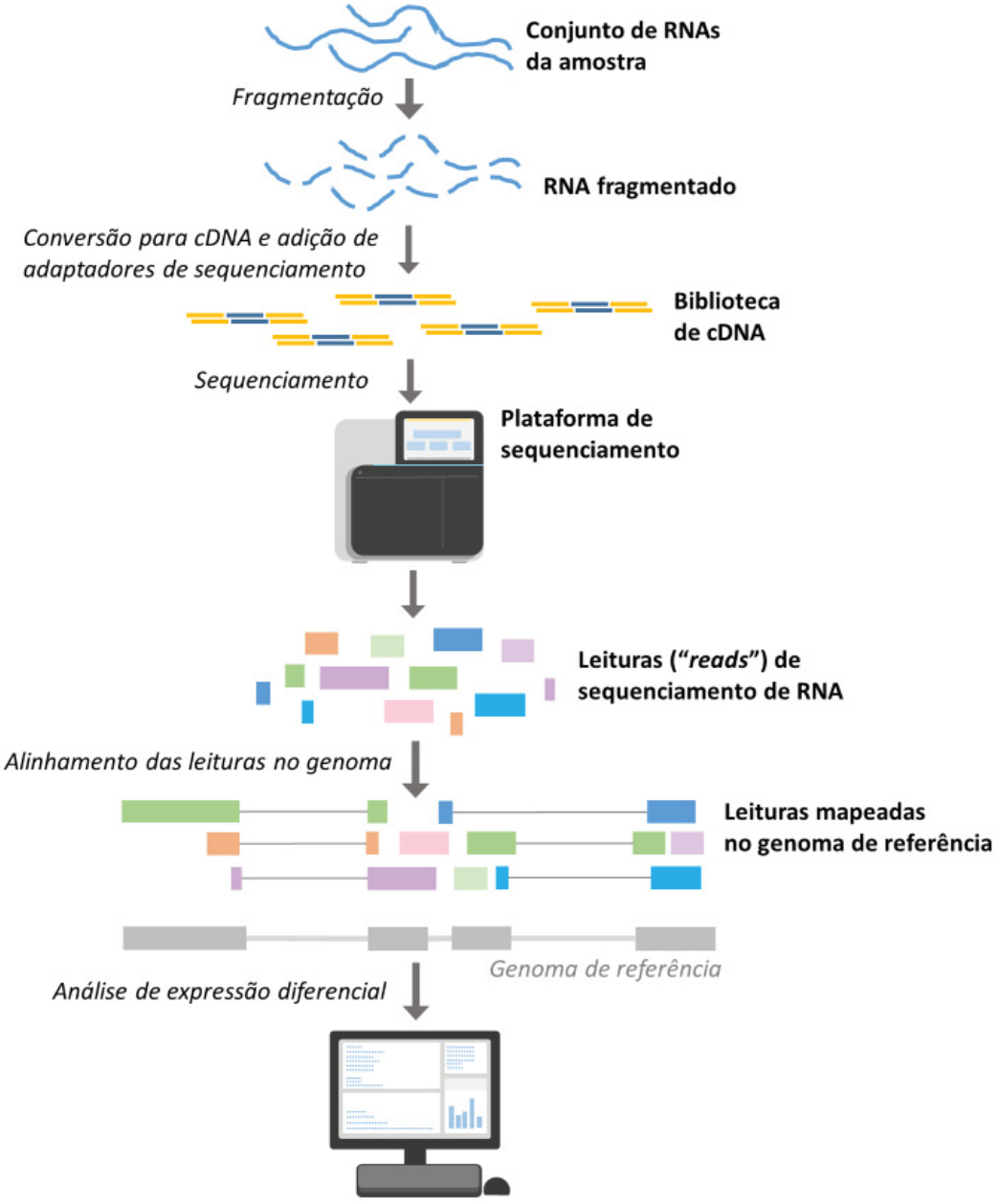


Figura 3: Protocolo padrão de sequenciamento de RNA (RNA-seq).

Fonte: Ilustração da autora (2020) conforme protocolo descrito por Wang e colaboradores (2009).

1.3.2.3 Transcriptoma do câncer cervical

Na literatura científica, há estudos que realizaram o sequenciamento de RNA de câncer cervical buscando traçar o seu perfil de expressão gênica e possivelmente encontrar novos candidatos a biomarcadores para suas condições. Foram encontrados genes diferencialmente expressos com perspectiva para serem

biomarcadores desse tipo tumoral, como a já mencionada p16 (ARONSON; FERNER, 2017; CUSCHIERI; WENTZENSEN, 2008; KLAES et al., 2001; QIN et al., 2019; VALENTI et al., 2017; WENTZENSEN; DOEBERITZ, 2007).

Porém, diferenciar os tipos histológicos através do perfil de expressão gênica se apresentou uma tarefa desafiante ainda com grandes meta-análises (KORI; YALCIN ARGA, 2018; SCHMITT et al., 2010). O perfil de expressão diferencial de miRNAs em câncer cervical e tecido saudável já foi abordado anteriormente por alguns trabalhos (LEE et al., 2008; WITTEN et al., 2010). Witten e colaboradores em 2010 mostraram a diferença do perfil de expressão de miRNAs entre câncer cervical e tecido normal, contudo obtiveram dificuldades para separar os tipos histológicos ADC e SCC utilizando a mesma técnica (WITTEN et al., 2010). O grupo de pesquisa do Atlas Genômico de Câncer (TCGA, do inglês “The Cancer Genome Atlas”) também evidenciou essa heterogeneidade molecular do câncer cervical, no qual subgrupos moleculares podem discordar do diagnóstico feito histologicamente (BURK et al., 2017).

A melhor caracterização do perfil de expressão dos subtipos de câncer cervical dá perspectivas principalmente para as pacientes de tumores ADC. O subtipo ADC apresenta um pior prognóstico e detém possibilidade de recorrência e/ou metástase (GADDUCCI; GUERRIERI; COSIO, 2019; WILLIAMS et al., 2015; ZHOU et al., 2018). Apesar das diferenças clínicas e histopatológicas entre ADC e SCC, ambos recebem os mesmos tratamentos, comprometendo a resposta de tumores ADC (WILLIAMS et al., 2015). Atualmente, o diagnóstico dos subtipos tumorais de câncer cervical é principalmente imuno-histoquímico (BUZA; HUI, 2017; KASPAR; CRUM, 2015). Portanto ainda faltam análises moleculares complementares para aumentar a precisão do diagnóstico da doença.

1.3.3 Processamento de *Splicing* em RNAs

O transcriptoma humano é composto por mais de 80 mil transcritos codificadores para proteína, gerando um número estimado de 250 mil até 1 milhão de proteínas (DE KLERK; ’T HOEN, 2015). Porém, esses transcritos e proteínas derivam de cerca de 20 mil genes humanos, o que indica uma imensa regulação transcricional,

pós-transcricional e traducional (DE KLERK; ’T HOEN, 2015). O avanço da transcriptômica e outras tecnologias ômicas permitiu à ciência entender mais a fundo como ocorrem esses mecanismos regulatórios, além de descobrir uma grande variedade de transcritos no transcriptoma de seres vivos.

O dogma central da biologia molecular expõe o fluxo de informações do código genético que pode ocorrer nos seres vivos, e foi inicialmente proposto por Francis Crick em 1970. O ácido nucleico, DNA, que armazena informação genética pode se duplicar no processo de Replicação de DNA (FRANCIS CRICK, 1970). O DNA também pode ser molde para a transcrição de uma molécula de RNA, que por sua vez poderá ser traduzida para uma proteína (FRANCIS CRICK, 1970). Sabe-se que há seres com enzimas capazes de realizar a transcrição reversa do RNA para DNA, e há casos de RNAs que também podem ser replicados.

A expressão gênica de genes que codificam proteínas se inicia com a transcrição de um gene no DNA para uma molécula de RNA. Em um primeiro momento, há a formação de um RNA mensageiro (mRNA) precursor, que posteriormente será processado tornando-se um mRNA maduro pronto para ser traduzido para proteína. Dentre esses processamentos pós-transcricionais, há o evento de *splicing* (Figura 4). Em células eucarióticas, o processo de *splicing* é definido pela excisão de regiões do transcrito que não serão posteriormente traduzidas, conhecidas como íntrons. Com essa remoção, permanecem apenas as regiões que efetivamente poderão ser traduzidas para proteína (sequência codificadora, CDS, do inglês “*coding sequencing*”), composta pelos exons (WANG; GERSTEIN; SNYDER, 2009). Essa remoção de íntrons e união de exons é orquestrada por um complexo de mais de 300 proteínas e cinco RNAs não codificadores (ncRNA) chamado de *spliceossomo* (URBANSKI; LECLAIR; ANCZUKÓW, 2018). O *spliceossomo* realiza a excisão de íntrons especificamente nas regiões denominadas de sítios doadores e receptores de *splicing* (PROUDFOOT; FURGER; DYE, 2002). Esses sítios de *splicing* são importantes sequências-sinal que orientam o local exato em que se encontra uma junção de exon com íntron que deverá sofrer *splicing* (PROUDFOOT; FURGER; DYE, 2002). Outras proteínas regulatórias participam do processo e estão envolvidas na modulação do *splicing*, ativando-o ou reprimindo-o pela ligação à elementos silenciadores ou acentuadores de *splicing* (URBANSKI; LECLAIR; ANCZUKÓW, 2018). Fatores de *splicing* podem se ligar

diretamente ao pré-RNA e regular os alvos seguintes da via de sinalização desse transcrito, regulação dependente da concentração do fator (URBANSKI; LECLAIR; ANCZUKÓW, 2018).

O processamento de *splicing* pode ocorrer em outros tipos de RNA além do RNA mensageiro, como RNAs não-codificadores (ncRNAs), RNAs ribossomais (rRNA), RNAs transportadores (tRNA), entre outros (LODISH et al., 2000; QUINN; CHANG, 2016). A ocorrência do *splicing* resulta em um transcrito maduro composto por uma ordem canônica de exons. Porém, também pode ocorrer o rearranjo de íntrons e exons e ordem diferente da usual, em uma vertente chamada de *splicing alternativo* (PROUDFOOT; FURGER; DYE, 2002; WANG; GERSTEIN; SNYDER, 2009).

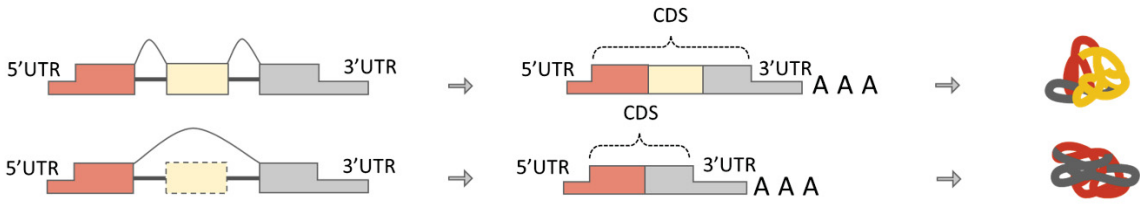


Figura 4: Representação de processamentos de *splicing* e *splicing* alternativo, de união dos íntrons (linhas cheias) e junção dos exons (retângulos) formando uma sequência codificadora (CDS) que poderá ser traduzida para proteína.

Fonte: Ilustração da autora (2020), conforme descrito por Proudfoot e colaboradores (2002) e Wang e colaboradores (2009).

1.3.3.1 *Splicing alternativo*

Descrito inicialmente em 1978 por Walter Gilbert, o processo de *splicing* alternativo apresenta a excisão de íntrons e união de exons em distintos rearranjos da ordem canônica, de forma que um único gene poderá originar mais de um transcrito diferente (GILBERT, 1978) (Figura 4). Quando traduzido, esses transcritos alternativos podem gerar proteínas diferentes, conhecidas como proteoformas (Figura 4). Dessa maneira, o processo de *splicing* alternativo pode aumentar ainda mais o repertório do transcriptoma (HALLEGGER; LLORIAN; SMITH, 2010). É um evento importante para

a complexidade dos organismos, tanto que mais de 90% dos genes humanos estão sob influência do processamento de *splicing* alternativo (BARBOSA-MORAIS et al., 2012; WANG; GERSTEIN; SNYDER, 2009).

1.3.3.2 *Splicing* alternativo em câncer

O evento de *splicing* é altamente regulado, tanto que disfunções no seu processamento ou na maquinaria de regulação são observadas em doenças como câncer, devido a alteração no padrão de *splicing* do tecido (PAJARES et al., 2007; SONG et al., 2018; URBANSKI; LECLAIR; ANCZUKÓW, 2018; VENABLES, 2006). Nesse quesito, a desregulação do *splicing* é importante para o desenvolvimento e manutenção de tumores, uma vez que podem ser expressas variantes que beneficiem a proliferação e migração celular, ou até evasão da resposta imune (URBANSKI; LECLAIR; ANCZUKÓW, 2018). Além de mutações gênicas que afetam a expressão de uma variante nas doenças, outra possibilidade é a disfunção da maquinaria de *splicing* que traz consequências para toda a cascata de sinalização do alvo desse processamento (URBANSKI; LECLAIR; ANCZUKÓW, 2018). Mudanças na concentração, localização, atividade ou composição dos fatores de *splicing* podem influenciar esse processamento nos tumores (PAJARES et al., 2007). Dessa forma, disfunções no processamento de *splicing* estão relacionadas à própria gênese e manutenção da massa tumoral, a sua progressão, invasão e até migração (SONG et al., 2018).

Por isso, as pesquisas oncológicas se propuseram a buscar esses tipos de transcritos diferencialmente expressos em câncer, ou ainda a utilizar reguladores de *splicing* na perspectiva de tratamento (URBANSKI; LECLAIR; ANCZUKÓW, 2018). Conforme a expressão diferencial de isoformas, as diferentes variantes de *splicing* têm potencial para serem novos alvos terapêuticos e biomarcadores (PAJARES et al., 2007; URBANSKI; LECLAIR; ANCZUKÓW, 2018). A presença de uma variante de *splicing* especificamente em um câncer ao invés do tecido saudável poderia ser um biomarcador de diagnóstico, prognóstico ou de predição (PAJARES et al., 2007). Por exemplo a variante MDM2 apresenta alta expressão aberrante em câncer de mama, associada a uma menor sobrevida de pacientes (PAJARES et al., 2007). Da mesma

maneira, foram encontrados eventos de *splicing* alternativo específicos de tecido tumoral e comuns para mais de um tipo de câncer diferente (TSAI et al., 2015).

Alguns trabalhos já demonstraram a presença de variantes de *splicing* diferencialmente expressas no transcriptoma de câncer cervical, auxiliando na caracterização dos subtipos tumorais e das lesões cervicais (KORI; YALCIN ARGA, 2018; SONG et al., 2007; WANG et al., 2017; WOERNER et al., 1995; XIAOGUANG et al., 2017). Meta-análises de dados de transcriptômica de câncer cervical puderam indicar novos candidatos a biomarcadores e alvos terapêuticos, além de biomarcadores conhecidos, oncogenes e supressores de tumor, entre outros (KORI; YALCIN ARGA, 2018). Identificação de variantes específicas em tecido tumoral e não em tecido saudável também são descobertas positivas para a melhor compreensão do perfil de expressão do câncer cervical (SONG et al., 2007). Há ainda possibilidade de lncRNAs terem atividade supressora contra o câncer cervical, como a inserção do lncRNA MEG3 em células tumorais cervicais levando à sua menor proliferação e maior taxa de morte celular (XIAOGUANG et al., 2017).

1.3.3.3 Programas para identificar variantes de *splicing*

Para encontrar novas variantes de *splicing*, é interessante utilizar uma abordagem de bioinformática por conta do enorme volume de dados que são produzidos através das técnicas de sequenciamento de RNA.

A montagem de variantes de *splicing* pode ser feita de duas formas: método baseado em uma anotação genômica de referência; e método *de novo* sem utilizar uma anotação de referência, em que os transcritos são montados conforme os algoritmos de cada programa (GARBER et al., 2011; HAAS; ZODY, 2010; SONG; SABUNCIYAN; FLOREA, 2016). Um ponto a se considerar quanto à utilização de programas baseados em anotação é que os transcritos que não estão presentes na referência são desconsiderados do resultado final. Por isso, a utilização de software com montagem *de novo* favorece a anotação de transcritos potencialmente novos, que ainda não foram anotados em referências genômicas (HAAS; ZODY, 2010; SONG; SABUNCIYAN; FLOREA, 2016).

O software CLASS2, do inglês "*Constraint-based Local Assembly and Selection of Splice variants*" (SONG; SABUNCIYAN; FLOREA, 2016) pode identificar variantes de splicing alternativo em dados de RNA-seq, construindo uma nova anotação gênica referente ao transcriptoma analisado. É um software que realiza montagem *de novo* dos transcritos (SONG; SABUNCIYAN; FLOREA, 2016). Ao não utilizar uma anotação genômica de referência, CLASS2 pode encontrar novos transcritos não-anotados e portanto potencialmente novos (SONG; SABUNCIYAN; FLOREA, 2016). É um programa comumente utilizado na reconstrução de anotação e com boa precisão na identificação dos transcritos quando comparado a outras abordagens (VENTURINI et al., 2018). Outro montador de anotação *de novo* conhecido e amplamente utilizado é Cufflinks (TRAPNELL et al., 2010, 2012). Entretanto, CLASS2 provou ser mais sensível para identificar variantes de *splicing* alternativo, detectando mais variantes que Cufflinks (SONG; SABUNCIYAN; FLOREA, 2016). Por isso escolhemos utilizar esse software para identificar as variantes de splicing nesse trabalho.

1.3.4 RNAs não-codificadores

A maior parte do transcriptoma humano é composta por RNAs não-codificadores (ncRNA), uma classe de transcritos que abrange diversas moléculas incluindo RNAs longos não-codificadores (lncRNA) (CHAN; TAY, 2018) e pequenos ncRNA, como microRNAs (miRNAs), pequenos RNAs de interferência (siRNA), entre outros (COLLINS; SCHÖNFELD; CHEN, 2011; KROL; LOEDIGE; FILIPOWICZ, 2010).

A maior classe de ncRNAs são os lncRNAs: longos transcritos de cerca de 200 nucleotídeos que não são traduzidos (WARD et al., 2015) ou que provavelmente não codificam proteínas funcionais (BATISTA; CHANG, 2013; ULITSKY; BARTEL, 2013). Há ainda subclasses de lncRNA, como os lncRNA intergênicos (lincRNA), os RNAs anti-senso (asRNA), pseudogenes, e RNAs circulares (circRNAS) (CHAN; TAY, 2018). Há crescentes evidências de que lncRNA apresentam padrões de expressão específicos em tecidos e até durante estágios do desenvolvimento (SONG et al., 2018), e muitos foram observados regulando expressão gênica (DERRIEN et al., 2012; ENGREITZ et al., 2016).

Como o próprio nome diz, pequenos ncRNA são compostos por cerca de 20 a 30 nucleotídeos e são não codificadores (COLLINS; SCHÖNFELD; CHEN, 2011). Dos pequenos ncRNA, os mais comuns são os miRNAs (KASHI et al., 2016). Esse tipo de ncRNA também exerce atividade regulatória, tanto de outros mRNA quanto silenciamento gênico mediado por cromatina (COLLINS; SCHÖNFELD; CHEN, 2011).

1.3.4.1 Processamento de ncRNAs

Apesar de não serem traduzidos ou não gerarem proteínas funcionais, os lncRNA são transcritos que estão sujeitos aos processamentos de RNA usuais, como o capeamento na extremidade 5', a poliadenilação na extremidade 3', e também o *splicing* (DERRIEN et al., 2012; QUINN; CHANG, 2016). De maneira semelhante, outros pequenos ncRNAs como miRNAs também podem sofrer capeamento na extremidade 5', adição de cauda poli-A na extremidade 3', e até *splicing* (HAMMOND, 2015; KROL; LOEDIGE; FILIPOWICZ, 2010).

A maior parte dos promotores em eucariotos é bidirecional, ou seja, a RNA polimerase II consegue transcrever um novo RNA partindo para ambas as direções 3' ou 5', de forma que RNAs mensageiros são transcritos quando no sentido senso de 5' para 3', e outros potenciais não codificadores RNAs geralmente para o anti-senso (COLLINS; SCHÖNFELD; CHEN, 2011; QUINN; CHANG, 2016). Com isso, os promotores são basicamente compartilhados. Os processamentos pós-transcricionais em lncRNAs e pequenos ncRNA podem ser distintos dos que ocorrem em outros tipos de transcritos (COLLINS; SCHÖNFELD; CHEN, 2011). Como exemplo, podemos citar os lncRNA MALAT1 e NEAT1 que são processados na extremidade 3' por uma RNAse P que gera pequenos subprodutos de RNA semelhantes ao RNA transportador (tRNA), além do transcrito maduro com uma extremidade 3' estável (QUINN; CHANG, 2016). Dependendo do lncRNA, esse pequeno subproduto de RNA pode ser estável e citoplasmático se derivado de MALAT1, ou instável se derivado de NEAT1 (QUINN; CHANG, 2016).

Conforme o nome, ncRNAs não são traduzidos (WARD et al., 2015), mas há outra definição de que provavelmente não codificam proteínas funcionais (BATISTA; CHANG, 2013; ULITSKY; BARTEL, 2013). Assim, ncRNA podem exibir capacidades

mínimas de codificação de proteína e portanto isso não deve ser uma característica biológica totalmente desconsiderada (DERRIEN et al., 2012; KASHI et al., 2016).

1.3.4.2 Regulação gênica com ncRNA

Os lncRNA e pequenos ncRNA têm papel na regulação gênica de outros transcritos e cada vez mais estudos apresentam possíveis funções biológicas, apesar do desafio de determinar funcionalidade em transcritos não-codificadores (WARD et al., 2015). Seus papéis regulatórios podem variar, desde pela ligação à complexos modificadores de histona ou a proteínas que ligam a DNA, como fatores de transcrição (LONG et al., 2017). Pequenos ncRNA como miRNA exercem atividade regulatória pós-transcricional quando ligam-se a sequências complementares nos transcritos-alvo, induzindo à sua degradação ou inibição da tradução do RNA alvo (KROL; LOEDIGE; FILIPOWICZ, 2010). Em virtude dessa capacidade regulatória com uma sequência-alvo, pequenos ncRNA como os siRNA podem ser utilizados em experimentos para diminuir a expressão de transcritos específicos, com aplicação em uma grande variedade de pesquisas (COLLINS; SCHÖNFELD; CHEN, 2011).

Até mesmo os lncRNA estão envolvidos na regulação do processamento de *splicing* alternativo, como por exemplo facilitando a ligação entre o fator de *splicing* em elementos regulatórios como silenciadores ou potencializadores (URBANSKI; LECLAIR; ANCZUKÓW, 2018). Até o próprio ato de transcrição de um lncRNA pode ser mais relevante do que o papel do próprio produto lncRNA na regulação (LONG et al., 2017). Exemplos de lncRNAs regulatórios conhecidos são H19 e XIST, que exercem importantes funções na determinação de *imprinting* epigenético e inativação do cromossomo X, respectivamente, inclusive apresentando disfunções em cânceres (BARTOLOMEI; TILGHMAN, 2014; CHAN; TAY, 2018; LOPES et al., 2003).

Diversos estudos apontam para a especificidade de expressão de lncRNAs em certos tecidos, células e sublocalizações celulares (DERRIEN et al., 2012; FLYNN; CHANG, 2014; PONJAVIC et al., 2009; SASAKI et al., 2007; WARD et al., 2015) e até ao longo do desenvolvimento (FLYNN; CHANG, 2014; LI et al., 2015). Há trabalhos que propõem que o perfil de expressão geral de lncRNA é ainda mais específico que de RNAs codificadores, observação persistente até quando corrigida para o menor

nível de expressão de lncRNAs (QUINN; CHANG, 2016). A sublocalização celular de um lncRNA é importante para poder entender a sua provável função, como lncRNA nucleares participando de modificação de histonas, ou lncRNAs citoplasmáticos contendo pequenas fases de leitura abertas, que possivelmente poderão ser traduzidas (LONG et al., 2017). Inclusive lncRNAs podem estar conservados entre espécies tanto para sua estrutura secundária quanto para sua função, mesmo quando sua sequência de nucleotídeos não é totalmente idêntica (DE RIE et al., 2017). Como no caso do lncRNA XIST, com a inativação do cromossomo X conservada em mamíferos (BARTOLOMEI; TILGHMAN, 2014), apesar da sequência de bases divergir entre espécies (NESTEROVA et al., 2001).

Muitos ncRNAs podem ser identificados no transcriptoma como transcritos anti-senso complementares de genes conhecidos (WARD et al., 2015) e por conta do seu processamento de RNA usual, como poliadenilação, esses transcritos aparecem em análises ômicas de RNA-seq (DERRIEN et al., 2012; KASHI et al., 2016; ULITSKY; BARTEL, 2013). Porém, ainda é desafiante discernir entre fragmentos alternativos de mRNA e sequências de ncRNAs (ULITSKY; BARTEL, 2013). Além disso, no geral a expressão de ncRNAs é menor que a expressão de mRNAs, em diversos tecidos humanos, o que também pode dificultar sua identificação (DERRIEN et al., 2012). Tanto que, em experimentos de sequenciamento, a necessidade de um valor de corte de leituras para determinar um transcrito pode levar à perda de informação sobre a presença de um lncRNA de baixa expressão (LONG et al., 2017).

1.3.4.3 ncRNAs em câncer

Sabe-se que ncRNAs podem estar presentes no contexto de diversas doenças, especialmente em cânceres (BATISTA; CHANG, 2013; BRUNNER et al., 2012; GUTSCHNER; DIEDERICHS, 2012; KOPP; MENDELL, 2018). Considerando sua expressão tecido-específica em tecido saudável, a mesma observação pode ser feita em tecidos tumorais (BRUNNER et al., 2012). Muitos estudos buscam a identificação das diferentes classes de ncRNAs em doenças como cânceres (BATISTA; CHANG, 2013; DENARO; MERLANO; LO NIGRO, 2019; DU; CHEN, 2018; ESTELLER, 2011; ULITSKY; BARTEL, 2013). Das mais diversas atividades desses

transcritos, foram observadas a influência da desregulação de lncRNA em câncer afetando proliferação celular tumoral, metástase e até apoptose (SONG et al., 2018). Como por exemplo a promoção de proliferação e migração celular tumoral causada pelo lncRNA MALAT1 (SONG et al., 2018; URBANSKI; LECLAIR; ANCZUKÓW, 2018). Há também o lncRNA LINC01133, que sequestra o fator de *splicing* SRSF6, promovendo metástase em modelos de camundongos de câncer coloretal (URBANSKI; LECLAIR; ANCZUKÓW, 2018). Outro exemplo de lncRNA com função descrita é HOTAIR, conhecido por estar superexpresso em cânceres de mama e até promovendo metástase (GUPTA et al., 2010).

Quanto ao câncer cervical, em seu transcriptoma é possível encontrar lncRNAs (AALIJAHAN; GHORBIAN, 2019; CHAN et al., 2016; DENARO; MERLANO; LO NIGRO, 2019; HOSSEINI et al., 2017; WANG et al., 2017; XIAOGUANG et al., 2017). Diversos lncRNA conhecidos exercem papéis funcionais importantes no câncer cervical, como HOTAIR, H19, NEAT1, entre outros (AALIJAHAN; GHORBIAN, 2019). Por exemplo, o lncRNA MALAT1 (ou NEAT2) é conhecido por estar superexpresso em diversos tipos de câncer, inclusive o câncer de colo de útero (GUTSCHNER; DIEDERICHS, 2012). Sua função reside no fato de regular o *splicing* alternativo através da manutenção dos níveis de fatores de *splicing*. Consequentemente, disfunções na sua expressão podem afetar importantes regulações na biologia de tumores (GUTSCHNER; DIEDERICHS, 2012). Também é importante lembrar de lncRNAs com atividade supressora de tumor, em câncer cervical, como MEG3 associado à menor proliferação e maior taxa de apoptose em células cervicais tumorais (XIAOGUANG et al., 2017). Dessa maneira, a identificação de novos lncRNAs em câncer cervical pode ser interessante para diagnóstico, biomarcadores e até alvos terapêuticos.

1.3.4.4 ncRNAs e *splicing* alternativo

Os ncRNAs, tanto lncRNA quanto pequenos ncRNA, podem sofrem o processamento de *splicing* (HAMMOND, 2015; KROL; LOEDIGE; FILIPOWICZ, 2010). Quanto aos lncRNA, a maior parte sofre *splicing,* porém apresentam diferenças estruturais quanto aos genes codificadores para proteína (DERRIEN et al., 2012). A

estrutura de lncRNAs é composta de somente dois exons em 42% dos transcritos, comparado com 6% dos genes codificadores de proteína (DERRIEN et al., 2012). Além disso, os lncRNA apresentam tanto exons quanto íntrons mais longos que o usual para genes codificadores, mas com sítios de *splicing* canônicos (GT/AG) na maioria dos casos (DERRIEN et al., 2012). Mais de 25% dos lncRNA sofrem o processamento de *splicing* alternativo, gerando ao menos dois transcritos diferentes por locus gênico (DERRIEN et al., 2012).

Ainda há necessidade de mais pesquisas a respeito da ocorrência de *splicing* alternativo em ncRNAs, inclusive em câncer cervical. As tecnologias atuais de sequenciamento e análise de expressão gênica diferencial trazem perspectiva para essa busca.

1.4 MATERIAL E MÉTODOS

Todas as análises realizadas nesse trabalho utilizaram da infraestrutura do laboratório de Bioinformática do Instituto Carlos Chagas (ICC) e da plataforma RPT04A - Bioinformática – RJ, da Rede de Plataformas Tecnológicas da Fundação Oswaldo Cruz (Fiocruz). Os programas utilizados para a execução das etapas foram escritos nas linguagens de programação Perl e R.

1.4.1 Conjunto de dados

As etapas de obtenção das amostras de tumores e extração do RNA foram feitas pelos grupos de pesquisa da Drª Luisa Lina Villa (Faculdade de Medicina da Universidade de São Paulo - USP), Drª Laura Sichero (Instituto do Câncer do Estado de São Paulo) e da Drª Patricia Savio de Araújo Souza (Universidade Federal do Paraná - UFPR). Essas etapas serão resumidas a seguir.

Nesse trabalho foi estudado o transcriptoma de tumores cervicais de 11 pacientes brasileiras positivas para o subtipo 16 de HPV. A coleta de tecido foi aprovada pelo Comitê de Ética em Pesquisa da Faculdade de Medicina da Universidade de São Paulo, sob registro no protocolo de pesquisa n° 033/16, intitulado

“Avaliação do perfil de expressão gênica de dois subtipos de câncer de colo uterino: carcinoma escamoso e adenocarcinoma” (Figura 25).

O material biológico foi coletado pelo Serviço Ginecológico e armazenado no Laboratório de Biologia Molecular (LBM) do Centro de Investigação Translacional em Oncologia (CTO) do Instituto do Câncer do Estado de São Paulo (ICESP). Dos 11 tumores amostrados, 8 foram confirmados histopatologicamente como carcinoma cervical de células escamosas (SCC) e 3 identificados histopatologicamente como adenocarcinoma cervical (ADC).

O transcriptoma das amostras de câncer cervical foi obtido pelo sequenciamento pareado de alta vazão do RNA, realizado nas plataformas do Laboratório Central de Tecnologias de Alto Desempenho (LaCTAD) da Universidade Estadual de Campinas (UNICAMP), no equipamento Illumina HiSeq 2500.

1.4.2 Controle de qualidade do sequenciamento e mapeamento dos alinhamentos

O controle de qualidade das leituras sequenciadas foi feito com **Trim Galore** (versão 0.4.0). Realizamos o mapeamento dos transcriptomas no genoma humano do Ensembl (ZERBINO et al., 2018) versão GRCh37/hg-19 utilizando o programa **HISAT2** (versão 2.1.0) (KIM; LANGMEAD; SALZBERG, 2015). Os arquivos BAM gerados das onze amostras foram utilizados nos próximos passos.

1.4.3 Identificação de variantes de *splicing* de genes não-codificadores

Para melhor compreensão e realização da pesquisa, dividimos o trabalho em três etapas:

- Na primeira etapa, realizamos a identificação das variantes de *splicing* e construção dos arquivos de anotação personalizados (Figura 5);

- Na segunda etapa, filtramos os arquivos de anotação para conter genes não anotados e somente genes potencialmente não codificadores (Figura 6);

- Na terceira etapa, analisamos a expressão diferencial com as anotações concatenadas (Figura 7).

1.4.3.1 Primeira etapa: identificação das variantes de *splicing* e construção das anotações sem referência;

Na primeira etapa desse trabalho, criamos os arquivos de anotação personalizados, contendo as coordenadas para as variantes de *splicing* identificadas sem uma anotação de referência, para podermos identificar transcritos potencialmente novos (Figura 5).

Com os arquivos de alinhamento BAM das onze amostras de câncer cervical, executamos a ferramenta **SAMTools (versão 1.5)** (LI et al., 2009) no modo ***View*** para filtrar somente as leituras mapeadas uma vez no genoma. É uma estratégia utilizada para lidar com "multi-leituras", leituras que mapeiam em diversas regiões do genoma (FINOTELLO; DI CAMILLO, 2015). Em seguida, executamos as funções ***Sort*** e ***Index*** do mesmo pacote **SAMTools** (LI et al., 2009) para ordenar e indexar os arquivos, respectivamente. Dessa forma, geramos arquivos de alinhamento ordenados de formato BAM e BAI das onze amostras tumorais.

A maneira convencional de analisar sequenciamentos em série é primeiramente processar os arquivos de alinhamento individualmente e então concatenar as anotações, criando uma única anotação referência de todas as amostras (CONESA et al., 2016; SONG et al., 2019). É uma maneira de analisar amostras em série visto que os montadores e identificadores de variantes de *splicing* mais comumente utilizados só recebem um único arquivo de entrada, como CLASS2 (SONG et al., 2019). Entretanto, essa metodologia pode prejudicar a precisão da identificação dos transcritos (CONESA et al., 2016; SONG et al., 2019). Além disso, no contexto de câncer cervical, os tipos ADC e SCC são molecularmente heterogêneos entre as amostras de diferentes pacientes (BURK et al., 2017).

Pensando nisso, nós nos propomos a buscar por variantes de *splicing* através de duas abordagens:

Equivalente ao método convencional de análise, na **abordagem A** fizemos a concatenação dos arquivos de **anotação** das diferentes amostras. Buscando encontrar transcritos presentes nos tipos histológicos, porém com expressão variável entre as amostras, nós desenhamos a **abordagem B** em que realizamos a concatenação dos arquivos de **alinhamento** das onze amostras antes de identificar os transcritos. Em detalhes:

- **Abordagem A:** Nessa abordagem, os onze arquivos de alinhamento foram executados individualmente com a ferramenta **CLASS2 (versão 1.7)** para identificar computacionalmente as variantes de *splicing* (SONG; SABUNCIYAN; FLOREA, 2016). Os onze arquivos de anotação em formato GTF gerados foram concatenados em um único arquivo de anotação com o programa **StringTie Merge (versão 1.3.3b)** (PERTEA et al., 2016). Dessa maneira geramos um único arquivo de anotação referência para todas as amostras.

- **Abordagem B:** Nessa abordagem, os onze arquivos de alinhamento foram concatenamos em um único arquivo formato BAM, utilizando o programa **SAMTools** no modo ***Merge*** (LI et al., 2009). Esse único arquivo de alinhamento foi usado como entrada para a execução de **CLASS2 (versão 1.7)** (SONG; SABUNCIYAN; FLOREA, 2016) gerando um arquivo de anotação formato GTF. Assim geramos um único arquivo de anotação referência para todas as amostras.

Para ambas as abordagens A e B, foram testados três valores de limiar para a execução de **CLASS2 (versão 1.7)** (SONG; SABUNCIYAN; FLOREA, 2016): 5%, 10% e 20% (parâmetros -F 0.05, -F 0.1 e -F 0.2 respectivamente). Esses valores de limite indicam que CLASS2 classifica como transcrito somente se a abundância de leituras usadas para formar o transcrito corresponde respectivamente a 5%, 10% ou 20% das leituras totais que compõem o gene (SONG; SABUNCIYAN; FLOREA, 2016). Portanto, ao utilizarmos as abordagens A e B e testarmos os valores de limiar de 5%, 10% e 20%, obtivemos no total **seis arquivos de anotação** diferentes, sendo identificados como as seis abordagens de: A 5%, A 10%, A 20%, B 5%, B 10% e B 20%.

1.4.3.1.1 Visualização das anotações

Para conferir visualmente a criação das anotações criadas no item anterior, utilizamos o programa **IGV (versão 2.4.1)** (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013). Primeiramente utilizamos sua ferramenta **IGVtools** para ordenar e indexar os arquivos de anotação de referência GTF. Para visualizar no IGV, colocamos as anotações A 5%, A 10%, A 20%, B 5%, B 10% e B 20%, juntamente com os arquivos de alinhamento das onze amostras como entrada no programa (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013). Dessa maneira pudemos visualizar o mapeamento das leituras dos arquivos de alinhamentos sobre as anotações criadas.

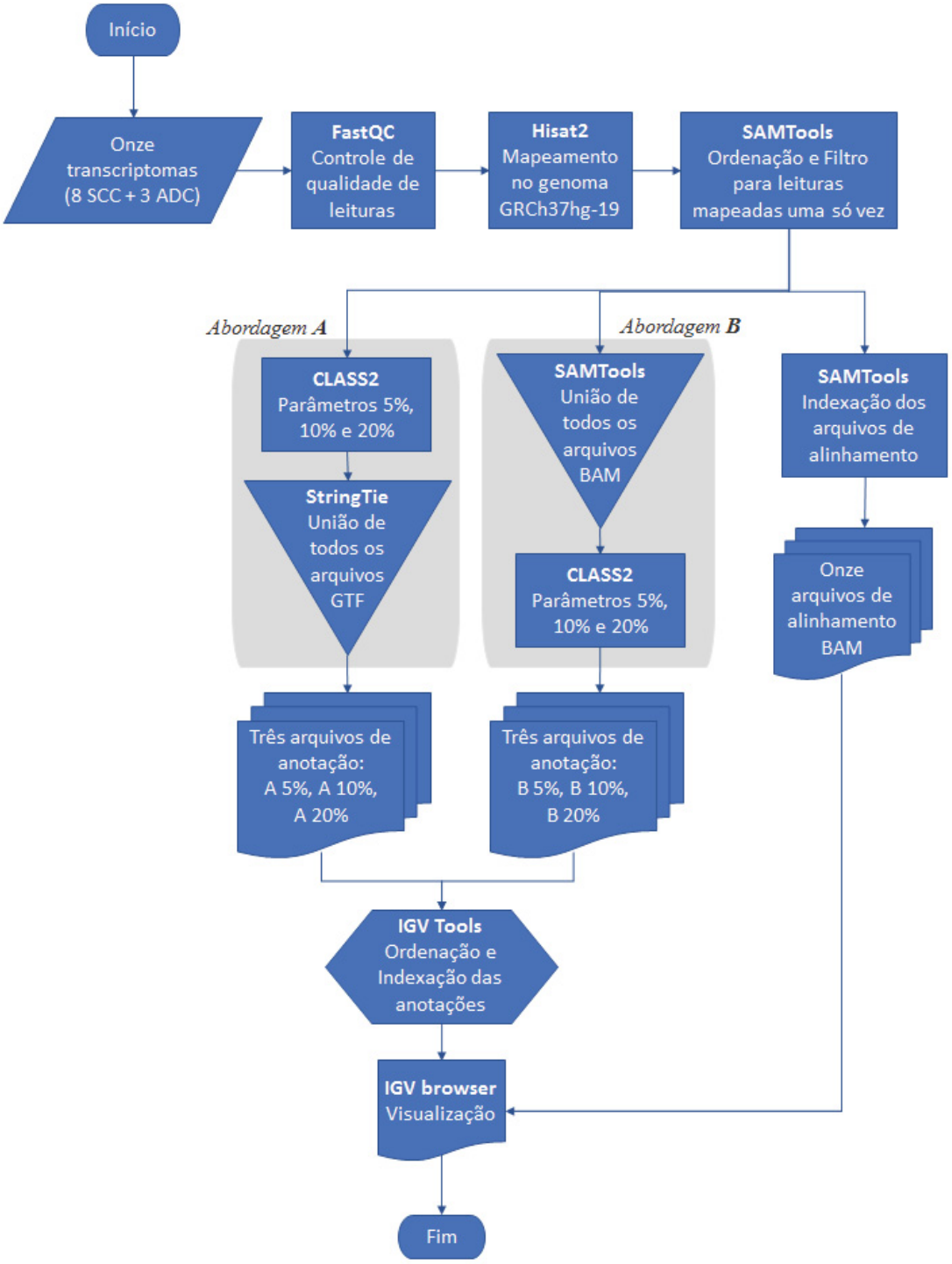


Figura 5: Fluxograma da primeira etapa do trabalho – Identificação das variantes de *splicing* e criação dos arquivos de anotação das seis abordagens A 5%, A 10%, A 20%, B 5%, B 10%, B 20%.

Fonte: A autora (2020).

1.4.4 Segunda etapa: identificação de transcritos potencialmente novos e não codificadores

Na segunda etapa do trabalho buscamos identificar os transcritos potencialmente novos encontrados através das nossas abordagens, assim como selecionamos somente genes sem potencial de codificação, para buscar por genes não codificadores (Figura 6).

1.4.4.1 Correção dos arquivos de anotação

Antes de prosseguir com as análises, foi necessário corrigir os arquivos de anotação gerados por **CLASS2** (SONG; SABUNCIYAN; FLOREA, 2016). Nessas anotações há a determinação do senso da fita a qual foi originado o transcrito. Através do uso de um sinal de mais (+) ou menos (-) é indicado se o transcrito foi sequenciado pela fita senso ou anti-senso, respectivamente. Porém, se o software **CLASS2** (SONG; SABUNCIYAN; FLOREA, 2016) não consegue especificar o senso da fita do transcrito, a orientação é identificada como um ponto (.). O ponto é um caractere que não é reconhecido pelos programas utilizados adiante no trabalho, como o software de contagem **HTSEQ** (ANDERS; PYL; HUBER, 2015). Portanto, criamos um programa em Perl para corrigir os arquivos de anotações e substituir os pontos pela determinante de fita senso (+).

1.4.4.2 Identificação dos transcritos potencialmente novos

O programa de identificação de variantes de *splicing* alternativo **CLASS2** (SONG; SABUNCIYAN; FLOREA, 2016) não utiliza uma anotação de referência para identificar os transcritos. Ao invés, os transcritos recebem identificadores genéricos, referentes ao número do cromossomo, ao número do gene identificado naquele cromossomo, e ao número do transcrito identificado naquele gene, como MSTRG.#.

Inicialmente nosso intuito era identificar quais transcritos já foram anotados previamente na referência **Ensembl** (ZERBINO et al., 2018) e quais eram

potencialmente novos, realizando então a avaliação da performance de montagem e identificação de transcritos.

Para fazer isso, utilizamos o programa **Grader**, do software **PsiCLASS** (SONG et al., 2019), de mesma autoria de **CLASS2** (SONG; SABUNCIYAN; FLOREA, 2016). Esse programa recebeu de entrada os seis arquivos de anotação personalizados (A 5%, A 10%, A 20%, B 5%, B 10% e B 20%) para serem comparados individualmente com a anotação de referência **Ensembl** versão **GRCh37/hg-19** (ZERBINO et al., 2018).

Como arquivo de saída, foi gerada uma lista chamada *Grader Precision* com a relação de quais transcritos do **Ensembl** (ZERBINO et al., 2018) eram identificados como equivalentes aos transcritos das anotações das abordagens (SONG et al., 2019). Além disso, também apontou o número de exons de cada transcrito identificado (SONG et al., 2019). Dessa forma, essa lista apresentou transcritos previamente anotados pelo **Ensembl**, transcritos não-anotados e seu número de exons.

Em uma primeira análise da quantidade de transcritos potencialmente novos e já anotados na referência, optamos por prosseguir com as análises seguintes somente com as abordagens A 20% e B 20% (Tabela 2). Essas abordagens apresentaram menor número de transcritos não anotados, portanto foram escolhidas para diminuir a possibilidade de transcritos falsos-positivos identificados, derivados de artefatos de mapeamento ou contaminações (CONESA et al., 2016).

Então, com o intuito de filtrar nossos arquivos de anotação das abordagens A 20% e B 20% para possuir somente transcritos potencialmente novos não-anotados pelo **Ensembl** (ZERBINO et al., 2018) e composto por mais de um exon, elaboramos um programa em linguagem **Perl** para realizar a filtragem desses transcritos.

Nosso programa construiu um novo arquivo de anotação das abordagens A 20% e B 20% contendo somente os transcritos com mais de um exon que não receberam um identificador **Ensembl** equivalente (ZERBINO et al., 2018) na análise anterior com **PsiCLASS Grader** (SONG et al., 2019).

1.4.4.3 Seleção de genes não codificadores

Visto que nosso trabalho foca em genes não-codificadores, foi necessário excluir os genes potencialmente codificadores dos nossos arquivos de anotação, tanto os genes não anotados provenientes das anotações das abordagens A 20% e B 20%, quanto os genes anotados da anotação do **Ensembl** (ZERBINO et al., 2018).

Para excluir os genes codificadores de proteína da referência **Ensembl** (ZERBINO et al., 2018), primeiro utilizamos a linguagem **R** para fazer a seleção de genes cujos biotipos não eram codificadores de proteína a partir do arquivo de anotação de formato GTF.

Depois, com essa lista de genes não codificadores, utilizamos a linguagem **Perl** para criar um novo arquivo de anotação do **Ensembl** (ZERBINO et al., 2018) contendo apenas genes não codificadores de proteína. Esse arquivo de anotação do **Ensembl** (ZERBINO et al., 2018) contendo apenas genes não codificadores será utilizado no final dessa segunda etapa.

Já para excluir os genes não anotados e codificadores de proteína, dos arquivos de anotação personalizados das abordagens A 20% e B 20%, foi preciso utilizar um software que realizasse essa análise de potencial de codificação.

Primeiramente foi necessário converter os arquivos de anotação de formato GTF para o formato FASTA, que é o formato de entrada necessário para o software de análise de potencial de codificação. Para converter as anotações de formato GTF para FASTA, utilizamos o programa **GFFread versão 0.11.6** (TRAPNELL et al., 2012). Usamos os parâmetros “-g” para indicar o genoma humano de referência em formato FASTA, no caso Ensembl hg-19, e “-w” para indicar o arquivo de saída.

Então calculamos o potencial de codificação de cada transcrito das anotações das abordagens utilizando o software **CPC versão 2 (CPC2)** (KANG et al., 2017). Os parâmetros utilizados foram “-i” para receber o arquivo de entrada em FASTA e “-o” para indicar o arquivo de saída. O software **CPC2** gerou uma lista com a relação dos transcritos e seu potencial de codificação, identificando-os como codificador ou não codificador (KANG et al., 2017).

Como nosso objetivo é identificar genes não codificadores, e a lista de potencial de codificação estava separada por transcritos, precisávamos ter somente genes os quais todos os transcritos eram não codificadores. Para isso, criamos um programa em **Perl** para excluir todos os genes que tinham ao menos um transcrito com potencial codificador das nossas anotações personalizadas.

Ao final desses passos de filtragem, obtivemos enfim dois arquivos de anotação, A 20% e B 20%, somente com genes não anotados no Ensembl e sem potencial de codificação. Além disso, temos a anotação de referência do **Ensembl** (ZERBINO et al., 2018) somente com genes não codificadores. Com a pretensão de identificar tanto transcritos não codificadores já anotados na referência e também os potencialmente novos, precisamos utilizar ambas as anotações nas contagens e análise de expressão diferencial da próxima etapa (CONESA et al., 2016). Para unir as anotações A 20% e B 20% com Ensembl, fizemos a concatenação dos arquivos das abordagens individualmente com a anotação do **Ensembl** (ZERBINO et al., 2018) utilizando o programa **Stringtie Merge (versão 1.3.3b)** (PERTEA et al., 2015). Dessa forma, obtivemos duas anotações, referentes à cada abordagem: Anotação A 20% e Ensembl, de genes sem potencial de codificação e potencialmente novos; e anotação B 20% e Ensembl, de genes sem potencial de codificação e potencialmente novos.

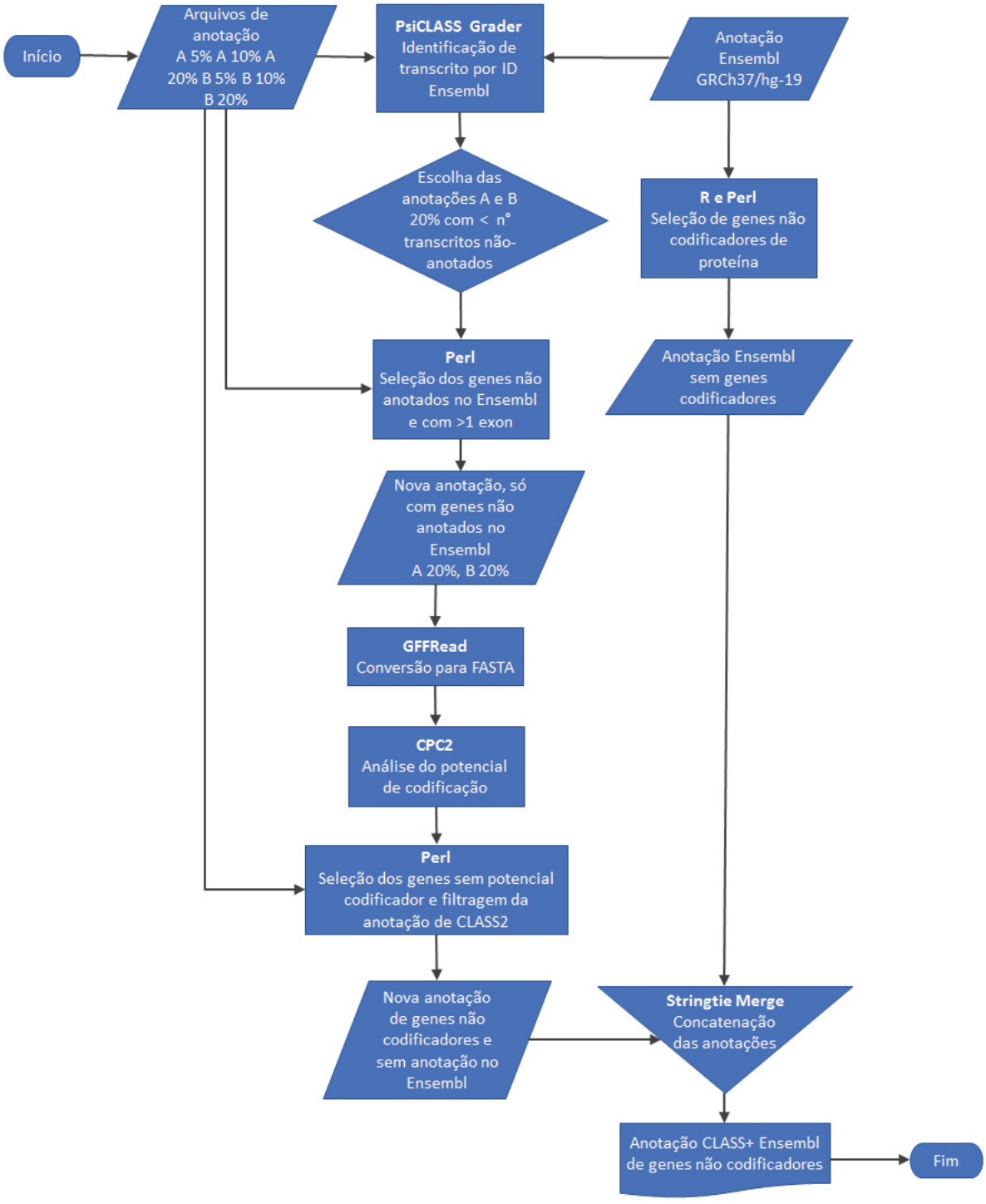


Figura 6: Fluxograma da segunda etapa do trabalho -Seleção de genes não codificadores e não anotados na referência Ensembl.

Fonte: A autora (2020).

1.4.5 Terceira etapa - Análise de expressão diferencial

Na última etapa, buscamos avaliar a expressão diferencial nos onze tumores cervicais dos transcritos não codificadores anotados previamente na referência **Ensembl** (ZERBINO et al., 2018) e também os encontrados potencialmente novos (Figura 7).

1.4.5.1 Contagem de leituras

Primeiramente realizamos a contagem das leituras sequenciadas nos onze alinhamentos de tumores cervicais, de acordo com as anotações concatenadas criadas na Segunda Etapa:

- Anotação A 20% com Ensembl, de genes não codificadores e potencialmente novos;
- Anotação B 20% com Ensembl, de genes não codificadores e potencialmente novos.

Para realizar a contagem, utilizamos a função ***Count*** do programa **HTSEQ (versão 0.11.1)** (ANDERS; PYL; HUBER, 2015) dando os onze arquivos de alinhamento como entrada juntamente com as anotações mencionadas anteriormente. Os parâmetros adicionais utilizados foram: “--format=BAM” para indicar o formato do arquivo de entrada; “--order=name” para indicar a ordenação do arquivo de entrada; “--stranded=reverse” para indicar o modo do sequenciamento das amostras como específico de fita reversa; “--mode=intersection-nonempty” para indicar ao programa como lidar com leituras que se sobrepõem; “--idattr=gene_id” para a contagem de leituras ser feita por gene.

1.4.5.2 Análise de expressão diferencial

Somente os números de contagens não são suficientes para comparar a expressão gênica entre amostras, pois podem ser influenciados por fatores adversos do sequenciamento, tamanho dos transcritos e número total de leituras mensuradas

(CONESA et al., 2016). Portanto, com as tabelas de contagens de leituras por gene, realizamos a análise de expressão diferencial com o programa **DESeq2 (versão 1.24.0)** (LOVE; HUBER; ANDERS, 2014) do ambiente de programação **R (versão 3.6.1)** (R CORE TEAM, 2018).

Para a análise do DESeq2, primeiramente realizamos uma filtragem prévia, para retirar os genes com baixas contagens, e mantermos somente as linhas com mais de 10 leituras. Para isso executamos “keep <- rowSums(counts(dds)) >= 10” e “dds <- dds[keep,]”.

Os dados de expressão diferencial foram gerados considerando p-valor ajustado menor ou igual a 0,05, logFC maior ou igual a 1,5 e menor ou igual a -1,5. Importante ressaltar que a taxa de falsos-positivos no pacote DESeq2 é definido como p-valor ajustado, mas em outros programas também é conhecido como FDR, do inglês “False Discovery Rate” (LOVE; HUBER; ANDERS, 2014).

Após calcular a expressão diferencial com DESeq2 e a fim de visualizar os genes diferencialmente expressos em *heatmap*, nós transformamos os dados de contagem para Rlog (logaritmo regularizado, do inglês “regularized logarithm”). Tal transformação foi feita com o intuito de remover a dependência da variância sobre a média de contagens, particularmente a alta variância do logaritmo da contagem quando a média é baixa (LOVE; HUBER; ANDERS, 2014). Essa transformação foi feita executando “rld <- rlog(dds, blind=FALSE)” e “rld <- normTransform(dds)”.

Para construir o *heatmap*, selecionamos os 50 genes de maior expressão diferencial e de menor p-valor ajustado, executando:

```
resSig <- subset(res, padj <= 0.05)
resSig <- subset(resSig, log2FoldChange >= 1.5 |
log2FoldChange <= -1.5)
resSig <- resSig[order(resSig$padj),]
res50 <- row.names(head(resSig, 50))
```

A construção do *heatmap* foi feita executando:

```
heatmap.2( assay(rld)[ topVarGenes, ], scale="row",
           trace="none", dendrogram="none",
           offsetRow = -0.3, offsetCol = -0.2,
           lmat=rbind( c(0, 3, 4), c(2,1,0) ),
           lwid=c(1, 5, 3),
           margins = c(6,10),
           xlab = "Amostras de câncer cervical",
```

```
            key.title = NA,
            Colv = FALSE,
            col = colorRampPalette( rev(brewer.pal(9,
"RdBu")))(255))
```

Para conferir a correta montagem dos transcritos, visualizamos genes de maior expressão diferencial e valor de *fold-change* utilizando o programa **IGV** (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013). Selecionamos alguns transcritos dentre os 50 genes mais diferencialmente expressos de cada abordagem, para poder conferir visualmente os alinhamentos sobre as anotações no IGV. O programa de visualização de anotações IGV provém uma anotação do RefSeq para comparar com os arquivos de entrada dados pelo usuário ou usuária, e nós também utilizamos essa referência do RefSeq para comparar com nossas anotações personalizadas (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013). Tais transcritos escolhidos serão descritos adiante, no tópico de resultados **2.4 Visualização dos genes mais diferencialmente expressos**.

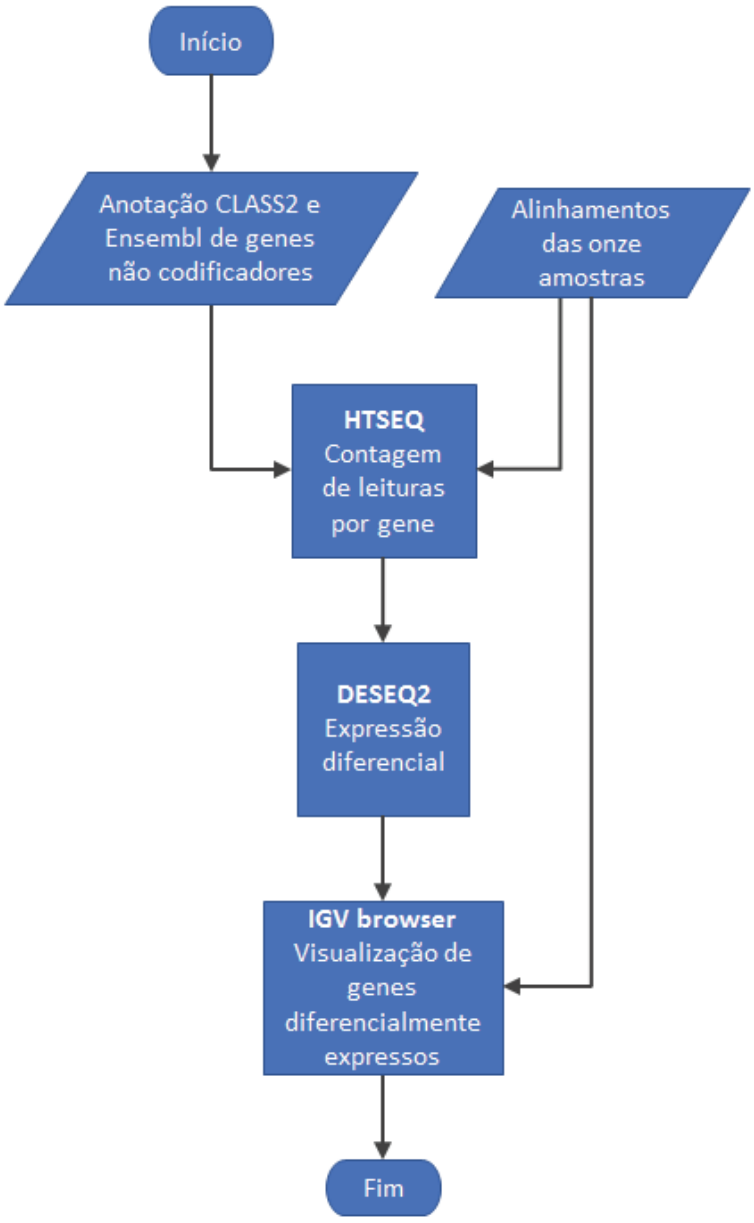


Figura 7: Fluxograma da terceira etapa do trabalho - análise de expressão diferencial.

Fonte: A autora (2020).

2 RESULTADOS

2.1 Controle de qualidade do sequenciamento

Para monitorar a qualidade do sequenciamento das amostras, fizemos o controle de qualidade das leituras. O sequenciamento das amostras gerou em média 85 milhões e 600 mil leituras pareadas para cada amostra (Tabela 1). Após a limpeza dos dados para remoção de leituras de baixa qualidade, restaram em média 85 milhões e 383 mil leituras para cada amostra. Destas, cerca de 69 milhões e 555 mil leituras por amostra foram mapeadas no genoma humano do Ensembl versão GRCh37/hg-19. Isso corresponde a 81,25% de leituras mapeadas no genoma humano (Tabela 1).

Tabela 1: Número de leituras pareadas antes e após controle de qualidade, e total mapeada no genoma humano GRCh37/hg-19.

Amostra	N° total de leituras pareadas	N° de leituras após controle de qualidade	Mapeamento com Ensembl GRCh37/hg-19	Porcentagem de leituras mapeadas no genoma humano
ADC_09	84 133 674	83 719 920	66 353 698	78,87%
ADC_10	92 988 595	92 657 439	75 006 643	80,62%
ADC_11	88 792 224	88 595 190	69 669 709	78,46%
SCC_01	91 316 248	91 050 622	72 716 269	79,63%
SCC_02	62 647 028	62 594 763	57 469 289	91,73%
SCC_03	62 355 298	62 252 599	55 325 137	88,72%
SCC_04	89 723 933	89 450 444	72 334 390	80,62%
SCC_05	105 033 114	104 877 150	85 001 982	80,93%
SCC_06	92 362 730	92 162 264	73 512 112	79,59%
SCC_07	88 678 949	88 445 693	71 269 598	80,37%
SCC_08	83 568 403	83 410 555	66 446 626	79,51%
Média	85 600 17	85 383 330	69 555 041	81,25%

Fonte: a autora (2020).

2.2 Identificação dos transcritos potencialmente novos e não codificadores

Em todas as anotações geradas com o uso de CLASS2 nós encontramos diversos transcritos não anotados nas referências do Ensembl de acordo com a identificação feita pelo software PsiClass Grader (SONG et al., 2019) (Tabela 2). No método A, entre 95% a 97% dos transcritos encontrados não foram identificados como anotados na referência Ensembl. No método B, ao utilizarmos os parâmetros 5% e 10% foram identificados 92% a 90% de transcritos potencialmente novos respectivamente, não anotados na referência Ensembl. A abordagem B 20% foi a que mais apresentou transcritos anotados na referência Ensembl, com cerca de 26% de transcritos previamente anotados e cerca de 73% de transcritos potencialmente novos (Figura 8 e Tabela 2).

Ao testarmos parâmetros diferentes para CLASS2, pudemos perceber a variação no número de transcritos identificados (Figura 8 e Tabela 2). Em virtude do alto número de transcritos potencialmente novos encontrados nas anotações, escolhemos continuar com os próximos passos desse trabalho somente com as abordagens A 20% e B 20%. Essas abordagens apresentaram o menor número de transcritos não anotados dentro de cada método. Essa escolha foi feita com o intuito de testar os dois métodos A e B de identificação de variantes de *splicing* e de obter menor taxa de falsos-positivos na análise de expressão diferencial.

Tabela 2: Número de transcritos identificados em cada abordagem, não anotados e potencialmente novos ou já anotados na referência do Ensembl versão GRCh37/hg-19

Abordagem	Transcritos não anotados, potencialmente novos	Transcritos anotados no Ensembl GRCh37/hg-19	Total de transcritos identificados
A 5%	149003 (97,33%)	4090 (2,67%)	153093
A 10%	126807 (96,75%)	4257 (3,25%)	131064
A 20%	103658 (95,97%)	4351 (4,03%)	108009
B 5%	100575 (92,51%)	8145 (7,49%)	108720
B 10%	81128 (90,97%)	8055 (9,03%)	89183
B 20%	65776 (73,75%)	7893 (26,25%)	73669

Fonte: a autora (2020).

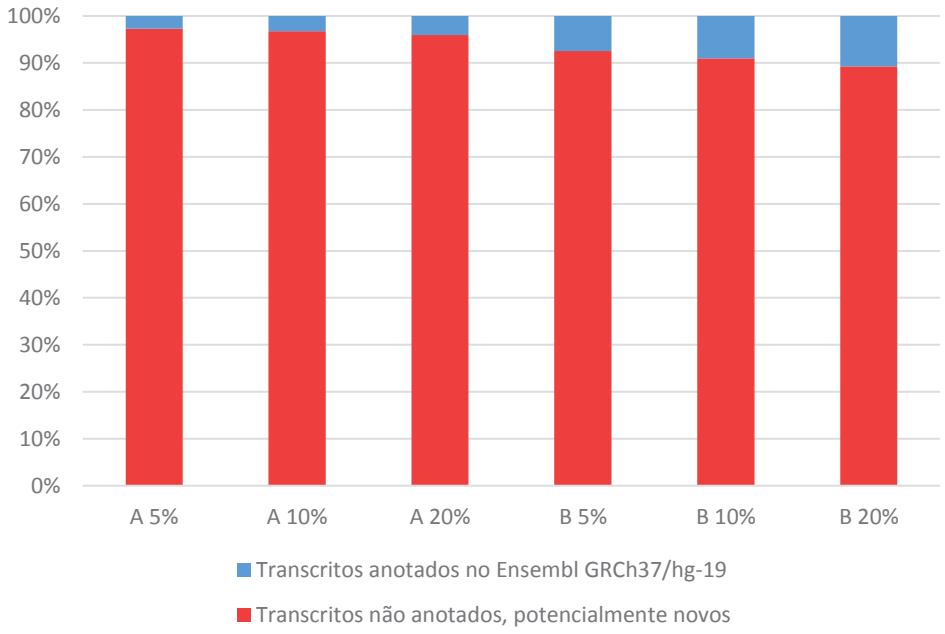


Figura 8: Proporção de transcritos anotados e não anotados encontrados com as abordagens.

Fonte: a autora (2020).

Para podermos identificar somente os transcritos não codificadores em nossas abordagens, averiguamos o potencial codificador dos transcritos não anotados e potencialmente novos através do software CPC2 (KANG et al., 2017). Ademais, somente os transcritos que apresentaram mais de um exon em sua estrutura foram considerados, a fim de diminuir a taxa de falsos-positivos identificados. Na abordagem A 20%, foram descobertos a maior quantidade de transcritos não anotados sem potencial codificador: cerca de 33 mil transcritos sem potencial codificador, totalizando 33,24% dos transcritos não anotados com mais de um exon identificados nessa abordagem (Tabela 3 e Figura 9). Na abordagem B 20%, a maior parte dos transcritos não anotados apresentaram potencial codificador: cerca de 57 mil transcritos com potencial codificador, totalizando 96,6%. Assim, na abordagem B 20% somente um pouco mais de 2 mil transcritos não-anotados foram considerados sem potencial codificador (Tabela 3 e Figura 9).

Para as análises seguintes, foram descartados os transcritos com potencial codificador e construímos as anotações das abordagens A 20% e B 20% somente com os transcritos sem potencial codificador (Tabela 3 e Figura 9).

Tabela 3: Quantidade de transcritos não-anotados com mais de um exon identificados em cada abordagem, com e sem potencial de codificação.

	Transcritos não anotados com > 1 exon	Transcritos sem potencial codificador	Transcritos com potencial codificador
A 20%	99966	33224 (33,24%)	66742 (66,76%)
B 20%	59924	2039 (3,4%)	57885 (96,6%)

Fonte: a autora (2020).

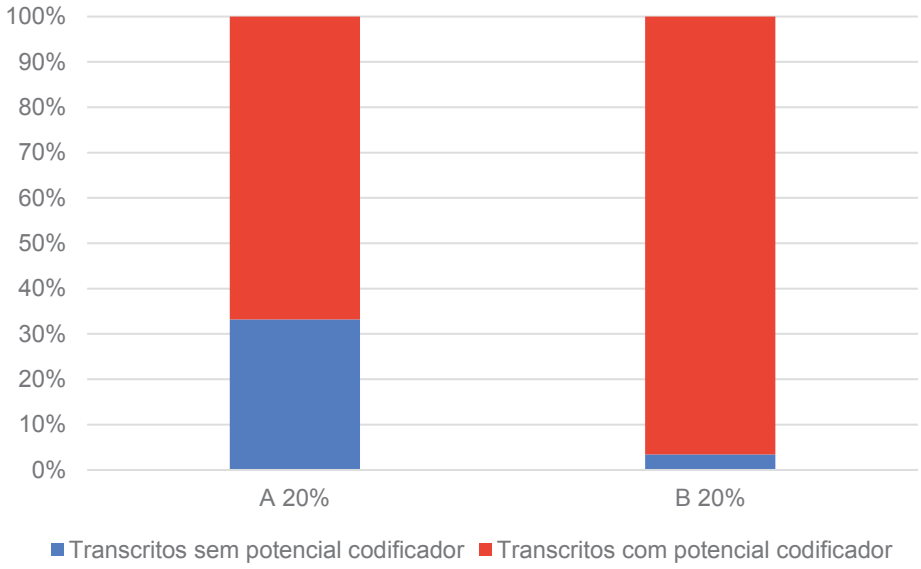


Figura 9: Proporção de transcritos não anotados com e sem potencial de codificação para proteína.

Fonte: a autora (2020).

2.3 Expressão diferencial dos genes nas abordagens A 20% e B 20%

A fim de mensurar a taxa expressão de genes não codificadores potencialmente novos e já anotados na referência Ensembl, realizamos a contagem de leituras dos onze tumores cervicais utilizando as anotações criadas previamente:

- Anotação A 20% concatenada com Ensembl, de genes não codificadores já anotados e também potencialmente novos;
- Anotação B 20% concatenada com Ensembl, de genes não codificadores já anotados e também potencialmente novos.

Para cada abordagem, nós fizemos as contagens de leituras por gene utilizando HTSEQ (ANDERS; PYL; HUBER, 2015) e em seguida mensuramos a expressão diferencial dos genes através do pacote DESEQ2 (LOVE; HUBER; ANDERS, 2014).

Ao longo das etapas, nós observamos uma redução na quantidade de transcritos encontrados diferencialmente expressos nas abordagens, conforme podemos ver na Figura 10. Na abordagem A 20%, dos 103 mil e 658 transcritos não anotados no Ensembl que foram identificados nessa abordagem, somente 284 transcritos sem potencial de codificação estavam diferencialmente expressos. Na abordagem B 20%, partindo de 65 mil e 776 transcritos identificados que não estavam anotados, somente 28 transcritos sem potencial de codificação estavam diferencialmente expressos (Figura 10).

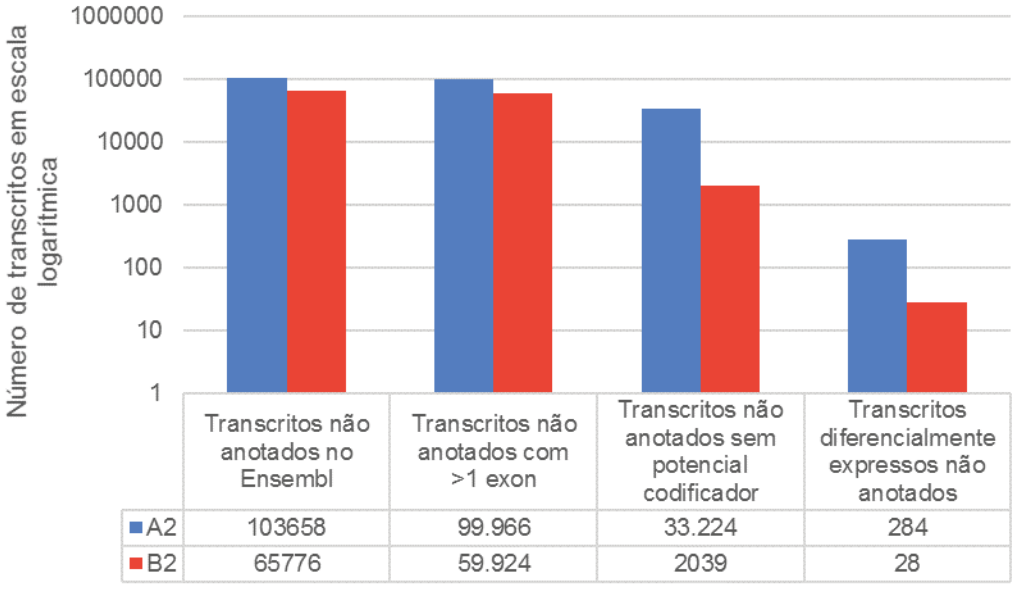


Figura 10: Redução da quantidade de transcritos identificados ao longo das etapas.

Fonte: a autora (2020).

Dos transcritos diferencialmente expressos encontrados nas abordagens A 20% e B 20%, obtivemos proporções diferentes da expressão dos mesmos em cada subtipo histológico de câncer cervical ADC e SCC (Figura 11). Dentre os transcritos mais expressos em SCC em A 20%, havia mais transcritos não anotados do que anotados nessa condição. Em contrapartida, para os transcritos mais expressos em SCC em B 20%, ocorreu o contrário: foram encontrados mais transcritos anotados do que não anotados (Figura 11).

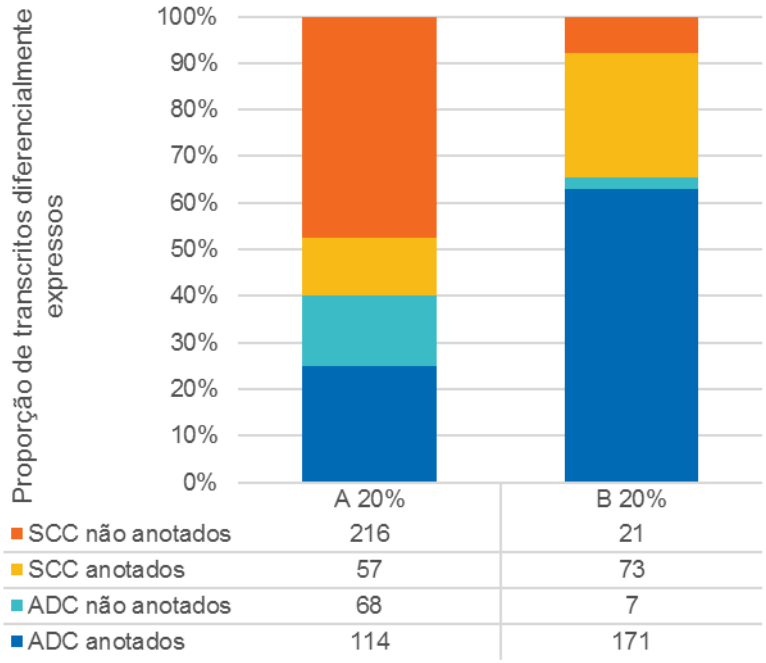


Figura 11: Quantidade de transcritos diferencialmente expressos nos subtipos de câncer cervical encontrados com as abordagens A 20% e B 20%.

Fonte: a autora (2020).

Dentre os transcritos diferencialmente expressos, para visualização nós construímos um *heatmap* selecionando os 50 genes mais diferencialmente expressos que apresentaram menor p-valor ajustado (também conhecido por FDR, do inglês “False Discovery Rate”), tanto na abordagem A 20% (Figura 12) quanto na abordagem B 20% (Figura 13). Os 50 genes mais diferencialmente expressos e seus respectivos valores de *fold-change*, p-valor e p-valor ajustado estão listados na Tabela 4 para abordagem A 20% e Tabela 5 para abordagem B 20%.

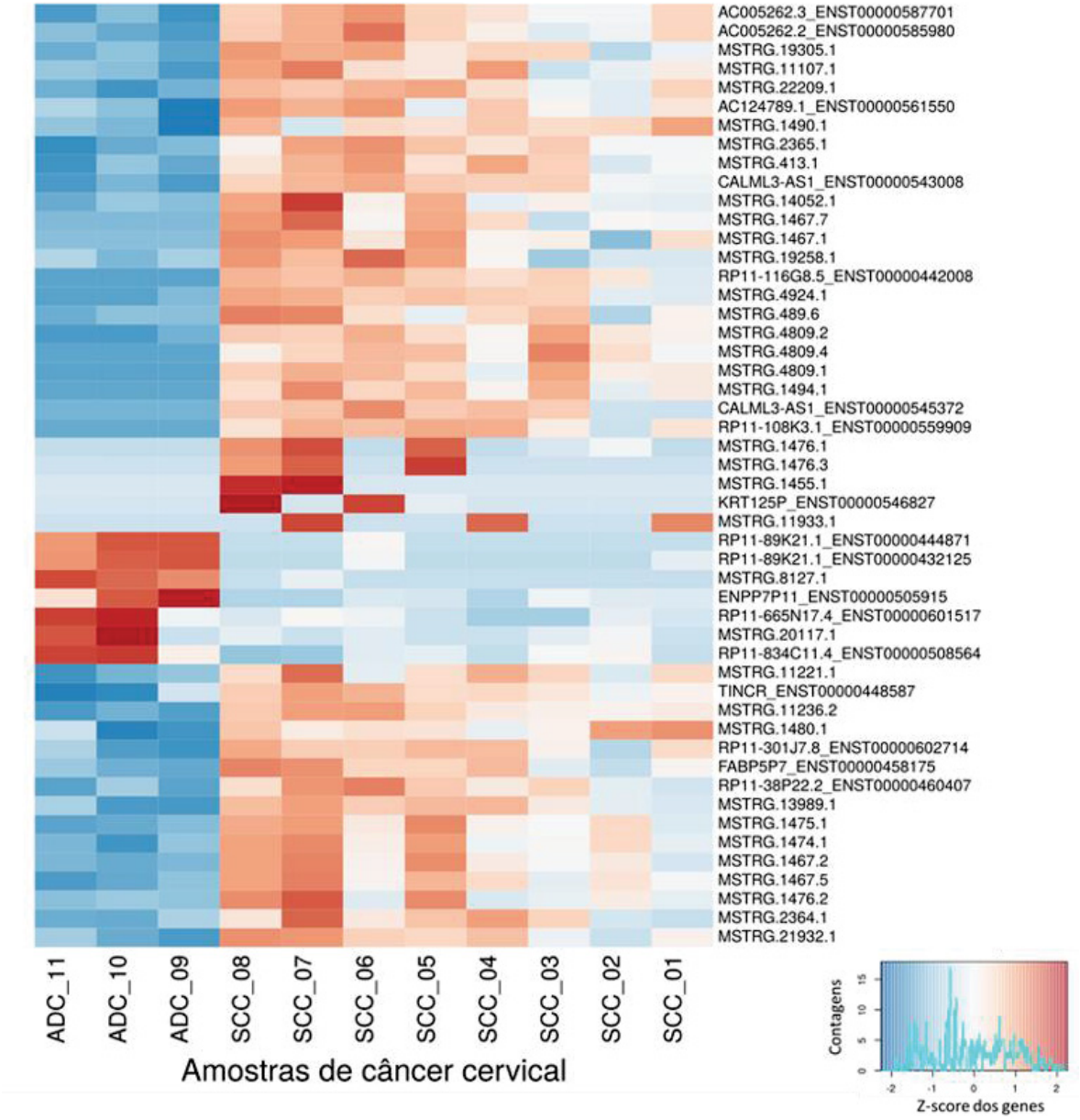


Figura 12: *Heatmap* da abordagem A 20%, representando os 50 genes mais diferencialmente expressos de menor p-valor ajustado.

Células avermelhadas indicam maior expressão e células azuladas indicam menor expressão, com variação de acordo com a intensidade de cor. Fonte: a autora (2020).

Tabela 4: 50 genes mais diferencialmente expressos de menor p-valor ajustado na abordagem A 20%, e seus respectivos valores de logFC, p-valor e p-valor ajustado.

Transcritos	logFC	p-valor	Taxa de Falso-positivo (p-valor ajustado ou FDR)
MSTRG.1476.1	23,92	1,98E-12	4,27E-09
MSTRG.1476.3	22,81	3,70E-11	3,71E-08
MSTRG.1455.1	22,06	1,57E-10	1,08E-07
KRT125P_ENST00000546827	20,83	1,52E-09	8,83E-07
MSTRG.11933.1	20,51	9,93E-11	7,51E-08
MSTRG.1467.1	12,58	1,05E-09	6,61E-07
MSTRG.1476.2	11,27	2,33E-11	2,51E-08
MSTRG.1474.1	11,02	1,30E-14	4,90E-11
MSTRG.1475.1	11,00	1,58E-16	7,92E-13
MSTRG.1467.7	10,59	1,21E-08	5,70E-06
MSTRG.1467.5	10,44	2,88E-14	8,69E-11
MSTRG.1467.2	10,25	7,09E-14	1,78E-10
RP11-116G8.5_ENST00000442008	10,09	2,92E-12	4,89E-09
MSTRG.2364.1	9,19	1,28E-11	1,49E-08
MSTRG.4924.1	9,15	1,33E-09	8,04E-07
MSTRG.4809.2	8,83	4,62E-10	3,03E-07
MSTRG.4809.4	8,65	5,40E-09	2,81E-06
MSTRG.19258.1	8,65	3,35E-07	1,07E-04
MSTRG.14052.1	8,41	1,44E-08	6,39E-06
MSTRG.413.1	8,23	5,49E-11	5,17E-08
CALML3-AS1_ENST00000543008	8,14	1,20E-17	9,07E-14
MSTRG.13989.1	8,14	1,03E-11	1,29E-08
MSTRG.11236.2	8,06	2,76E-19	4,16E-15
MSTRG.4809.1	7,74	2,54E-08	1,06E-05
CALML3-AS1_ENST00000545372	7,74	4,74E-07	1,42E-04
MSTRG.22209.1	7,51	8,82E-09	4,43E-06
MSTRG.1494.1	7,40	1,98E-07	7,29E-05
RP11-108K3.1_ENST00000559909	7,36	2,16E-07	7,61E-05
MSTRG.11221.1	7,09	7,55E-11	6,32E-08
MSTRG.2365.1	7,03	6,39E-12	9,63E-09
MSTRG.489.6	6,23	3,01E-07	9,87E-05
TINCR_ENST00000448587	6,07	4,98E-08	1,97E-05
MSTRG.21932.1	5,40	4,03E-09	2,17E-06
FABP5P7_ENST00000458175	5,38	2,14E-08	9,22E-06
MSTRG.11107.1	5,23	2,33E-07	7,99E-05
AC005262.3_ENST00000587701	5,07	7,58E-12	1,04E-08
MSTRG.1490.1	4,93	6,21E-11	5,51E-08

MSTRG.1480.1	4,84	2,17E-07	7,61E-05
MSTRG.19305.1	4,75	4,08E-08	1,66E-05
RP11-38P22.2_ENST00000460407	4,62	7,39E-08	2,86E-05
RP11-301J7.8_ENST00000602714	4,60	1,99E-09	1,11E-06
AC124789.1_ENST00000561550	4,52	4,07E-07	1,28E-04
AC005262.2_ENST00000585980	3,41	4,40E-07	1,35E-04
RP11-834C11.4_ENST00000508564	-4,45	1,14E-08	5,54E-06
RP11-665N17.4_ENST00000601517	-6,00	1,47E-07	5,55E-05
ENPP7P11_ENST00000505915	-6,28	1,35E-08	6,18E-06
RP11-89K21.1_ENST00000432125	-6,80	9,97E-11	7,51E-08
MSTRG.8127.1	-7,50	2,47E-12	4,66E-09
RP11-89K21.1_ENST00000444871	-7,93	1,37E-10	9,86E-08
MSTRG.20117.1	-8,71	2,67E-07	8,92E-05

Fonte: a autora (2020).

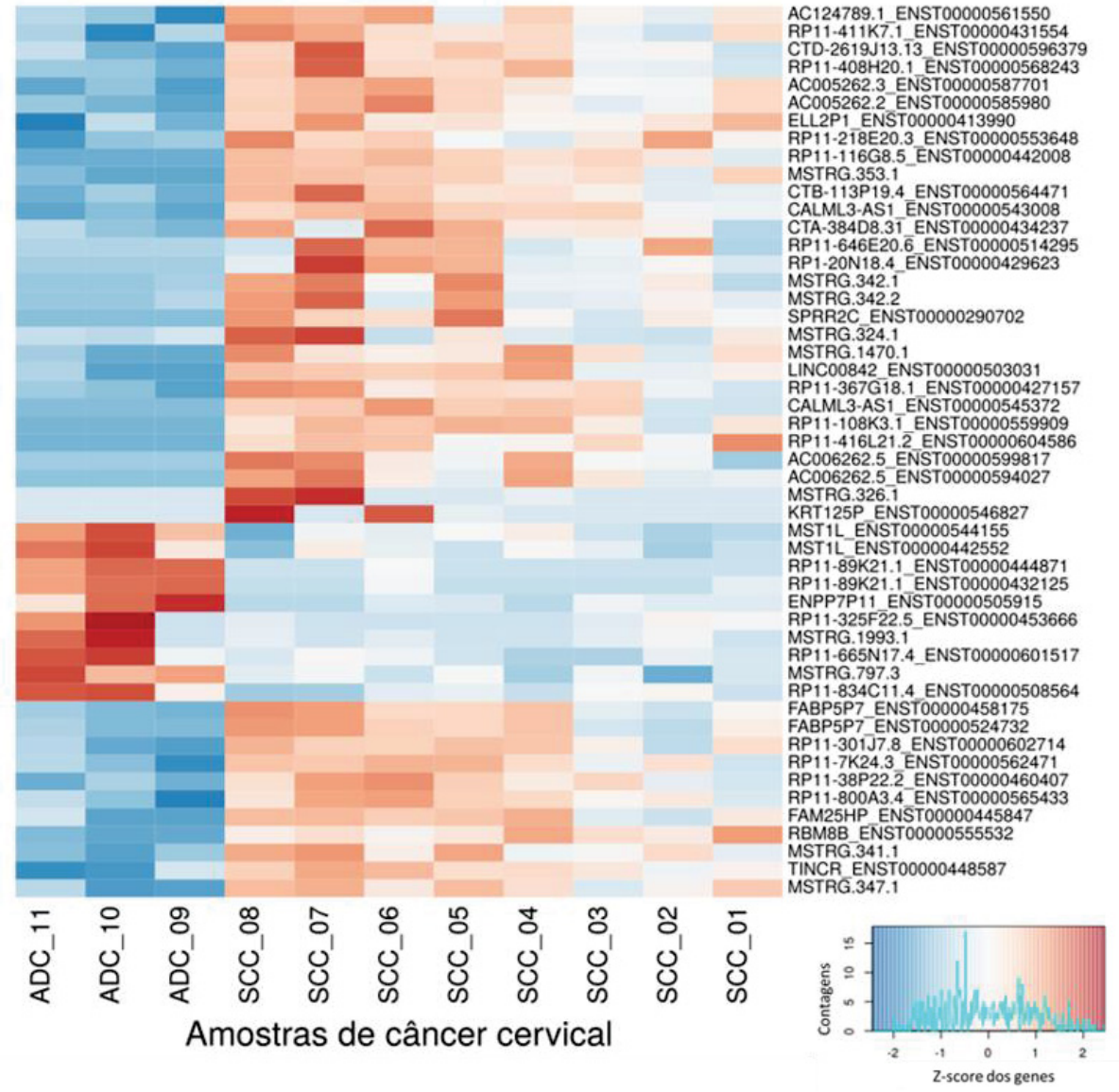


Figura 13: *Heatmap* da abordagem B 20%, representando os 50 genes mais diferencialmente expressos de menor p-valor ajustado.

Células avermelhadas indicam maior expressão e células azuladas indicam menor expressão, com variação de acordo com a intensidade de cor. Fonte: a autora (2020).

Tabela 5: 50 genes mais diferencialmente expressos de menor p-valor ajustado na abordagem B 20%, e seus respectivos valores de logFC, p-valor e p-valor ajustado

Transcritos	logFC	p-valor	Taxa de Falso-positivo (p-valor ajustado ou FDR)
MSTRG.1476.1	23,92	1,98E-12	4,27E-09
MSTRG.1476.3	22,81	3,70E-11	3,71E-08
MSTRG.1455.1	22,06	1,57E-10	1,08E-07
KRT125P_ENST00000546827	20,83	1,52E-09	8,83E-07
MSTRG.11933.1	20,51	9,93E-11	7,51E-08
MSTRG.1467.1	12,58	1,05E-09	6,61E-07
MSTRG.1476.2	11,27	2,33E-11	2,51E-08
MSTRG.1474.1	11,02	1,30E-14	4,90E-11
MSTRG.1475.1	11,00	1,58E-16	7,92E-13
MSTRG.1467.7	10,59	1,21E-08	5,70E-06
MSTRG.1467.5	10,44	2,88E-14	8,69E-11
MSTRG.1467.2	10,25	7,09E-14	1,78E-10
RP11-116G8.5_ENST00000442008	10,09	2,92E-12	4,89E-09
MSTRG.2364.1	9,19	1,28E-11	1,49E-08
MSTRG.4924.1	9,15	1,33E-09	8,04E-07
MSTRG.4809.2	8,83	4,62E-10	3,03E-07
MSTRG.4809.4	8,65	5,40E-09	2,81E-06
MSTRG.19258.1	8,65	3,35E-07	1,07E-04
MSTRG.14052.1	8,41	1,44E-08	6,39E-06
MSTRG.413.1	8,23	5,49E-11	5,17E-08
CALML3-AS1_ENST00000543008	8,14	1,20E-17	9,07E-14
MSTRG.13989.1	8,14	1,03E-11	1,29E-08
MSTRG.11236.2	8,06	2,76E-19	4,16E-15
MSTRG.4809.1	7,74	2,54E-08	1,06E-05
CALML3-AS1_ENST00000545372	7,74	4,74E-07	1,42E-04
MSTRG.22209.1	7,51	8,82E-09	4,43E-06
MSTRG.1494.1	7,40	1,98E-07	7,29E-05
RP11-108K3.1_ENST00000559909	7,36	2,16E-07	7,61E-05
MSTRG.11221.1	7,09	7,55E-11	6,32E-08
MSTRG.2365.1	7,03	6,39E-12	9,63E-09
MSTRG.489.6	6,23	3,01E-07	9,87E-05
TINCR_ENST00000448587	6,07	4,98E-08	1,97E-05
MSTRG.21932.1	5,40	4,03E-09	2,17E-06
FABP5P7_ENST00000458175	5,38	2,14E-08	9,22E-06
MSTRG.11107.1	5,23	2,33E-07	7,99E-05
AC005262.3_ENST00000587701	5,07	7,58E-12	1,04E-08
MSTRG.1490.1	4,93	6,21E-11	5,51E-08
MSTRG.1480.1	4,84	2,17E-07	7,61E-05

MSTRG.19305.1	4,75	4,08E-08	1,66E-05
RP11-38P22.2_ENST00000460407	4,62	7,39E-08	2,86E-05
RP11-301J7.8_ENST00000602714	4,60	1,99E-09	1,11E-06
AC124789.1_ENST00000561550	4,52	4,07E-07	1,28E-04
AC005262.2_ENST00000585980	3,41	4,40E-07	1,35E-04
RP11-834C11.4_ENST00000508564	-4,45	1,14E-08	5,54E-06
RP11-665N17.4_ENST00000601517	-6,00	1,47E-07	5,55E-05
ENPP7P11_ENST00000505915	-6,28	1,35E-08	6,18E-06
RP11-89K21.1_ENST00000432125	-6,80	9,97E-11	7,51E-08
MSTRG.8127.1	-7,50	2,47E-12	4,66E-09
RP11-89K21.1_ENST00000444871	-7,93	1,37E-10	9,86E-08
MSTRG.20117.1	-8,71	2,67E-07	8,92E-05

Fonte: a autora (2020).

Na abordagem A 20%, dos 50 genes mais diferencialmente expressos, a maioria eram genes não anotados. Somente 17 genes foram encontrados anotados na referência do Ensembl, contra 33 transcritos não anotados potencialmente novos (Figura 12 e Tabela 6).

Em contrapartida, na abordagem B 20%, dos 50 genes mais diferencialmente expressos, a maior parte eram genes já anotados na referência Ensembl. Foram encontrados 40 genes anotados previamente contra 10 transcritos não anotados potencialmente novos (Figura 13 e Tabela 6). A Figura 14 apresenta visualmente a proporção de genes anotados e não anotados que foram encontrados nos heatmap das abordagens A 20% e B 20%.

Tabela 6: Genes anotados e potencialmente novos dentre os 50 genes mais diferencialmente expressos das abordagens A 20% e B 20%

	Genes anotados no Ensembl	**Genes potencialmente novos**
A 20%	17	33
B 20%	40	10

Fonte: a autora (2020).

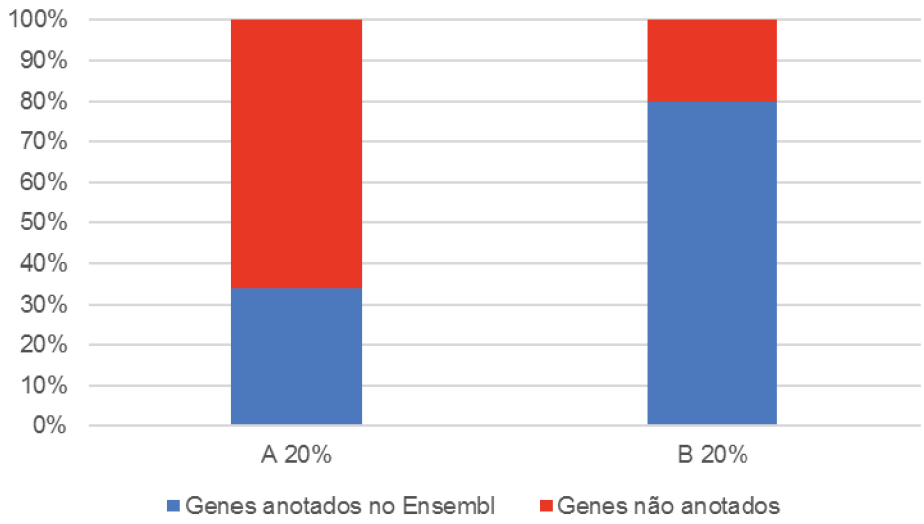


Figura 14: Proporção de genes anotados e não anotados dentre os 50 genes diferencialmente expressos dos heatmaps das abordagens A 20% e B 20%.

Fonte: a autora (2020).

Também observamos os mesmos genes ou transcritos sendo encontrados diferencialmente expressos em ambas as abordagens A 20% e B 20%. Dentre os 50 genes diferencialmente expressos apresentados nos heatmap das abordagens A 20% e B 20%, 17 transcritos já anotados apareceram em ambas as abordagens. Estes genes estão listados na Tabela 7, assim como seus respectivos valores de logFC, p-valor e p-valor ajustado (Tabela 7).

Tabela 7: Dezessete transcritos dentre os 50 mais diferencialmente expressos em comum nas duas abordagens A 20% e B 20%, com seus respectivos valores de logFC, p-valor e p-valor ajustado.

	A 20%			B 20%		
Transcritos em comum nas duas abordagens	**logFC**	**p-valor**	**Taxa de Falso-positivo (p-valor ajustado ou FDR)**	**logFC**	**p-valor**	**Taxa de Falso-positivo (p-valor ajustado ou FDR)**
AC005262.2_ENST00000585980	3,41	4,40E-07	1,35E-04	3,40	3,12E-07	1,17E-04
AC005262.3_ENST00000587701	5,07	7,58E-12	1,04E-08	5,07	3,96E-12	8,01E-09
AC124789.1_ENST00000561550	4,52	4,07E-07	1,28E-04	4,51	2,76E-07	1,07E-04
CALML3-AS1_ENST00000543008	8,14	1,20E-17	9,07E-14	8,14	5,13E-18	5,18E-14
CALML3-AS1_ENST00000545372	7,74	4,74E-07	1,42E-04	7,74	3,70E-07	1,33E-04
ENPP7P11_ENST00000505915	-6,28	1,35E-08	6,18E-06	-6,26	9,16E-09	6,61E-06
FABP5P7_ENST00000458175	5,38	2,14E-08	9,22E-06	5,36	1,28E-08	8,10E-06
KRT125P_ENST00000546827	20,83	1,52E-09	8,83E-07	20,83	1,53E-09	1,29E-06
RP11-108K3.1_ENST00000559909	7,36	2,16E-07	7,61E-05	7,36	1,82E-07	7,65E-05
RP11-116G8.5_ENST00000442008	10,09	2,92E-12	4,89E-09	10,09	2,09E-12	5,68E-09
RP11-301J7.8_ENST00000602714	4,60	1,99E-09	1,11E-06	4,59	1,17E-09	1,07E-06
RP11-38P22.2_ENST00000460407	4,62	7,39E-08	2,86E-05	4,62	4,93E-08	2,49E-05

RP11-665N17.4_ENST00000601517	-6,00	1,47E-07	5,55E-05	-6,00	8,90E-08	4,28E-05
RP11-834C11.4_ENST00000508564	-4,45	1,14E-08	5,54E-06	-4,44	1,16E-08	7,83E-06
RP11-89K21.1_ENST00000432125	-6,80	9,97E-11	7,51E-08	-6,79	4,73E-11	6,84E-08
RP11-89K21.1_ENST00000444871	-7,93	1,37E-10	9,86E-08	-7,92	1,87E-11	3,15E-08
TINCR_ENST00000448587	6,07	4,98E-08	1,97E-05	6,09	2,55E-08	1,43E-05

Fonte: a autora (2020).

2.4 Visualização dos genes mais diferencialmente expressos

A fim de conferir mais detalhadamente alguns dos 50 transcritos mais diferencialmente expressos, visualizamos no programa IGV os alinhamentos das onze amostras de câncer cervical juntamente com as anotações criadas:

- A 20% concatenada com Ensembl, de genes não codificadores já anotados e também potencialmente novos;
- B 20% concatenada com Ensembl, de genes não codificadores já anotados e também potencialmente novos.

Dos 50 transcritos mais diferencialmente expressos na abordagem **A 20%**, visualizamos individualmente no programa IGV os seguintes:

Dois transcritos mais expressos em ADC:
- **MSTRG.20117.1** (Tabela 8, Figura 15)
- **MSTRG.8127.1** (Tabela 8, Figura 16)

Três transcritos mais expressos em SCC:
- **MSTRG.1476.1** (Tabela 8, Figura 17)
- **MSTRG.1476.2** (Tabela 8, Figura 17)
- **MSTRG.1476.3** (Tabela 8, Figura 17)

Os valores de logFC, de p-valor e de p-valor ajustado desses transcritos estão listados na Tabela 8.

Tabela 8: Transcritos escolhidos para serem analisados individualmente, dentre os 50 mais diferencialmente expressos da abordagem A 20%.

Transcrito escolhido	logFC	p-valor	p-valor ajustado
MSTRG.20117.1	-8,71	2,67E-07	8,92E-05
MSTRG.8127.1	-7,50	2,47E-12	4,66E-09
MSTRG.1476.1	23,92	1,98E-12	4,27E-09
MSTRG.1476.2	11,27	2,33E-11	2,51E-08
MSTRG.1476.3	22,81	3,70E-11	3,71E-08

Fonte: a autora (2020).

O transcrito **MSTRG.20117.1** aparece na lista dos 50 genes mais diferencialmente expressos da abordagem A 20%, conforme visto no *heatmap* da Figura 12 e nos valores de logFC, p-valor e p-valor ajustado da Tabela 8. Com a visualização no IGV pudemos ver que esse mesmo transcrito também é identificado no *heatmap* de 50 genes mais diferencialmente expressos da abordagem B 20%, porém com o identificador MSTRG.1993.1 (Figura 13 e Tabela 5). O alinhamento das leituras nas anotações de ambas as abordagens podem ser vistas na Figura 15.

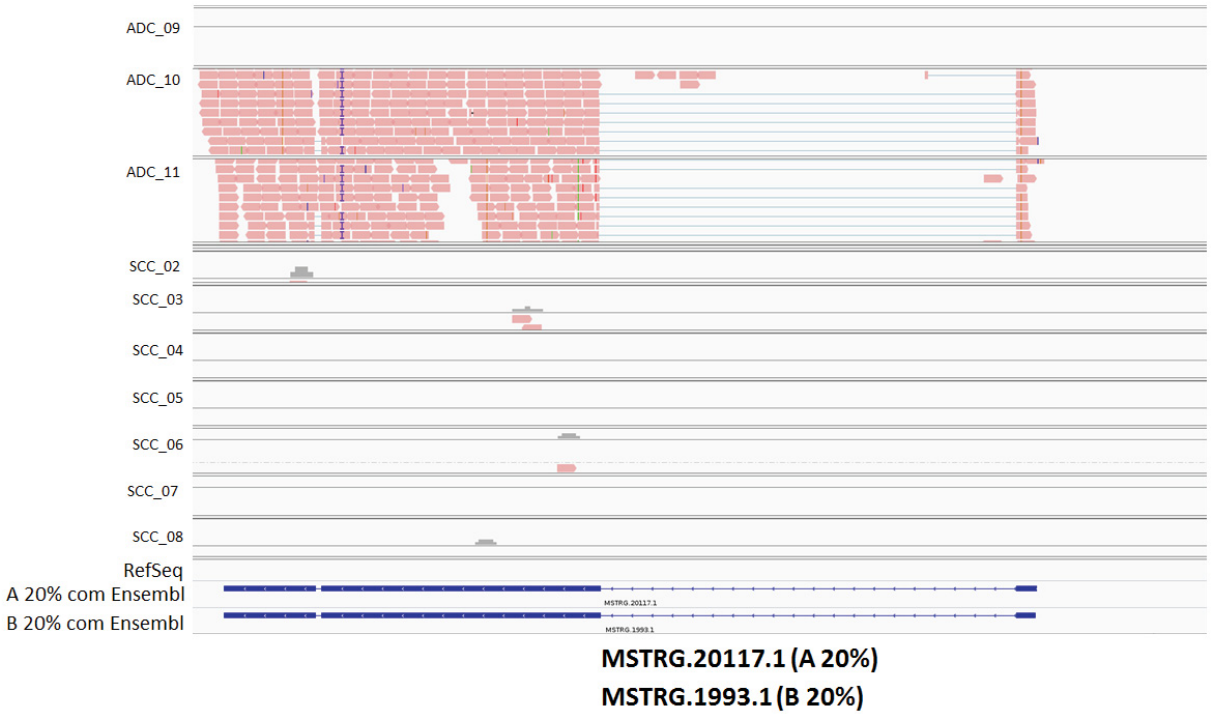


Figura 15: Transcrito diferencialmente expresso MSTRG.20117.1 ou MSTRG.1993.1 identificado nas abordagens A 20% e B 20% respectivamente.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços vermelhos são leituras pareadas anti-senso.

Fonte: a autora (2020).

O transcrito **MSTRG.8127.1** foi encontrado entre os 50 genes mais diferencialmente expressos da abordagem A 20%, conforme visto no *heatmap* da Figura 12 e nos valores de logFC, p-valor e p-valor ajustado da Tabela 8. A abordagem B 20% não retornou nenhuma anotação para esse mesmo *locus*, assim como a anotação de RefSeq também não. O alinhamento das leituras sobre a anotação A 20% pode ser visto na Figura 16.

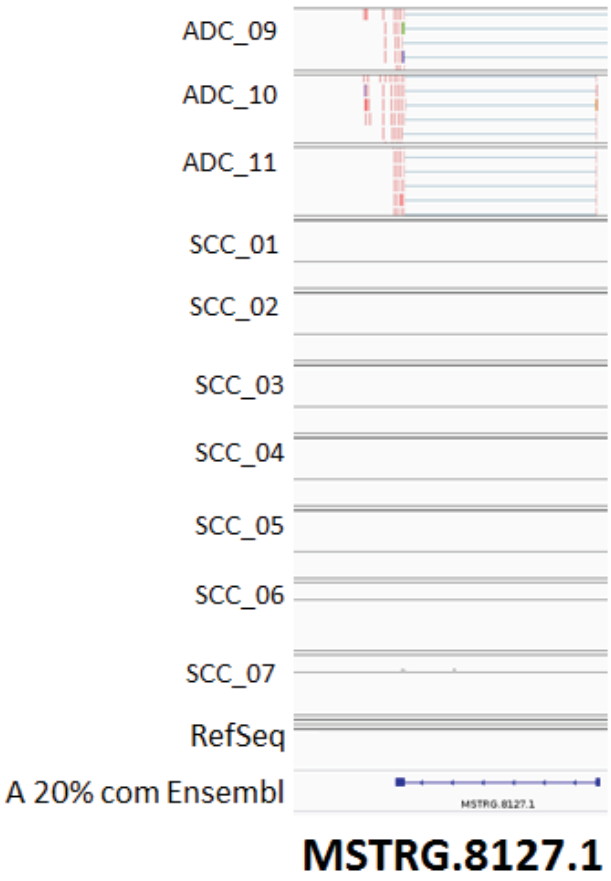


Figura 16: Transcrito MSTRG.8127.1 diferencialmente expresso na abordagem A 20%.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços vermelhos são leituras pareadas anti-senso.

Fonte: a autora (2020).

O gene **MSTRG.1476** e seus **transcritos 1, 2 e 3** apareceram na lista dos 50 genes mais diferencialmente expressos da abordagem A 20%, conforme visto no *heatmap* da Figura 12 e nos valores de logFC, p-valor e p-valor ajustado da Tabela 8. Através da visualização no IGV pudemos ver que esses mesmos transcritos são os identificados na abordagem B 20% como MSTRG.342.1 e MSTRG.342.2 e MSTRG.342.3, e inclusive eles também estão entre os 50 genes mais diferencialmente expressos da abordagem B 20% (Figura 13 e Tabela 5). Estão alinhados nos *locus* gênicos de SPRR2B, SPRR2E, SPRR2F e SPRR2C anotados no RefSeq. Com a anotação do RefSeq podemos ver que o transcrito MSTRG.1476.1 é um falso positivo, pois corresponde ao gene já anotado SPRR2B. Esses transcritos e seus alinhamentos podem ser conferidos na Figura 17.

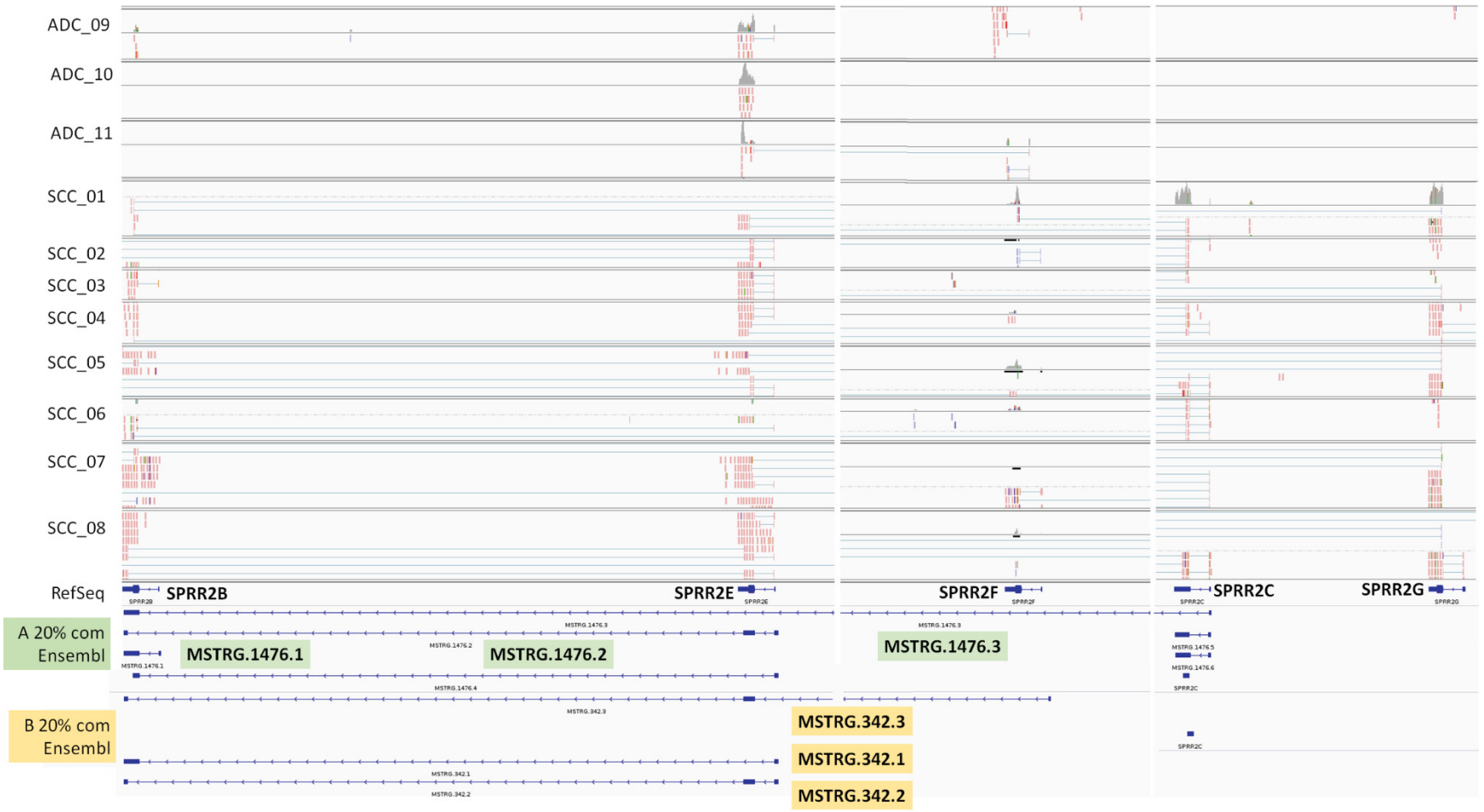


Figura 17: Genes MSTRG.1476 e MSTRG.342 não anotados no Ensembl, presentes nos 50 genes mais diferencialmente expressos nas abordagens A 20% e B 20% respectivamente.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços vermelhos são leituras pareadas anti-senso.

Fonte: A autora (2020).

A respeito da **abordagem B 20%**, escolhemos alguns dos 50 genes mais diferencialmente expressos (Tabela 5) encontrados nessa abordagem para a visualização com o programa IGV:

- **MSTRG.797.3** (Tabela 9, Figura 18);
- **MSTRG.347.1** (Tabela 9, Figura 19);

Os valores de logFC, p-valor e p-valor ajustado desses transcritos escolhidos estão listados na Tabela 9.

Tabela 9: Transcritos escolhidos para serem analisados individualmente, a partir dos 50 transcritos diferencialmente expressos de menor p-valor ajustado da abordagem B 20%

Transcrito escolhido	logFC	p-valor	p-valor ajustado
MSTRG.347.1	6,74	2,25E-12	5,68E-09
MSTRG.797.3	-2,22	9,80E-07	2,36E-04

Fonte: a autora (2020).

O transcrito **MSTRG.797.3** (Tabela 9) está entre os 50 genes mais diferencialmente expressos da abordagem B 20%. A visualização da anotação no IGV mostrou que na abordagem A 20% esse transcrito foi identificado como MSTRG.1935.6. Porém, MSTRG.1935.6 não está diferencialmente expresso na abordagem A 20%. O alinhamento de suas leituras e as anotações A 20% e B 20% podem ser vistas na Figura 18.

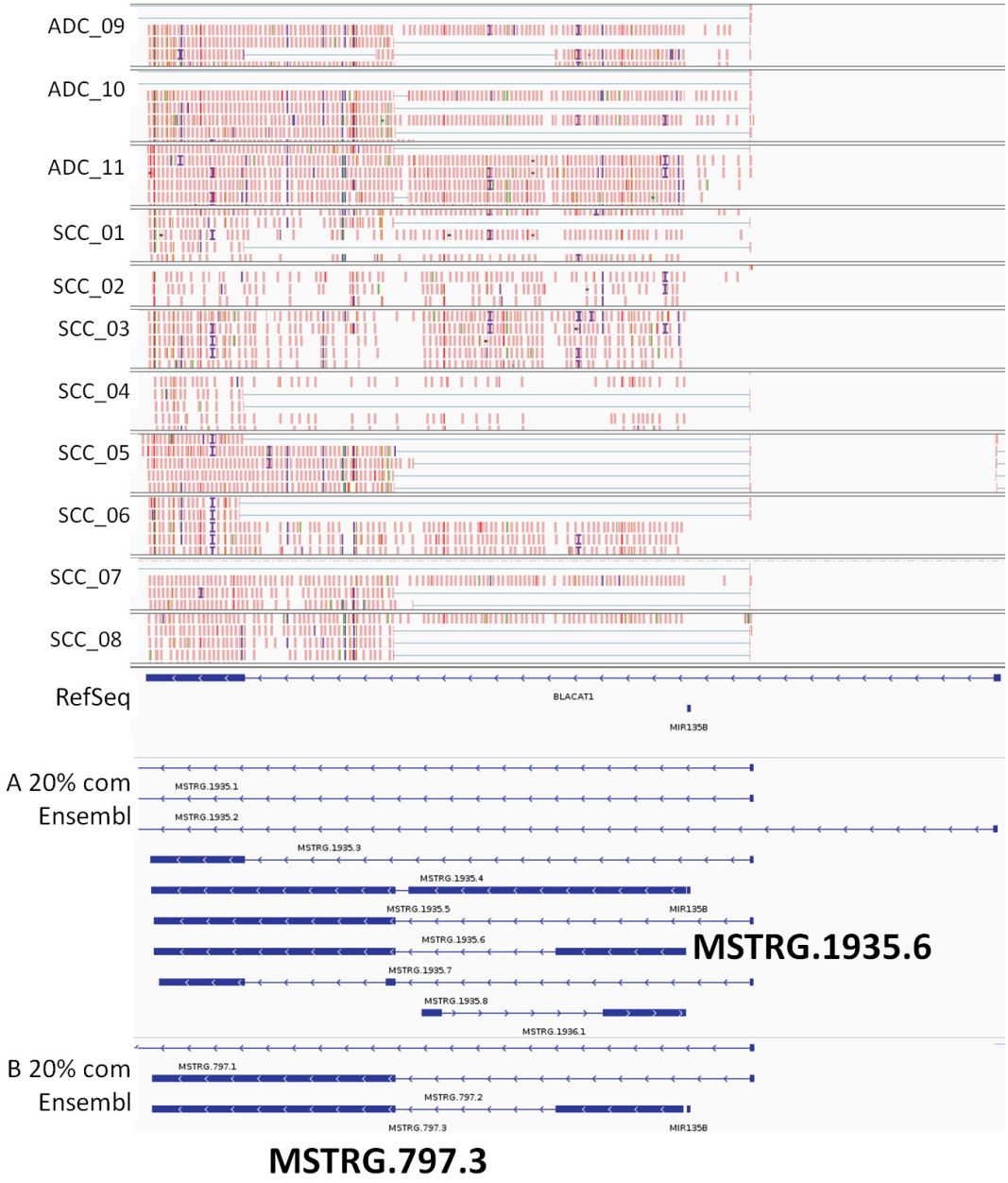


Figura 18: Transcrito MSTRG.797.3 entre os 50 genes mais diferencialmente expressos na abordagem B 20%, é também identificado como MSTRG.1935.6 na abordagem A 20%, não estando diferencialmente expresso em A 20%.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços vermelhos são leituras pareadas anti-senso. Fonte: a autora (2020).

O transcrito **MSTRG.347.1** (Tabela 9) está presente nos 50 genes mais diferencialmente expressos da abordagem B 20%. Sua visualização no IGV mostrou que esse gene não está presente na anotação do Ensembl, porém está presente na anotação de referência RefSeq como S100A8, conforme visto na Figura 19.

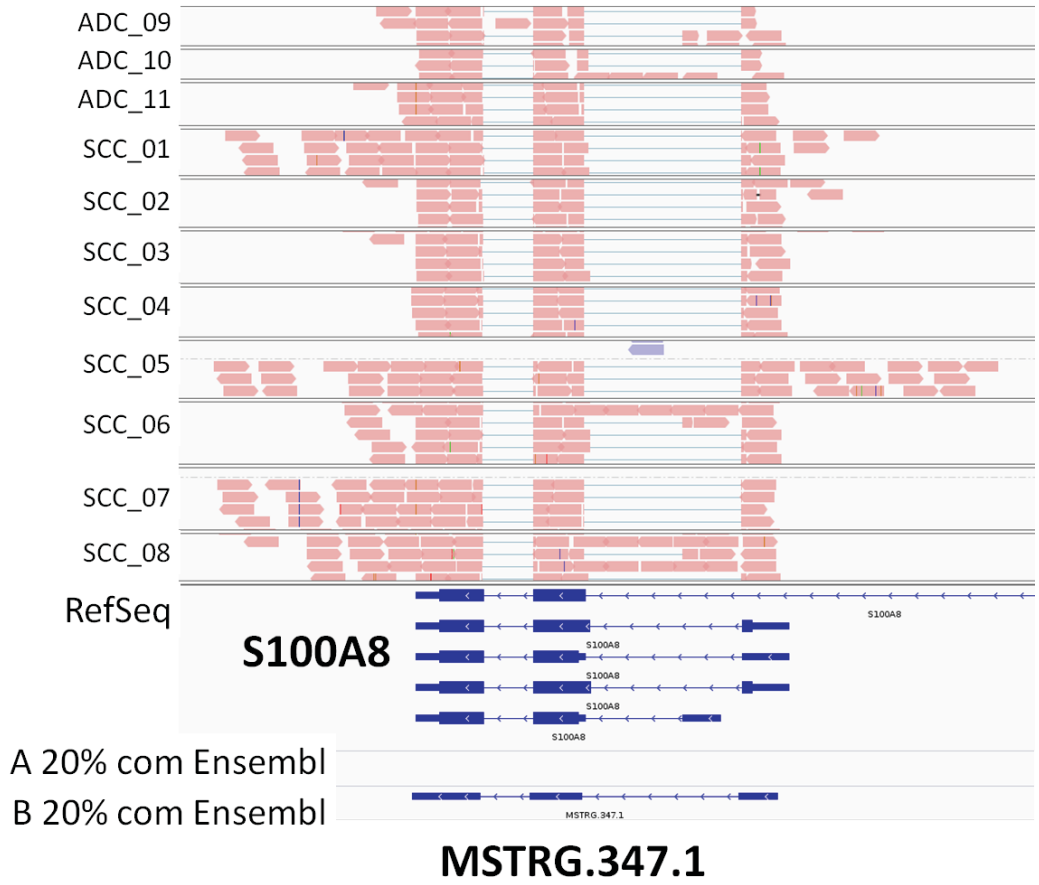


Figura 19: Transcrito MSTRG.347.1 não anotado no Ensembl e anotado no RefSeq como S100A8. Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços vermelhos são leituras pareadas anti-senso, e traços azuis claros são leituras senso. Fonte: a autora (2020).

Dos 17 genes que apareceram nos 50 genes mais diferencialmente expressos em **ambas as abordagens A 20% e B 20%**, escolhemos os seguintes genes para conferir as anotações e os alinhamentos:

- **TINCR** (Tabela 10, Figura 20 e Figura 21);
- **CALML3-AS1** (transcritos ENST00000543008 e ENST00000545372, Tabela 10, Figura 22 e Figura 23);
- **RP11-89K21.1** (transcritos ENST00000432125 e ENST00000444871, Tabela 10, Figura 24).

Seus valores de logFC, p-valor e p-valor ajustado estão listados na Tabela 10.

Tabela 10: Transcritos escolhidos para análise individual, encontrados entre os 50 mais diferencialmente expressos de ambas as abordagens A 20% e B 20%, com seus respectivos valores de logFC, p-valor e p-valor ajustado.

Transcritos em comum nas duas abordagens	A 20%			B 20%		
	logFC	p-valor	p-valor ajustado	logFC	p-valor	p-valor ajustado
TINCR ENST00000448587	6,07	4,98E-08	1,97E-05	6,09	2,55E-08	1,43E-05
CALML3-AS1 ENST00000543008	8,14	1,20E-17	9,07E-14	8,14	5,13E-18	5,18E-14
CALML3-AS1 ENST00000545372	7,74	4,74E-07	1,42E-04	7,74	3,70E-07	1,33E-04
RP11-89K21.1 ENST00000432125	-6,80	9,97E-11	7,51E-08	-6,79	4,73E-11	6,84E-08
RP11-89K21.1 ENST00000444871	-7,93	1,37E-10	9,86E-08	-7,92	1,87E-11	3,15E-08

Fonte: a autora (2020).

Para o gene **TINCR** (Figura 20 e Figura 21), encontramos os mesmos três transcritos já anotados no Ensembl tanto na abordagem B 20% quanto na abordagem A 20%. Especialmente o transcrito ENST00000448587 do gene TINCR foi encontrado mais diferencialmente expresso em ambas as abordagens (Tabela 10). A relação dos transcritos pode ser vista na Figura 20, e as leituras alinhadas na anotação podem ser conferidas na Figura 21.

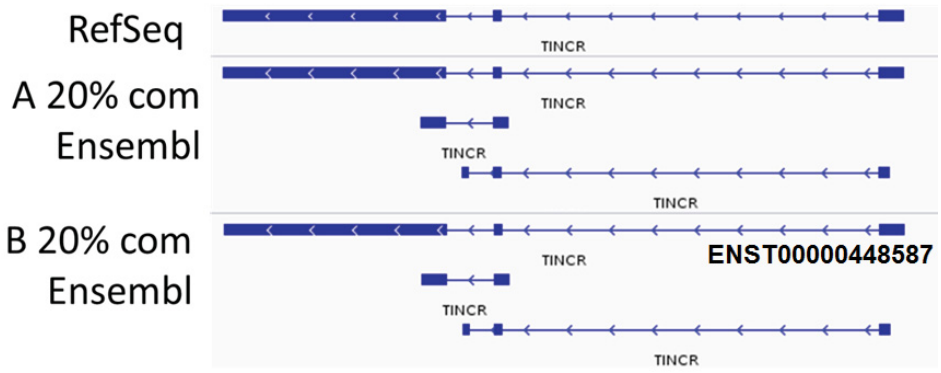


Figura 20: Transcritos de TINCR diferencialmente expressos em ambas as abordagens A 20% e B 20%.

Na anotação, exon são os retângulos azuis conectados pelas linhas que são os íntrons.

Fonte: a autora (2020).

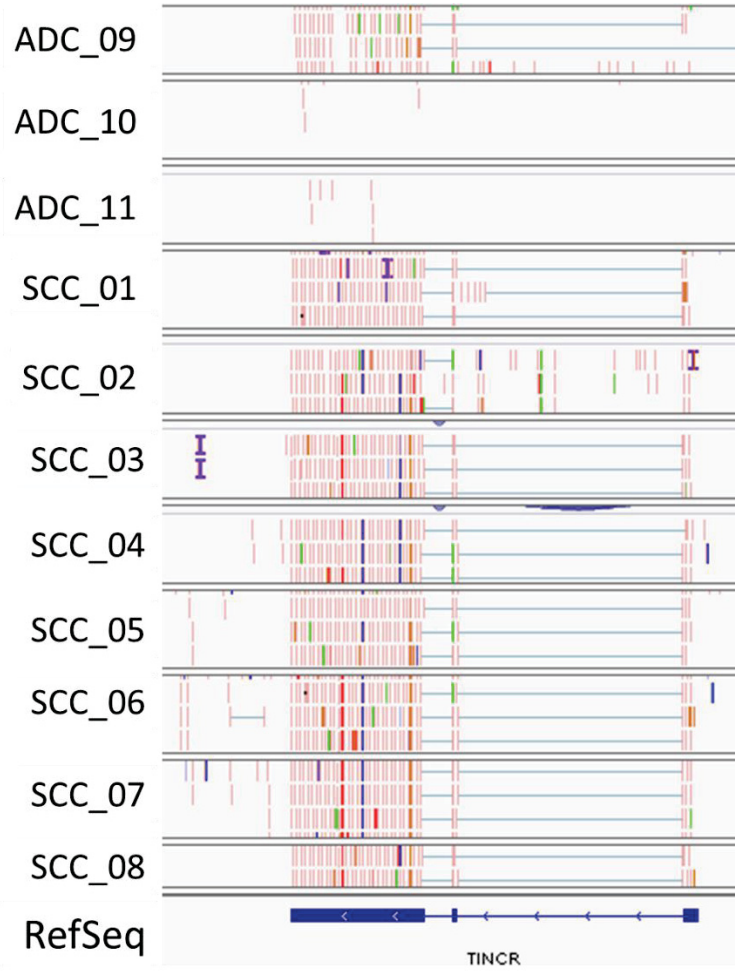


Figura 21: Leituras alinhadas sobre TINCR, gene diferencialmente expresso encontrado nas duas abordagens A 20% e B 20%.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços vermelhos são leituras pareadas anti-senso.

Fonte: a autora (2020).

Para o gene **CALML3-AS1** (Figura 22), em ambas as abordagens A 20% e B 20% foram identificados 4 de seus transcritos já anotados na referência Ensembl (Figura 22). Dois transcritos em específico foram encontrados como diferencialmente expressos em ambas as abordagens, sendo eles de identificadores ENST00000543008 e ENST00000545372 (Tabela 10, Figura 22). Na abordagem A 20% e adjacente ao gene CALML3-AS1 também pode ser identificado um transcrito não anotado na referência, de identificador **MSTRG.2365.1** (Figura 22). Este transcrito não anotado MSTRG.2365.1, adjacente ao gene CALML3-AS1, apresentou muitas leituras alinhadas à sua sequência, conforme visto na Figura 23. Entretanto, ele não está

presente nos 50 genes diferencialmente expressos da abordagem A 20%, e por isso não foi mencionado anteriormente. Porém, esse transcrito está diferencialmente expresso na abordagem A 20%, com valores de logFC de 7,03, p-valor de 6,39E-12, e p-valor ajustado de 9,63E-09.

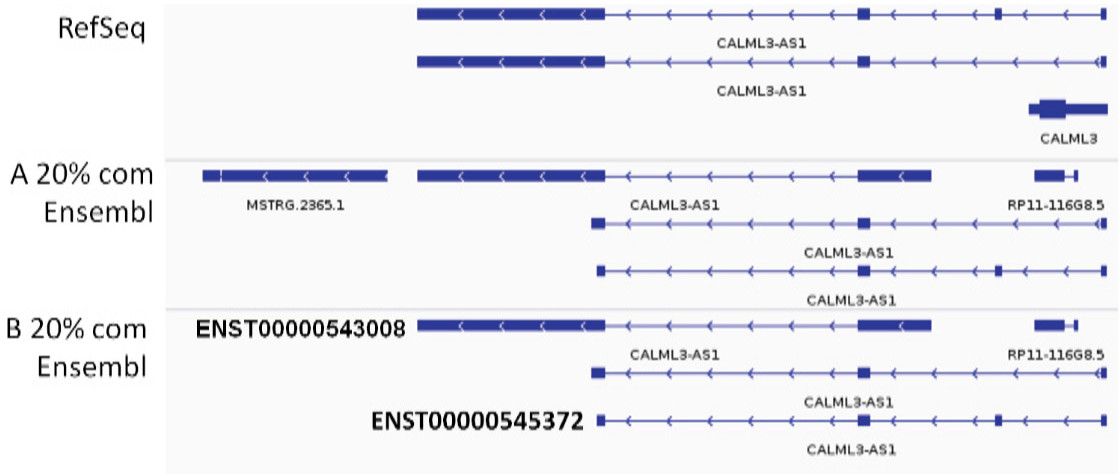


Figura 22: Transcritos de CALML3-AS1 (ENST00000543008 e ENST00000545372) diferencialmente expressos nas abordagens A 20% e B 20%.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons.

Fonte: a autora (2020).

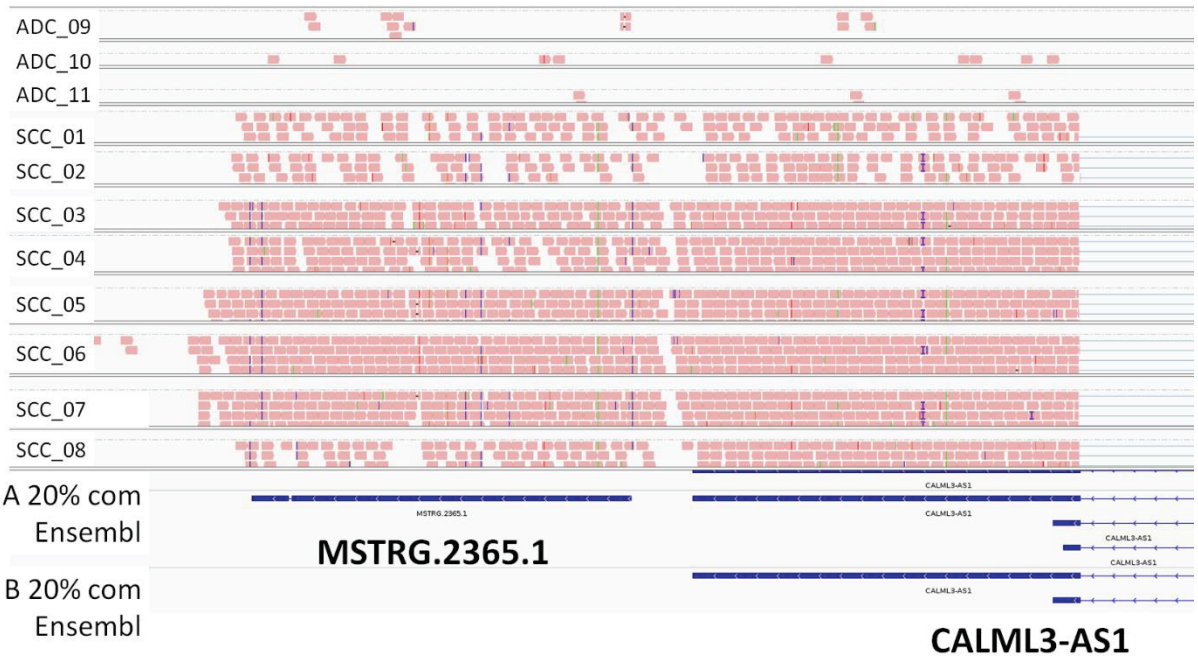


Figura 23: Transcrito não anotado MSTRG.2365.1 adjacente ao gene CALML-AS1, encontrado com a abordagem A 20%.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os traços avermelhados são leituras pareadas anti-senso.

Fonte: a autora (2020).

A respeito do gene **RP11-89K21.1**, em ambas as abordagens A 20% e B 20% foram encontrados os mesmos 6 transcritos já anotados na referência Ensembl (Figura 24). Na referência RefSeq esse gene é anotado como LINC01833. Dois transcritos diferentes do mesmo gene RP11-89K21.1 foram identificados como diferencialmente expressos em ambas as abordagens A 20% e B 20%, sendo eles ENST00000432125 e ENST00000444871 (Tabela 10). O alinhamento das leituras e as anotações podem ser vistas na Figura 24.

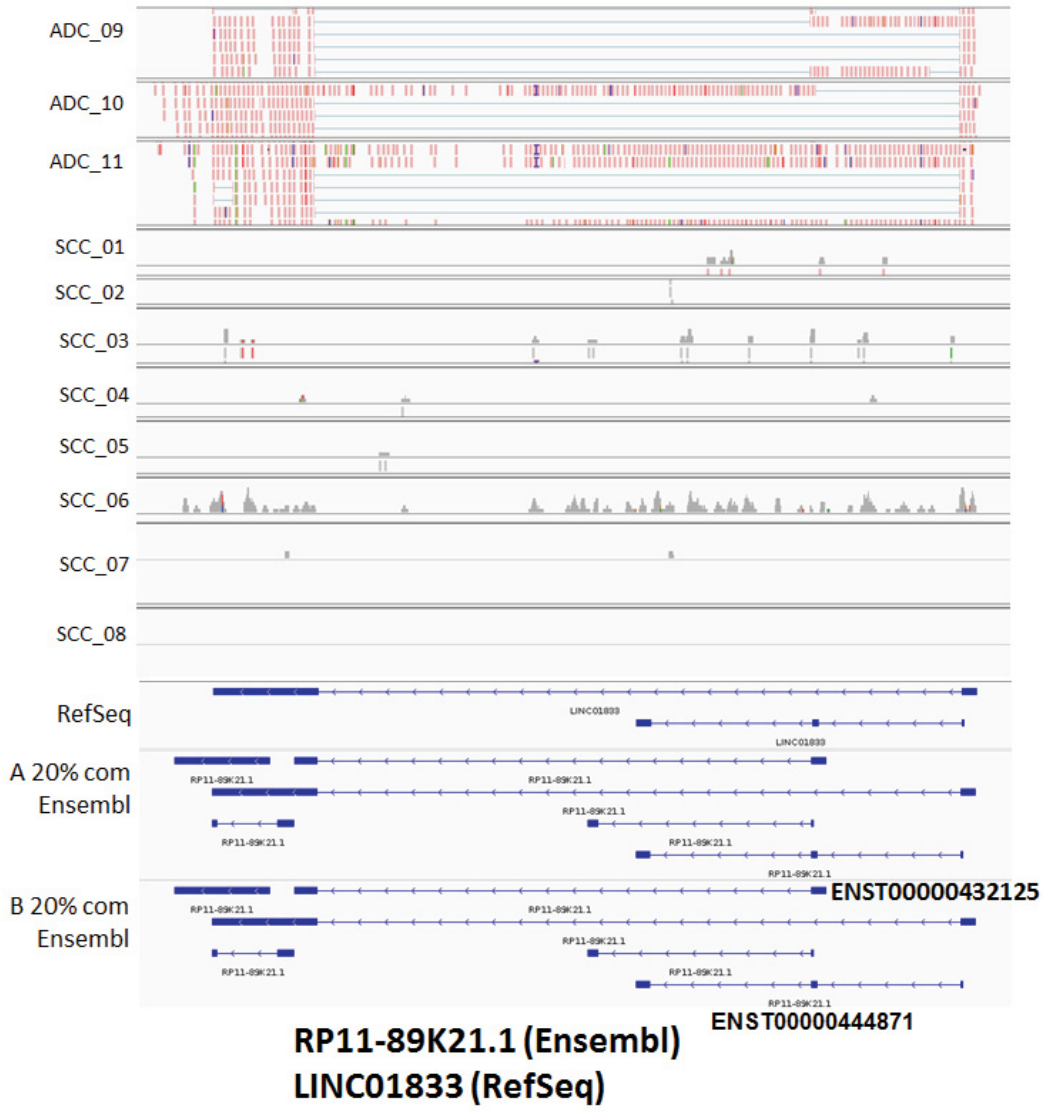


Figura 24: Gene RP11-89K21.1 (LINC01833) diferencialmente expresso em ambas as abordagens A 20% e B 20%, conforme visualização no IGV.

Na anotação, exons são os retângulos azuis conectados pelas linhas que são os íntrons. Os pequenos traços avermelhados são leituras anti-senso.

Fonte: a autora (2020).

3 DISCUSSÃO

Nesse trabalho, nós investigamos a presença de variantes de *splicing* alternativo de genes não codificadores em tumores cervicais dos tipos ADC e SCC. Especialmente, buscamos identificar variantes não anotadas na referência do Ensembl (ZERBINO et al., 2018), e portanto transcritos propostos como potencialmente novos. Nós nos propusemos a executar o identificador de variantes de *splicing* CLASS2 (SONG; SABUNCIYAN; FLOREA, 2016) de duas maneiras, para analisar qual seria a abordagem ideal para buscar esses transcritos em mais de um arquivo de alinhamento. Os dois métodos serão discutidos especificamente no tópico **3.2 Diferentes abordagens para identificar transcritos alternativos.**

3.1 Controle de qualidade e amostragem no RNA-seq

Ainda é um desafio a correta construção de transcritos derivados de um sequenciamento de RNA de pequenas leituras (VENTURINI et al., 2018). Com a popularização da análise de RNA-seq, surgiram muitas variações na maneira de analisar os dados de sequenciamento (CONESA et al., 2016). Cientistas podem adaptar os protocolos para seus objetos de estudo e seus próprios objetivos com a pesquisa (CONESA et al., 2016; FINOTELLO; DI CAMILLO, 2015). De todas as maneiras devem ser feitas verificações constantes a respeito da qualidade dos dados que foram gerados e estão sendo manipulados, a fim de assegurar reprodutibilidade e confiabilidade dos seus resultados (CONESA et al., 2016).

A avaliação inicial da qualidade do sequenciamento se dá pela quantidade de leituras mapeadas no genoma humano (CONESA et al., 2016). Espera-se que a proporção de leituras mapeadas no genoma humano atinja entre 70% a 90% de cobertura (CONESA et al., 2016). Nosso sequenciamento atingiu essa quantidade de leituras mapeadas desejadas para tal, o que garantiu a qualidade dos dados para a continuação da pesquisa e confiabilidade dos resultados (CONESA et al., 2016).

Para um projeto de anotação é indicado trabalhar com múltiplos conjuntos de transcriptomas, utilizando diferentes amostras, diferentes parâmetros de execução, ou diferentes métodos de montagem de transcritos (VENTURINI et al., 2018). Além disso, ao utilizar mais amostras de sequenciamento, é possível anotar uma maior quantidade de transcritos que podem estar expressos em diferentes amostragens (VENTURINI et al., 2018). Consequentemente, dessa forma irá ser feita uma anotação mais completa do referido objeto de estudo (VENTURINI et al., 2018). Nós escolhemos trabalhar com onze amostras de câncer cervical, sendo que a maioria dos tumores eram diagnosticados como SCC. O subtipo ADC é realmente o mais raro dos tumores cervicais (GADDUCCI; GUERRIERI; COSIO, 2019), e consequentemente a coleta de amostras desse tipo é mais limitada. Entretanto o número mínimo de três réplicas biológicas para ADC foi atingido para podermos estudar esse tipo histológico com confiabilidade na pesquisa (CONESA et al., 2016).

3.2 Diferentes abordagens para identificar transcritos alternativos

Convencionalmente, o método empregado para analisar o sequenciamento de mais de uma amostra é processar os alinhamentos individualmente e só então unir as anotações (SONG et al., 2019). Os trabalhos de transcriptômica geralmente analisam múltiplas amostras, enquanto que os montadores mais utilizados e populares recebem de entrada somente um arquivo, como Cufflinks e CLASS2 (SONG et al., 2019; TRAPNELL et al., 2010). Por causa disso, os transcritos parciais das amostras são então concatenados para gerar uma única anotação de referência que represente todas as amostras (CONESA et al., 2016; SONG et al., 2019). Porém, há prejuízos na precisão da identificação dos transcritos ao realizar a pesquisa dessa maneira (CONESA et al., 2016; SONG et al., 2019). De qualquer forma, identificar novos transcritos utilizando sequenciamento de pequenas leituras é considerado um dos maiores desafios na análise de RNA-seq (CONESA et al., 2016). Cada programa de montagem tem suas características positivas e negativas quanto à construção de transcritos e por isso é interessante utilizar mais de um software ou método (VENTURINI et al., 2018).

Portanto, em nosso trabalho buscamos averiguar qual a melhor maneira de identificar variantes de *splicing* alternativo derivadas de genes de ncRNAs em dados de RNA-seq de várias amostras de câncer cervical, utilizando o método convencional ou uma alternativa. Por isso, traçamos nossos métodos em duas abordagens, testando três parâmetros diferentes: o método convencional como abordagem A de parâmetros A 5%, A 10% e A 20%, e o método alternativo como abordagem B de parâmetros B 5%, B 10% e B 20%. Resumidamente, o método convencional averiguamos na abordagem A, no qual nós concatenamos os arquivos de **anotação** das diferentes amostras após identificar as variantes de *splicing*, enquanto que o método alternativo conferimos com a abordagem B, no qual nós concatenamos os arquivos de **alinhamento** das onze amostras, antes de identificar as variantes de *splicing*.

Nós propomos a abordagem B com a concatenação dos alinhamentos dos transcriptomas pensando no próprio contexto dos transcriptomas dos subtipos histológicos ADC e SCC de câncer cervical. Uma vez que os tipos ADC e SCC apresentam uma heterogeneidade molecular entre as amostras de diferentes mulheres (BURK et al., 2017), a concatenação dos diferentes arquivos de alinhamentos pode selecionar transcritos cuja expressão varia entre as amostras, mas está presente dentro dos tipos histológicos.

Além disso, os parâmetros de porcentagem são determinados para o software CLASS2 considerar um transcrito como novo ou não (SONG; SABUNCIYAN; FLOREA, 2016). Os parâmetros de 5%, 10% ou 20% indicam que CLASS2 irá considerar aquele transcrito como novo somente se a quantidade de leituras alinhadas naquele transcrito corresponde a 5%, 10% ou 20% das leituras totais alinhadas sobre o gene (SONG; SABUNCIYAN; FLOREA, 2016). Para diminuir a anotação de transcritos falsos-positivo, nós escolhemos continuar nossa pesquisa somente com o parâmetro de 20% de CLASS2, criando anotações personalizadas de A 20% e B 20%. As abordagens A e B em específico serão discutidas nos tópicos a seguir.

3.2.1 Sobre a abordagem A

A abordagem A se aproxima do método convencional de análise de RNA-seq, como mencionado anteriormente (CONESA et al., 2016). Nessa abordagem nós utilizamos o programa CLASS2 (SONG; SABUNCIYAN; FLOREA, 2016) para identificar as variantes de *splicing* alternativo nos arquivos de alinhamento das amostras individuais, e então concatenamos as anotações geradas em um único arquivo de anotação. Como o programa CLASS2 não recebe uma anotação de referência para identificar os transcritos já anotados, como por exemplo anotação do Ensembl ou RefSeq, é possível identificar variantes potencialmente novas que ainda não foram anotadas (SONG; SABUNCIYAN; FLOREA, 2016).

Na abordagem A, pudemos observar que a identificação de transcritos alternativos potencialmente novos foi muito maior que na abordagem B. Com a abordagem A pudemos identificar cerca de 95% de transcritos potencialmente novos no total, enquanto na abordagem B foram 75% de transcritos não anotados na referência do Ensembl. Entretanto, a montagem de transcritos não anotados e potencialmente novos pode ser fruto da própria falta de precisão de identificar tais transcritos (CONESA et al., 2016; SONG et al., 2019). Por isso, escolhemos trabalhar somente com a abordagem A 20% que apresentou menor número de transcritos não anotados. Além disso, consideramos somente aqueles transcritos que apresentaram mais de um exon em sua estrutura, a fim de diminuir essa relação de falsos-positivos.

A maioria dos transcritos identificados na abordagem A apresentaram potencial codificador, conforme determinado pelo software CPC2 que apresenta acurácia para analisar ncRNAs (KANG et al., 2017). Por causa do escopo do nosso trabalho, trabalhamos somente com os 33% de transcritos não identificados que não apresentaram potencial codificador da abordagem A 20%, o que correspondeu a 33 mil e 224 transcritos.

Por consequência do maior número de transcritos não anotados, na abordagem A 20% foram apresentados mais genes não anotados nos 50 genes diferencialmente expressos de menor p-valor ajustado. Destes 50, 33 não estavam anotados na referência do Ensembl. Apesar de não estarem anotados, o padrão de

expressão de muitos desses transcritos permitiu distinguir os subtipos SCC de ADC, o que traz boas perspectivas para complementar o perfil de expressão desses tumores. No tópico adiante **3.4 - Novas variantes de splicing de ncRNAs** discutimos especificamente sobre os transcritos avaliados individualmente encontrados através dessa abordagem A 20%.

3.2.2 Sobre a abordagem B

A abordagem B é um pouco diferente do método convencional para analisar RNA-seq descrito anteriormente (CONESA et al., 2016). Ao invés de concatenarmos as anotações, em nossa abordagem B concatenamos os alinhamentos de todas as onze amostras em um único arquivo formato BAM. Esse alinhamento único foi dado de entrada para o identificador de variantes de *splicing* CLASS2, gerando consequentemente um arquivo de anotação referente à todas as amostras.

Desde o começo do trabalho, as abordagens A 20% e B 20% apresentaram diferenças notáveis. Nossos resultados mostraram que a identificação de transcritos não anotados e potencialmente novos foi muito menor na abordagem B do que na abordagem A. Tanto nos valores brutos quanto na análise de expressão diferencial, fica evidente que a abordagem B apresentou muito menos candidatos à novas anotações de transcritos.

Na abordagem B, cerca de 26% dos transcritos encontrados já estavam anotados na referência Ensembl. Como mencionado previamente, ainda realizamos uma filtragem dos transcritos não anotados para considerar somente os estruturados com mais de um exon, a fim de diminuir a possibilidade de identificar falsos-positivos (CONESA et al., 2016; SONG et al., 2019). Além disso, a maioria dos transcritos potencialmente novos tinham potencial codificador e, portanto, foram descartados por não entrarem no escopo de nossa pesquisa. Por isso, um pouco mais de 2 mil transcritos potencialmente novos e sem potencial codificador foram considerados para as nossas análises de expressão diferencial, em contraste com os mais de 33 mil potenciais novos encontrados com A 20%.

Por consequência do menor número de transcritos potencialmente novos, a análise de expressão diferencial da abordagem B 20% apresentou poucos genes não-anotados nos 50 genes mais diferencialmente expressos. Destes, somente 10 genes potencialmente novos foram encontrados nessa abordagem. Isso nos leva a crer que esse método B, de concatenação dos arquivos de alinhamento, não é o mais indicado para encontrar variantes de *splicing* alternativo de ncRNAs potencialmente novas.

No tópico numerado **3.4 - Novas variantes de *splicing* de ncRNAs** discorremos especificamente sobre os transcritos encontrados através da abordagem B 20% e também da A 20%.

3.3 Identificação de variantes de *splicing* em múltiplas amostras

Em geral, a expressão média de ncRNAs é menor que mRNAs (DERRIEN et al., 2012; ULITSKY; BARTEL, 2013). Enquanto que a montagem de pequenas leituras de RNA-seq é um constante desafio, geralmente somente os transcritos mais abundantes serão montados (HAAS; ZODY, 2010). Assim, parâmetros muito restritos podem descartar potenciais ncRNA novos e não anotados que apresentam baixa expressão (LONG et al., 2017). A execução de CLASS2 permite definir quão sensível será a identificação dos transcritos alternativos, em que o parâmetro padrão indicado para descobrir isoformas de baixa expressão é de que a quantidade de leituras que corroboram com a anotação daquele transcrito corresponde a 1% das leituras totais daquele gene todo (SONG; SABUNCIYAN; FLOREA, 2016). Nós optamos por utilizar o parâmetro de 20% a fim de encontrar transcritos cuja diferença de expressão entre os tipos histológicos fosse mais evidente, uma vez que nossa triagem com as abordagens de parâmetro 5% e 10% apresentaram resultados pouco satisfatórios de grande quantidade de transcritos não anotados.

Ainda que ncRNAs apresentam menor expressão no geral (DERRIEN et al., 2012; ULITSKY; BARTEL, 2013), parâmetros muito abrangentes podem apresentar falta de precisão na identificação dos transcritos e por conseguinte anotar mais falsos-positivos (CONESA et al., 2016; SONG et al., 2019). Apesar de utilizarmos os

parâmetros de 20% para fazer a contagem de leituras e expressão diferencial, conseguimos encontrar variantes potencialmente novas e sem potencial codificador utilizando altos parâmetros de filtro de qualidade e expressão.

Em ambas as abordagens pudemos encontrar transcritos não anotados sem potencial codificador. Porém, como discorrido nos tópicos anteriores, pudemos concluir que a abordagem A 20% se mostrou mais indicada para buscar novas variantes de *splicing* de ncRNAs em múltiplas amostras. Apesar da abordagem B 20% mostrar um padrão de expressão relevante para distinguir os subtipos ADC de SCC, só mostrou mais do mesmo, no sentido de que uma análise de RNA-seq utilizando uma anotação genômica como referência já seria suficiente (BURK et al., 2017; DERRIEN et al., 2012). A abordagem A 20% conseguiu apresentar mais variantes potencialmente novas mesmo com todos os parâmetros de filtragem e diminuição de falsos-positivos empregados.

A recente publicação do grupo desenvolvedor de CLASS2 também traz novas perspectivas para a identificação de variantes de *splicing* em mais de uma amostra. Em novembro de 2019, Song e colaboradores apresentaram PsiCLASS, que promete ser um montador de transcritos eficiente ao analisar múltiplas amostras de RNA-seq simultaneamente (SONG et al., 2019). Em virtude da publicação do referido programa durante a finalização de nosso trabalho, utilizamos somente o seu pacote adicional Grader para identificar a presença de anotação ou não para os transcritos encontrados em nossas abordagens. Porém, segundo os autores, o novo software PsiCLASS não é indicado para buscar variantes de *splicing* em uma quantidade muito grande de sequenciamentos como provenientes de repositórios (SONG et al., 2019). Em contrapartida, nossa abordagem A 20% é passível de automatização para mais amostras, entretanto deverá ser visto qual o limite do software para o número de anotações possíveis a serem concatenadas após a identificação das variantes de *splicing.*

3.4 Novas variantes de *splicing* de ncRNAs

Os RNAs sem potencial de codificação compõem a maior parte do transcriptoma humano (CHAN; TAY, 2018) e por isso, são bons candidatos a biomarcadores no caso de doenças como câncer, pois também são encontrados nesses transcriptomas (BATISTA; CHANG, 2013; BRUNNER et al., 2012; GUTSCHNER; DIEDERICHS, 2012; KOPP; MENDELL, 2018). Considerando que a descoberta de novos transcritos permite contribuir para a melhor anotação genômica do organismo sendo estudado (WANG; GERSTEIN; SNYDER, 2009), descobrir novos transcritos não anotados em transcriptomas tumorais trazem boas perspectivas para biomarcadores ou complemento ao perfil de expressão do tumor. Principalmente, variantes de *splicing* tem grande potencial como biomarcadores de doenças (PAJARES et al., 2007; URBANSKI; LECLAIR; ANCZUKÓW, 2018), uma vez que o processamento de *splicing* afeta a maioria dos genes humanos (BARBOSA-MORAIS et al., 2012; WANG; GERSTEIN; SNYDER, 2009) e atinge cerca de 25% dos lncRNAs (DERRIEN et al., 2012), além de pequenos ncRNA (COLLINS; SCHÖNFELD; CHEN, 2011; KROL; LOEDIGE; FILIPOWICZ, 2010). Além disso, muitos estudos mostraram variantes de *splicing* alternativo diferencialmente expressas em tumores ou exercendo funções importantes na biologia tumoral (PAJARES et al., 2007; SONG et al., 2018; URBANSKI; LECLAIR; ANCZUKÓW, 2018; VENABLES, 2006). Tecnologias de alta vazão como sequenciamento de RNA auxiliam na busca por candidatos a biomarcadores tumorais, uma vez que permite definir um perfil de expressão composto por múltiplos genes, de acordo com a progressão da doença ou até resposta para o tratamento (PAJARES et al., 2007). Entretanto há pouco desenvolvimento e teste clínico de tratamentos moleculares cujo alvo são variantes de *splicing* alternativo, ainda mais em câncer (PAJARES et al., 2007). Por isso é importante buscar por variantes de *splicing* alternativo de ncRNA em tumores, principalmente as que ainda não foram anotadas em referências.

Nesse trabalho nós conseguimos mostrar que há muitos transcritos não anotados e, portanto, potencialmente novos, sendo descobertos tanto da maneira convencional de análise (abordagem A) quanto da nossa alternativa proposta (abordagem B). Ademais, ao buscar por variantes não anotadas, nós resgatamos a

informação da presença de transcritos potencialmente novos diferencialmente expressos nos transcriptomas, tal qual não seria possível de realizar da maneira tradicional em que se realiza a anotação contra uma referência prévia (WANG; GERSTEIN; SNYDER, 2009).

Em nosso trabalho, a concatenação das anotações com a referência do Ensembl e a conferência manual no visualizador IGV e sua referência RefSeq padrão também foram importantes para avaliar os genes potencialmente novos com maior precisão. Ao avaliarmos manualmente os genes não codificadores com maior variação em nossas anotações no programa IGV, pudemos confirmar a presença de transcritos não-anotados nas referências RefSeq e Ensembl. Para analisar detalhadamente, escolhemos alguns transcritos da lista de 50 genes mais diferencialmente expressos encontrados nas abordagens A 20% e B 20%, os quais serão discutidos a seguir.

3.4.1 Novo transcrito MSTRG.20117.1

O transcrito MSTRG.20117.1 foi encontrado dentre os 50 genes mais diferencialmente expressos da abordagem A 20%, apresentando logFC de -8,71, p-valor de 2,67E-07, e p-valor ajustado de 8,92E-05. Quando visualizamos esse transcrito no IGV, pudemos perceber que o mesmo foi encontrado através da abordagem B 20% também, sob o identificador MSTRG.1993.1. Inclusive, o transcrito MSTRG.1993.1 aparece dentre os 50 genes mais diferencialmente expressos da abordagem B 20%, com valores de logFC de -8,71, p-valor de 1,22E-07, e p-valor ajustado de 5,62E-05. No mesmo *locus* não foram encontradas nenhuma anotação tanto na referência do Ensembl quanto na referência do RefSeq (ZERBINO et al., 2018). Isso nos leva a crer que esse transcrito tem potencial a ser descrito como proveniente de um novo gene.

Nós realizamos a filtragem dos transcritos não anotados através do seu potencial codificador utilizando o software CPC2 (KANG et al., 2017). Ao visualizarmos com o programa IGV pudemos ver as leituras pareadas alinhadas sobre a anotação no sentido anti-senso. Visto que a maior parte dos ncRNAs, tanto lncRNA quanto

pequenos ncRNAs, são transcritos a partir da fita anti-senso, isso é um bom indício do biotipo desse transcrito não anotado (COLLINS; SCHÖNFELD; CHEN, 2011; QUINN; CHANG, 2016). Além dos valores significativos da expressão diferencial desse transcrito, consideramos o transcrito MSTRG.20117.1 com potencial a ser descrito como oriundo de novo gene de ncRNA.

3.4.2 Novo transcrito MSTRG.8127.1

O transcrito MSTRG.8127.1 foi encontrado entre os 50 genes mais diferencialmente expressos da abordagem A 20%. Ao visualizarmos no IGV, observamos que no mesmo *locus* não apareceu nenhuma anotação nas referências do RefSeq, do Ensembl e nem da abordagem B 20%.

Possivelmente os altos parâmetros de 20% de CLASS2 podem ter impedido o encontro de transcritos não-anotados nesse *locus* para a abordagem B. É fato que parâmetros muito restritos podem descartar esses transcritos não anotados e potencialmente novos que estão com baixa expressão (LONG et al., 2017), pois em suma serão anotados somente os transcritos mais abundantes (HAAS; ZODY, 2010). Aliás, é um fator muito importante a se considerar em nosso trabalho, pois a expressão de ncRNAs é geralmente menor que outros RNAs codificadores (DERRIEN et al., 2012; ULITSKY; BARTEL, 2013).

Contudo, na mesma balança temos que considerar que parâmetros muito abrangentes podem anotar mais falsos-positivos (CONESA et al., 2016; SONG et al., 2019). Por todo o exposto, é compreensível que encontrar novos transcritos potencialmente novos em sequenciamentos de pequenas leituras é um dos maiores desafios das análises de RNA-seq (CONESA et al., 2016).

A abordagem B 20% realmente apresentou menos transcritos não anotados que A 20% no geral. Porém, a quantidade de transcritos não anotados de B 5% era mais próxima à quantidade de transcritos não anotados de A 20%, cerca de 100 mil transcritos. Possivelmente teríamos encontrado os mesmos transcritos

potencialmente novos se comparássemos as abordagens pela quantidade de transcritos não anotados, ao invés de compararmos os mesmos parâmetros do programa. Aparentemente o modo distinto de execução e concatenação dos métodos A e B influenciou mais na quantidade e variedade de transcritos não anotados encontrados, do que realmente os valores de parâmetros iguais. Para verificar isso, a análise de expressão diferencial e visualização das anotações deve ser feita também com a abordagem B 5% e B 10%, em busca desses transcritos de menor expressão.

De qualquer maneira, o transcrito MSTRG.8127.1 apareceu dentre os 50 genes mais diferencialmente expressos da abordagem A 20% e com alta confiabilidade de acordo com seus valores de logFC de -7,50, p-valor de 2,47E-12, e p-valor ajustado de 4,66E-09.

Portanto, de acordo com essas proposições, nós consideramos o transcrito MSTRG.8127.1 como potencial para ser descrito como oriundo de um novo gene de ncRNA.

3.4.3 Novos transcritos MSTRG.1476.2 e MSTRG.1476.3

A partir dos genes mais diferencialmente expressos da abordagem A 20%, escolhemos no caso o gene MSTRG.1476 e seus transcritos 1, 2 e 3. O transcrito MSTRG.1476.1 apresentou logFC de 23,92, p-valor de 1,98E-12, e p-valor ajustado de 4,27E-09, porém a visualização no IGV nos permitiu identifica-lo como um falso-positivo, pois estava alinhado sobre o gene SPRR2B na anotação do RefSeq. Como esse gene não estava presente na anotação do Ensembl, esse transcrito foi erroneamente considerado como não anotado. Enquanto isso, os outros transcritos MSTRG.1476.2 apresentou logFC de 11,27, p-valor de 2,33E-11, e p-valor ajustado de 2,51E-08, enquanto que MSTRG.1476.3 apresentou logFC de 22,81, p-valor de 3,70E-11, e p-valor ajustado de 3,71E-08.

Ao visualizar os transcritos MSTRG.1476.2 e MSTRG.1476.3 com o IGV, pudemos perceber que são os mesmos transcritos que na abordagem B 20% foram

identificados como MSTRG.342.1, MSTRG.341.2 e MSTRG.342.3. Toda a estrutura desses transcritos com seus longos íntrons não aparecem anotadas nas referências do Ensembl. Entretanto, seus exons se alinham com os genes anotados na referência do RefSeq SPRR2B, SPRR2E, SPRR2F e SPRR2C. De fato, lncRNAs apresentam regiões intrônicas maiores do que o usual para genes que codificam para proteína (DERRIEN et al., 2012), então há a possibilidade de existir esse gene que incorpore os exons de SPRR2B, SPRR2E, SPRR2F e SPRR2C e as longas regiões intrônicas entre eles. Além disso, a transcrição de genes não-codificadores geralmente ocorre com o uso de promotores em comum com genes codificadores de proteína, porém no sentido anti-senso (QUINN; CHANG, 2016). Isso nos levou a supor que os transcritos potencialmente novos de MSTRG.1476 compartilham promotores em comum com os genes SPRR2B, SPRR2E, SPRR2F e SPRR2C.

Uma vez que a referência que utilizamos do Ensembl não apresenta os genes SPRR2B, SPRR2E e SPRR2F, mas a referência RefSeq sim, também nos leva a considerar utilizar mais de uma anotação para confirmar a descoberta de novos transcritos. Incorporar o RefSeq no método que estamos propondo iria complementar a confiabilidade da descoberta de transcritos potencialmente novos.

O alinhamento das leituras e o fato dos transcritos MSTRG.1476.2 e 3, e MSTRG.342.2 e 3, terem aparecido nos 50 genes mais diferencialmente expressos nas duas abordagens A e B 20% também suportam para sua existência. Além disso, a ocorrência do *splicing* alternativo em lncRNA geralmente originam ao menos dois transcritos distintos por locus gênico (DERRIEN et al., 2012), tal qual podemos ver pela anotação e expressão diferencial de dois transcritos dos mesmos genes MSTRG.1476 e MSTRG.342. Portanto, descartando ainda o transcrito falso positivo MSTRG.1476.1, nós consideramos os transcritos MSTRG.1476.2 e MSTRG.1476.3 com potencial a serem descritos como oriundos de novos genes de ncRNAs.

3.4.4 Potencialmente novo transcrito MSTRG.797.3

O transcrito MSTRG.797.3 foi encontrado entre os 50 genes mais diferencialmente expressos da abordagem B 20%, com logFC de -2,22, p-valor de 9,80E-07, e p-valor ajustado de 2,36E-04. Ao visualizarmos no IGV, pudemos observar que esse mesmo transcrito foi identificado com a abordagem A 20% como MSTRG.1935.6. Porém, o transcrito MSTRG.1935.6 não está diferencialmente expresso na abordagem A 20%. Nenhum transcrito foi anotado pela referência do Ensembl, mas na anotação da referência de RefSeq aparece o transcrito BLACAT1 sobre o mesmo *locus*.

O gene BLACAT1 aparece na anotação completa do Ensembl GRCh37/hg-19, sob o nome de LEMD1 (ZERBINO et al., 2018). Porém LEMD1 não é um gene do qual são transcritos somente ncRNAs, pois o mesmo gera transcritos codificadores também. Em nossa pesquisa utilizamos a anotação do genoma humano de versão GRCh37/hg-19 do Ensembl (ZERBINO et al., 2018) assim como outros trabalhos de busca por ncRNAs o fizeram (HAN et al., 2016; LUYKX et al., 2019). A anotação do Ensembl foi utilizada para que pudéssemos identificar somente os transcritos que ainda não estavam anotados nessa referência, ou seja, potencialmente novos ncRNAs. Uma diferença na anotação do Ensembl que utilizamos em nosso trabalho é que nós descartamos todos os genes com biotipo de codificador de proteína, que por ventura transcrevessem ao menos um RNA codificador. É provável que com isso, LEMD1 foi excluído e não apareceu nas nossas anotações A 20% ou B 20%.

Portanto, o potencialmente novo transcrito MSTRG.347.1 na verdade é um transcrito não codificador derivado de um gene que transcreve variantes codificadoras, os quais estão todas já anotadas no Ensembl. Portanto, nós não consideramos o transcrito MSTRG.797.3 como potencial transcrito a ser descrito como novo ncRNA.

3.4.5 Potencialmente novo transcrito MSTRG.347.1

Identificamos o transcrito MSTRG.347.1 nos 50 genes diferencialmente expressos de menor p-valor ajustado na abordagem B 20%, com logFC de 6,74, p-valor de 2,25E-12, e p-valor ajustado de 5,68E-09. Apesar de não estar anotado na referência do Ensembl, ao visualizarmos no IGV e com acesso à anotação do RefSeq, pudemos ver que tal *locus* gênico está associado com o gene S100A8. É um caso semelhante ao transcrito anterior MSTRG.797.3. Apesar de não aparecer na anotação que utilizamos (sem genes que transcrevam ao menos um transcrito codificador) S100A8 está presente na anotação completa do Ensembl, sendo um gene de biotipo codificador, que transcreve tanto transcritos não codificadores quanto codificadores. Portanto, não consideramos o transcrito MSTRG.347.1 como potencial transcrito a ser descrito como novo ncRNA.

3.4.6 Transcrito MSTRG.2365.1 adjacente ao gene CALML-AS1

Ao analisarmos o gene CALML3-AS1 que apareceu diferencialmente expresso na abordagem B 20%, encontramos um transcrito não anotado na outra abordagem A 20%. Esse transcrito, de identificador MSTRG.2365.1, não estava anotado tanto na referência do Ensembl quanto na referência do RefSeq. Tal transcrito MSTRG.2365.1 encontra-se adjacente ao *locus* gênico de CALML3-AS1 e está diferencialmente expresso da abordagem A 20%, com valores de logFC de 7,03, p-valor de 6,39E-12, e p-valor ajustado de 9,63E-09.

Existe a possibilidade da expressão de um gene estar correlacionada com a expressão de genes próximos, como CALML3-AS1 nesse caso (ENGREITZ et al., 2016). Os transcritos do gene CALML3-AS1 estão diferencialmente expressos na abordagem A 20% e até na abordagem B 20% (CALML3-AS1_ENST00000543008 com logFC de 8,14, p-valor de 1,20E-17 e p-valor ajustado de 9,07E-14 na abordagem A 20%; CALML3-AS1_ENST00000545372 com logFC de 7,74, p-valor de 4,74E-07, p-valor ajustado de 1,42E-04, na abordagem A 20%). Da mesma maneira, o transcrito

MSTRG.2365.1 também está diferencialmente expresso, com logFC de 7,03, p-valor de 6,39E-12, e p-valor ajustado de 9,63E-09.

Ademais, a orientação anti-senso indica seu potencial biotipo de ncRNA, uma vez que esse tipo de transcrito geralmente é gerado a partir da fita anti-senso (COLLINS; SCHÖNFELD; CHEN, 2011; QUINN; CHANG, 2016). Por essas razões propostas, consideramos o transcrito MSTRG.2365.1 com potencial para ser descrito como oriundo de novo gene de ncRNA.

3.4.7 Biotipo dos transcritos potencialmente novos

Uma vez que nosso parâmetro para identificar variantes de *splicing* não anotadas foi através do potencial de codificação, os transcritos potencialmente novos encontrados nesse trabalho são classificados como ncRNA, sem especificidade se lncRNA, miRNA ou outros.

São necessárias mais pesquisas e outros experimentos de sequenciamento para avaliar especificamente os transcritos encontrados nesse trabalho e classifica-los como lncRNA, miRNA, entre outros (KASHI et al., 2016). Inclusive, experimentos como de perfil ribossomal poderiam até elucidar o potencial codificador desses ncRNAs (KASHI et al., 2016).

3.5 Relação dos transcritos com câncer cervical

3.5.1 TINCR, CALML3-AS1 e RP11-89K21.1 (LINC01833)

Através das nossas análises encontramos alguns genes anotados que estavam diferencialmente expressas entre as amostras e entre os tipos histológicos de câncer cervical ADC e SCC. Sendo esses especialmente os genes TINCR,

CALML3-AS1 e RP11-89K21.1. Todos esses genes foram encontrados diferencialmente expressos em SCC e ADC em ambas as abordagens A 20% e B 20%.

Encontramos o gene TINCR diferencialmente expresso em ambas as abordagens A 20% e B 20% (na abordagem A 20%, logFC de 6,07, p-valor de 4,98E-08, p-valor ajustado de 1,97E-05; na abordagem B 20%, logFC de 6,09, p-valor de 2,55E-08, p-valor ajustado de 1,43E-05).

TINCR é descrito como ncRNA indutor de diferenciação tecidual e é classificado como lincRNA no Ensembl (ZERBINO et al., 2018). TINCR também é conhecido como PLAC2 ou LINC00036. Sua expressão é alta em tecido epidermal e placentário saudável (FAGERBERG et al., 2014). Em tecido tumoral, foi observada a alta expressão de TINCR em carcinoma escamoso oral, tanto em tecido quanto em linhagem celular (CHEN et al., 2019).

Já em carcinoma escamoso cervical, o mesmo gene também foi encontrado regulando positivamente a proliferação celular (HOU et al., 2019) inclusive sofrendo regulação por um miRNA. miRNAs são pequenos RNAs de cerca de 22 nucleotídeos que exercem atividade regulatória pós-transcricional quando ligam-se a sequências complementares nos transcritos-alvo. Assim, podem induzir a degradação ou inibição da tradução daquele RNA (KROL; LOEDIGE; FILIPOWICZ, 2010). Portanto, é compreensível encontrar esse gene diferencialmente expresso em nossas amostras de carcinoma escamoso SCC. Além disso, consideramos que TINCR é um bom candidato a biomarcador de SCC.

O gene **CALML3-AS1** está entre os 50 genes de maior expressão diferencial de ambas as abordagens A 20% e B 20%, especificamente dois de seus transcritos de identificadores ENST00000543008 e ENST00000545372. Ambos transcritos estão diferencialmente expressos na abordagem A 20% (CALML3-AS1_ENST00000543008 com logFC de 8,14, p-valor de 1,20E-17 e p-valor ajustado de 9,07E-14; CALML3-AS1_ENST00000545372 com logFC de 7,74, p-valor de 4,74E-07, p-valor ajustado de 1,42E-04) e na abordagem B 20% (CALML3-AS1_ENST00000543008 com logFC de 8,14, p-valor de 5,13E-18, p-valor ajustado de 5,18E-14; CALML3-

AS1_ENST00000545372 com logFC de 7,74, p-valor de 3,70E-07, p-valor ajustado de 1,33E-04).

É um ncRNA de pouca publicação, porém estudos recentes apontam para sua importância em câncer de colo de útero. Experimentos celulares e de *knockdown* apontaram para um importante papel de CALML3-AS1 na progressão de câncer cervical em uma recente publicação (LIU et al., 2019). Sua superexpressão foi associada com maior crescimento e metástase em tumores cervicais de pacientes (LIU et al., 2019). Porém, esse trabalho não focou nos tipos histológicos desse tumor e como CALML3-AS1 se comporta especificamente em SCC e ADC. Seria interessante repetir os experimentos celulares e de *knockdown* propostos pelo grupo considerando os tipos ADC e SCC. Em nossas análises, seu padrão de maior expressão está mais associado ao tipo histológico SCC. Portanto, consideramos o gene CALML3-AS1 como um bom candidato a biomarcador para o subtipo SCC, juntamente com o transcrito adjacente MSTRG.2365.1 que já foi elucidado anteriormente.

O gene **RP11-89K21.1** está entre os 50 genes mais diferencialmente expressos de ambas as abordagens A 20% e B 20%. Inclusive, dois de seus transcritos aparecem entre os 50 genes mais diferencialmente expressos das duas abordagens, sendo seus identificadores ENST00000432125 e ENST00000444871, no caso ENST00000432125 com logFC de -6,80, p-valor de 9,97E-11, p-valor ajustado de 7,51E-08 na abordagem A 20%, e com logFC de -6,79, p-valor de 4,73E-11, e p-valor ajustado de 6,84E-08 na abordagem B 20%; e o transcrito ENST00000444871 com logFC de -7,93, p-valor de 1,37E-10, p-valor ajustado de 9,86E-08 na abordagem A 20%, e com logFC de -7,92, p-valor de 1,87E-11, p-valor ajustado de 3,15E-08 na abordagem B 20%.

Esse gene também é conhecido como **LINC01833** (do inglês, *Long Intergenic Non-Protein Coding RNA 1833*). Diferente dos outros genes TINCR e CALML3-AS1, o lincRNA RP11-89K21.1 está mais diferencialmente expresso em ADC em ambas as abordagens A 20% e B 20%, como visto nos *heatmaps* e pelos seus valores de logFC mencionados anteriormente.

Há poucos trabalhos mencionando o lincRNA RP11-89K21.1 e seu contexto biológico. Esse lincRNA RP11-89K21.1 já foi encontrado em outro trabalho relacionado com adenocarcinoma (CHEN et al., 2016). Chen e colaboradores em 2016 apresentaram RP11-89K21.1 como expresso aberrantemente em adenocarcinoma de pulmão, dentre outros lncRNAs diferencialmente expressos. No caso, RP11-89 K21.1 estava menos expresso em adenocarcinoma de pulmão do que amostras de pulmão sadios, de acordo com experimentos de PCR em tempo real e microarranjo (CHEN et al., 2016). Inclusive Chen e colaboradores trazem RP11-89 K21.1 como importante lncRNA na via de sinalização de WNT em câncer (CHEN et al., 2016).

No contexto do câncer cervical, o lincRNA RP11-89K21.1 também foi mencionado por Boers e colaboradores. Em sua pesquisa, Boers e colaboradores buscaram candidatos a marcadores de metilação para lesões cervicais e neoplasias cervicais (BOERS et al., 2016). Metilação é um processo de marcação do DNA durante o estabelecimento do *imprinting* genômico, mecanismo epigenético de regulação gênica que pode alterar a expressão gênica de um alelo parental específico (ISHIDA; MOORE, 2013; REIK; WALTER, 2001). De acordo com o grupo, o lincRNA RP11-89K21.1 foi selecionado incialmente através de um sequenciamento do tipo *MetylCap-seq*, referente ao sequenciamento do perfil de metilação do DNA em todo o genoma em amostras de lesões cervicais (BOERS et al., 2016). Esse sequenciamento revelou 176 regiões diferencialmente metiladas, correspondendo a 164 genes, dentre os quais estava incluso o lncRNA RP11-89K21.1 (BOERS et al., 2016). Porém, lncRNA RP11-89K21.1 não passou pelas validações iniciais, sendo descartado logo na primeira validação experimental de PCR específico de metilação com tecido tumoral (BOERS et al., 2016). Outras análises de expressão diferencial desse lncRNA não foram feitas pelo grupo (BOERS et al., 2016).

Apesar de poucas publicações mencionando RP11-89K21.1, o fato desse lncRNA ter sido encontrado no contexto de adenocarcinoma e de lesões cervicais traz boas perspectivas para ser um candidato a biomarcador para ADC. Portanto, com esse trabalho propomos que os transcritos ENST00000432125 e ENST00000444871 do gene RP11-89K21.1 são bons candidatos a biomarcadores de ADC.

Além disso, propomos que os transcritos ENST00000543008 e ENST00000545372 do gene CALML3-AS1 e o gene TINCR são bons candidatos a biomarcadores de SCC.

3.5.2 Novos transcritos

Os transcritos potencialmente novos discutidos anteriormente que encontramos em nosso trabalho apresentam padrão de expressão que pode estar associado tanto ao subtipo SCC quanto ao ADC. Os padrões de expressão diferencial de todos os novos transcritos **MSTRG.1476.2** (logFC de 11,27, p-valor de 2,33E-11, e p-valor ajustado de 2,51E-08), **MSTRG.1476.3** (logFC de 22,81, p-valor de 3,70E-11, e p-valor ajustado de 3,71E-08) e **MSTRG.2365.1** (logFC de 7,03, p-valor de 6,39E-12, e p-valor ajustado de 9,63E-09) estão associados positivamente com o subtipo SCC. Portanto, são bons candidatos a biomarcadores em SCC.

Enquanto isso, os padrões de expressão diferencial de todos os novos transcritos **MSTRG.20117.1** (logFC de -8,71, p-valor de 2,67E-07, e p-valor ajustado de 8,92E-05) e **MSTRG.8127.1** (logFC de -7,50, p-valor de 2,47E-12, e p-valor ajustado de 4,66E-09) estão associados negativamente com o subtipo SCC, portanto positivamente com o subtipo ADC. Assim, são bons candidatos a biomarcadores em ADC.

A descoberta dessas variantes de *splicing* diferencialmente expressas em SCC e ADC permite complementar o padrão de expressão visto especificamente nesses subtipos. Conforme a confirmação experimental dessas variantes possa ser feita, e a confirmação dos resultados em um maior número amostral, há perspectiva de serem definidos como biomarcadores para o tipo histológico SCC e ADC (ARONSON; FERNER, 2017).

Além disso, é interessante pontuar sobre o transcrito MSTRG.2365.1 que está adjacente ao gene CALML3-AS1, ambos diferencialmente expressos em SCC. Uma vez que a expressão de CALML2-AS1 e MSTRG.2365.1 pode estar correlacionada

em virtude da alta proximidade de seus *locus* (ENGREITZ et al., 2016) e que CALML3-AS1 foi visto atuando na progressão de câncer cervical (LIU et al., 2019), seria interessante buscar por esse transcrito potencialmente novo MSTRG.2365.1 em um número amostral maior de tumores SCC. Confirmando seu padrão de expressão em mais amostras, juntamente com CALML3-AS1 o transcrito MSTRG.2365.1 pode indicar potencial biomarcador para essa condição (QIN et al., 2019; ARONSON; FERNER, 2017; WENTZENSEN; VON KNEBEL DOEBERITZ, 2007). Ao realizar a análise em um maior número de amostras podemos aumentar tanto a confiabilidade de nossos resultados assim como descobrir um padrão de expressão diferencial mais evidente (BURK et al., 2017).

As variações genéticas podem influenciar na expressão gênica e de variantes de *splicing* (PICKRELL et al., 2010). Inclusive, nos gráficos estilo *heatmap* das duas abordagens A 20% e B 20% pudemos perceber a variação no padrão de expressão dentre as amostras de cada tipo histológico de câncer cervical ADC e SCC. No caso, poucos genes apresentando o mesmo padrão de expressão em todas as amostras de um tipo histológico ADC ou SCC. Essa heterogeneidade de padrão de expressão dos tipos histológicos de câncer cervical também é observada em análises de expressão diferencial de muitas amostras (BURK et al., 2017; KORI; YALCIN ARGA, 2018; SCHMITT et al., 2010; WITTEN et al., 2010).

Pensando nisso, uma perspectiva futura para nosso trabalho é buscar esses transcritos encontrados potencialmente novos em dados do repositório *The Cancer Genome Atlas* (TCGA). O TCGA é um repositório muito conhecido que contém uma enorme quantidade de dados coletados de tumores e pacientes do mundo todo, depositados no banco por grupos de pesquisa e cientistas diversos (BURK et al., 2017). Informações clínicas como sobrevida, presença ou não de metástase e até fatores comportamentais importantes, como hábito tabagista, podem ser encontradas no TCGA (BURK et al., 2017). Todos esses dados auxiliam pesquisadores e pesquisadoras a compreender o contexto biológico do perfil de expressão do tumor (BURK et al., 2017).

Apesar de muito rico em informação, raramente os dados do repositório do TCGA serão transcriptomas tumorais de mulheres brasileiras e, portanto, podem não refletir realmente as características do câncer cervical específico de nosso país. Ressaltamos que é importante analisar tumores de pacientes da própria nação, cidade ou comunidade, pois pode-se encontrar importantes variações, relevantes para a população estudada, até por que os índices de incidência e mortalidade também podem variar regionalmente no Brasil (INCA, 2018).

A variabilidade de expressão entre amostras de um mesmo tipo histológico também deve ser considerada no quesito de políticas públicas de saúde para a nossa população brasileira. A incidência de câncer cervical dentro do Brasil é muito discrepante entre as regiões, sendo o tipo tumoral mais incidente na região Norte e ao mesmo tempo o quarto mais incidente na região Sul (INCA, 2018). A desigualdade regional do câncer cervical também segue padrão semelhante nos índices de mortalidade (RIBEIRO et al., 2016). Assim como as regiões Norte e Nordeste são as que mais têm incidência desse câncer, também são lá em que mais morrem mulheres vítimas da doença (RIBEIRO et al., 2016). Apesar de poucos trabalhos publicados com esse tema, também deve ser levado em conta a incidência de câncer cervical nos recortes sociais de minorias, como mulheres indígenas, quilombolas, negras e periféricas. Sabendo que tanto a genética quanto as características comportamentais e de infecção são fatores de risco para o desenvolvimento de câncer cervical (CASTELLSAGUÉ; MUÑOZ, 2003; KIM et al., 2012; KJELLBERG et al., 2000), é imprescindível considerar o contexto da mulher paciente.

A busca por novos biomarcadores para os subtipos de câncer cervical precisa estar atenta à diversidade de mulheres, e a descoberta de novos biomarcadores promissores precisa ser muito bem confirmada para que possa enfim voltar para os cuidados com a saúde da população em forma de diagnóstico ou tratamento. Nossa busca por novas variantes de *splicing* diferencialmente expressas em câncer cervical em mulheres da região Sudeste nos permitiu compreender melhor o padrão de expressão desses tipos de tumores em específico. Porém é importante que essas variantes sejam procuradas em maiores escalas, tanto em tumores de mais mulheres do Sudeste e de outras regiões do país, quanto a nível latino-americano e global.

3.6 CONCLUSÃO

Os ncRNAs estão ganhando cada vez mais atenção nas pesquisas biomédicas e vêm se destacando como elementos importantes em diversos cânceres. O desenvolvimento das tecnologias ômicas de sequenciamento, mais precisamente do sequenciamento de RNA, permitiu a busca por esses tipos de transcritos com maior precisão e alcance.

Nosso objetivo específico inicial era comparar a identificação de novas variantes de *splicing* de genes não codificadores de proteínas com a ferramenta CLASS2 e utilizando duas abordagens e três parâmetros de limiar distintos. Nesse trabalho, observamos que os parâmetros de limiar muito baixos retornavam muito mais transcritos não-anotados, concluímos então que o parâmetro de 20% de CLASS2 é melhor para obter menor taxa de falsos-positivo. Ainda observamos que a abordagem B não identificou tantos transcritos não-anotados quanto a abordagem A, portanto concluímos que a abordagem A é mais eficiente para identificar novos transcritos não anotados. Não obstante, deve ser repetida a análise completa com a abordagem B 5% para investigar se um parâmetro de menor restrição no método B em específico permite identificar mais transcritos potencialmente novos.

Assim, dos métodos testados nesse trabalho, a metodologia proposta como abordagem A 20% é a estratégia mais eficiente para encontrar novas variantes de *splicing* alternativo de ncRNAs, não-anotadas na referência Ensembl, comumente utilizada nas análises de sequenciamentos.

Nosso outro objetivo específico era identificar a expressão diferencial de novas variantes de *splicing* de transcritos ncRNA, entre os tipos histológicos SCC e ADC de câncer cervical. Nós encontramos mais de 30 transcritos não codificadores, não anotados e potencialmente novos diferencialmente expressos entre ADC e SCC. Destes, apresentamos especificamente os transcritos **MSTRG.1476.2** (logFC de 11,27, p-valor de 2,33E-11, e p-valor ajustado de 2,51E-08), **MSTRG.1476.3** (logFC de 22,81, p-valor de 3,70E-11, e p-valor ajustado de 3,71E-08) e **MSTRG.2365.1** (logFC de 7,03, p-valor de 6,39E-12, e p-valor ajustado de 9,63E-09) como bons candidatos a biomarcadores positivos em SCC. Mais ainda, apresentamos os transcritos

MSTRG.20117.1 (logFC de -8,71, p-valor de 2,67E-07, e p-valor ajustado de 8,92E-05) e **MSTRG.8127.1** (logFC de -7,50, p-valor de 2,47E-12, e p-valor ajustado de 4,66E-09) como bons candidatos a biomarcadores positivos para ADC. Além disso, propomos que esses transcritos MSTRG.20117.1, MSTRG.8127.1, MSTRG.1476.2, MSTRG.1476.3, e MSTRG.2365.1 têm potencial para serem descritos como oriundos de novos genes de ncRNAs.

O depósito de uma enorme quantidade de dados em bancos de dados biológicos relacionados a ncRNAs, *splicing* alternativo e câncer possibilita investigar e corroborar as descobertas encontradas nesse trabalho, especialmente se forem dados regionais brasileiros. Assim sendo, é uma perspectiva futura de prosseguir com essa linha de pesquisa analisando os sequenciamentos de RNA de dezenas de tumores cervicais disponíveis em dados públicos nos repositórios apropriados ou também em mais amostras de pacientes brasileiras. Nesses, poderemos buscar os eventos de *splicing* alternativo em ncRNAs em um maior número amostral e consequentemente dar mais robustez aos nossos resultados.

REFERÊNCIAS

(INCA), Instituto Nacional de Câncer. **Estimativa de Câncer no Brasil**. Disponível em: <https://www.inca.gov.br/numeros-de-cancer>. Acesso em: 27 fev. 2019.

AALIJAHAN, H.; GHORBIAN, S. Long non-coding RNAs and cervical cancer. **Experimental and Molecular Pathology**, v. 106, p. 7–16, 1 fev. 2019.

ANDERS, S.; PYL, P. T.; HUBER, W. HTSeq--a Python framework to work with high-throughput sequencing data. **Bioinformatics (Oxford, England)**, v. 31, n. 2, p. 166–9, 15 jan. 2015.

ARONSON, J. K.; FERNER, R. E. Biomarkers—a general review. **Current Protocols in Pharmacology**, v. 2017, n. March, p. 9.23.1-9.23.17, 2017.

BARBOSA-MORAIS, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. **Science**, v. 338, n. 6114, p. 1587–1593, 2012.

BARTOLOMEI, M. S.; TILGHMAN, S. M. Genomic Imprinting in Mammals. **Annual Review of Genetics**, v. 31, n. 1, p. 493–525, 2014.

BATISTA, P. J.; CHANG, H. Y. Long noncoding RNAs: cellular address codes in development and disease. **Cell**, v. 152, n. 6, p. 1298–307, 14 mar. 2013.

BOERS, A. et al. Discovery of new methylation markers to improve screening for cervical intraepithelial neoplasia grade 2/3. **Clinical Epigenetics**, v. 8, n. 1, p. 1–16, 2016.

BRUNNER, A. L. et al. Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. **Genome biology**, v. 13, n. 8, p. R75, 28 ago. 2012.

BURK, R. D. et al. Integrated genomic and molecular characterization of cervical cancer. **Nature**, v. 543, n. 7645, p. 378–384, 2017.

BUZA, N.; HUI, P. Immunohistochemistry in gynecologic pathology an example-based practical update. **Archives of Pathology and Laboratory Medicine**, v. 141, n. 8, p. 1052–1071, 2017.

CASTELLSAGUÉ, X.; MUÑOZ, N. Chapter 3: Cofactors in human papillomavirus carcinogenesis--role of parity, oral contraceptives, and tobacco smoking. **Journal of the National Cancer Institute. Monographs**, n. 31, p. 20–28, 2003.

CHAN, J. F. W. et al. Differential cell line susceptibility to the emerging Zika virus: implications for disease pathogenesis, non-vector-borne human transmission and animal reservoirs. **Emerging microbes & infections**, v. 5, n. July, p. e93, 2016.

CHAN, J. J.; TAY, Y. Noncoding RNA: RNA regulatory networks in cancer. **International Journal of Molecular Sciences**, v. 19, n. 5, 2018.

CHEN, F. et al. LncRNA PLAC2 activated by H3K27 acetylationpromotes cell proliferation and invasion via the activation ofWnt/β-catenin pathway in oral squamous cell carcinoma. **International Journal of Oncology**, v. 54, n. 4, p. 1183–1194, 1 abr. 2019.

CHEN, J. et al. Detection and Analysis of Wnt Pathway Related lncRNAs Expression Profile in Lung Adenocarcinoma. **Pathology and Oncology Research**, v. 22, n. 3, p. 609–615, 2016.

COLLINS, L. J.; SCHÖNFELD, B.; CHEN, X. S. **The epigenetics of non-coding RNA**. First Edit ed. [s.l.] Elsevier Inc., 2011.

CONESA, A. et al. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v. 17, n. 1, p. 13, 26 dez. 2016.

CUSCHIERI, K.; WENTZENSEN, N. Human papillomavirus mRNA and p16 detection as biomarkers for the improved diagnosis of cervical neoplasia. **Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology**, v. 17, n. 10, p. 2536–45, out. 2008.

DE KLERK, E.; 'T HOEN, P. A. C. **Alternative mRNA transcription, processing, and translation: Insights from RNA sequencingTrends in Genetics**, 2015.

DE RIE, D. et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. **Nature Biotechnology**, v. 35, n. 9, p. 872–878, 1 set. 2017.

DENARO, N.; MERLANO, M. C.; LO NIGRO, C. Long noncoding RNAs as regulators of cancer immunity. **Molecular Oncology**, v. 13, n. 1, p. 61–73, 2019.

DERRIEN, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. **Genome research**, v. 22, n. 9, p. 1775–89, set. 2012.

DU, H.; CHEN, Y. Competing endogenous RNA networks in cervical cancer: function, mechanism and perspective. **Journal of Drug Targeting**, v. 0, n. 0, p. 1–15, 2018.

EBI EMBL. Disponível em: <https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/primary-and-secondary-databases>.

ENGREITZ, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. **Nature**, v. 539, n. 7629, p. 452–455, 2016.

ESPINDOLA, F. S. et al. Recursos de bioinformática aplicados às ciências ômicas como genômica, transcriptômica, proteômica, interatômica e metabolômica. **Bioscience Journal**, v. 26, n. 3, p. 463–477, 2010.

ESTELLER, M. Non-coding RNAs in human disease. **Nature Reviews Genetics**, v.

12, n. 12, p. 861–874, 18 dez. 2011.

FAGERBERG, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. **Molecular and Cellular Proteomics**, 2014.

FINOTELLO, F.; DI CAMILLO, B. Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. **Briefings in Functional Genomics**, v. 14, n. 2, p. 130–142, 2015.

FLYNN, R. A.; CHANG, H. Y. Long noncoding RNAs in cell-fate programming and reprogramming. **Cell stem cell**, v. 14, n. 6, p. 752–61, 5 jun. 2014.

FRANCIS CRICK. The Central Dogma of molecular biology. **Nature**, v. 227, p. 561–563, 1970.

GADDUCCI, A.; GUERRIERI, M. E.; COSIO, S. Adenocarcinoma of the uterine cervix: Pathologic features, treatment options, clinical outcome and prognostic variables. **Critical Reviews in Oncology/Hematology**, v. 135, n. January, p. 103–114, 2019.

GARBER, M. et al. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nature Methods**, v. 8, n. 6, p. 469–477, 2011.

GHITTONI, R. et al. Role of human papillomaviruses in carcinogenesis. **Ecancermedicalscience**, v. 9, p. 526, 2015.

GIEN, L. T.; BEAUCHEMIN, M. C.; THOMAS, G. Adenocarcinoma: A unique cervical cancer. **Gynecologic Oncology**, v. 116, n. 1, p. 140–146, 2010.

GILBERT, W. Why genes in pieces? **Nature**, v. 271, n. 5645, p. 501, 1978.

GOEDERT, L. et al. Long Noncoding RNAs in HPV-Induced Oncogenesis. **Advances in Tumor Virology**, v. 6, p. 1–9, 2016.

GUPTA, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. **Nature**, v. 464, n. 7291, p. 1071–6, 15 abr. 2010.

GUTSCHNER, T.; DIEDERICHS, S. The hallmarks of cancer: a long non-coding RNA point of view. **RNA biology**, v. 9, n. 6, p. 703–19, jun. 2012.

HAAS, B. J.; ZODY, M. C. Advancing RNA-Seq analysis. **Nature Biotechnology**, v. 28, n. 5, p. 421–423, 2010.

HALLEGGER, M.; LLORIAN, M.; SMITH, C. W. J. Alternative splicing: global insights. **FEBS Journal**, v. 277, n. 4, p. 856–866, fev. 2010.

HAMMOND, S. M. An overview of microRNAs. **Advanced Drug Delivery Reviews**, v. 87, p. 3–14, 2015.

HAN, S. et al. Long noncoding RNA identification: Comparing machine learning based tools for long noncoding transcripts discrimination. **BioMed Research International**, v. 2016, 2016.

HOSSEINI, E. S. et al. Dysregulated expression of long noncoding RNAs in gynecologic cancers. p. 1–13, 2017.

HOU, A. et al. LncRNA terminal differentiation-induced ncRNA (TINCR) sponges miR-302 to upregulate cyclin D1 in cervical squamous cell carcinoma (CSCC). **Human Cell**, v. 32, n. 4, p. 515–521, 2019.

RIBEIRO, B. et al. Desigualdades regionais na mortalidade por câncer de colo de útero no brasil: Tendências e projeções até o ano 2030. **Ciencia e Saude Coletiva**, v. 21, n. 1, p. 253–262, 2016.

ISHIDA, M.; MOORE, G. E. The role of imprinted genes in humans. **Molecular Aspects of Medicine**, v. 34, n. 4, p. 826–840, 2013.

IVANOVA, T. A. et al. Up-regulation of expression and lack of 5' CpG island hypermethylation of p16 INK4a in HPV-positive cervical carcinomas. **BMC cancer**, v. 7, p. 47, 14 mar. 2007.

KANG, Y. J. et al. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. **Nucleic Acids Research**, 2017.

KASHI, K. et al. Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1859, n. 1, p. 3–15, 2016.

KASPAR, H. G.; CRUM, C. P. The utility of immunohistochemistry in the differential diagnosis of gynecologic disorders. **Archives of Pathology and Laboratory Medicine**, v. 139, n. 1, p. 39–54, 2015.

KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature methods**, v. 12, n. 4, p. 357–60, abr. 2015.

KIM, J. et al. Human papillomavirus genotypes and cofactors causing cervical intraepithelial neoplasia and cervical cancer in Korean women. **International Journal of Gynecological Cancer**, v. 22, n. 9, p. 1570–1576, 2012.

KJELLBERG, L. et al. Smoking, diet, pregnancy and oral contraceptive use as risk factors for cervical intra-epithelial neoplasia in relation to human papillomavirus infection. **British Journal of Cancer**, v. 82, n. 7, p. 1332–1338, 2000.

KLAES, R. et al. Overexpression of p16INK4A as a specific marker for dysplastic and neoplastic epithelial cells of the cervix uteri. **International Journal of Cancer**, v. 92, n. 2, p. 276–284, 15 abr. 2001.

KOPP, F.; MENDELL, J. T. Functional Classification and Experimental Dissection of

Long Noncoding RNAs. **Cell**, v. 172, n. 3, p. 393–407, 2018.

KORI, M.; YALCIN ARGA, K. Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective. **PloS one**, v. 13, n. 7, p. e0200717, 2018.

KROL, J.; LOEDIGE, I.; FILIPOWICZ, W. The widespread regulation of microRNA biogenesis, function and decay. **Nature Reviews Genetics**, v. 11, n. 9, p. 597–610, 27 set. 2010.

LEE, J. W. et al. Altered MicroRNA expression in cervical carcinomas. **Clinical Cancer Research**, v. 14, n. 9, p. 2535–2542, 2008.

LI, F. et al. Spatiotemporal-specific lncRNAs in the brain, colon, liver and lung of macaque during development. **Molecular BioSystems**, v. 11, n. 12, p. 3253–3263, 10 nov. 2015.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics (Oxford, England)**, v. 25, n. 16, p. 2078–9, 15 ago. 2009.

LIU, C. N. et al. Upregulation of lncRNA CALML3-AS1 promotes cell proliferation and metastasis in cervical cancer via activation of the Wnt/β-catenin pathway. **European Review for Medical and Pharmacological Sciences**, v. 23, n. 13, p. 5611–5620, 2019.

LODISH, H.; BERK, A.; ZIPURSKY, S. Processing of rRNA and tRNA - Molecular Cell Biology. In: **4ª edição**. [s.l: s.n.].

LONG, Y. et al. How do lncRNAs regulate transcription? **Science Advances**, v. 3, n. 9, 2017.

LOPES, S. et al. Epigenetic modifications in an imprinting cluster are controlled by a hierarchy of DMRs suggesting long-range chromatin interactions. **Human Molecular Genetics**, v. 12, n. 3, p. 295–305, 2003.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology**, v. 15, n. 12, p. 550, 5 dez. 2014.

LUYKX, J. J. et al. Coding and non-coding RNA abnormalities in bipolar disorder. **Genes**, v. 10, n. 11, p. 1–14, 2019.

MANZONI, C. et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. **Briefings in Bioinformatics**, v. 19, n. 2, p. 286–302, 1 mar. 2018.

MCCREDIE, M. R. et al. Natural history of cervical neoplasia and risk of invasive cancer in women with cervical intraepithelial neoplasia 3: a retrospective cohort study. **The Lancet Oncology**, v. 9, n. 5, p. 425–434, 2008.

National Cancer Institute. Disponível em: <www.cancer.gov>.

NESTEROVA, T. B. et al. **Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequenceGenome Research**, maio 2001.

PAJARES, M. J. et al. Alternative splicing: an emerging topic in molecular and clinical oncology. **Lancet Oncology**, v. 8, n. 4, p. 349–357, 2007.

PERTEA, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. **Nature biotechnology**, v. 33, n. 3, p. 290–5, mar. 2015.

PERTEA, M. et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. **Nature Protocols**, v. 11, n. 9, p. 1650–1667, 11 set. 2016.

PICKRELL, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. **Nature**, v. 464, n. 7289, p. 768–772, 2010.

PONJAVIC, J. et al. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. **PLoS genetics**, v. 5, n. 8, p. e1000617, ago. 2009.

PROUDFOOT, N. J.; FURGER, A.; DYE, M. J. Integrating mRNA Processing with Transcription. **Cell**, v. 108, p. 501–512, 2002.

QIN, S. et al. Identifying Molecular Markers of Cervical Cancer Based on Competing Endogenous RNA Network Analysis. 2019.

QUINN, J. J.; CHANG, H. Y. Unique features of long non-coding RNA biogenesis and function. **Nature Reviews Genetics**, v. 17, n. 1, p. 47–62, 15 jan. 2016.

REIK, W.; WALTER, J. Genomic imprinting: parental influence on the genome : Article : Nature Reviews Genetics. **Nature Reviews Genetics**, v. 2, n. 1, p. 21–32, 2001.

RONNETT, B. M. Endocervical adenocarcinoma: selected diagnostic challenges. **Modern Pathology**, v. 29, n. S1, p. S12–S28, 30 jan. 2016.

SASAKI, Y. T. F. et al. Identification and characterization of human non-coding RNAs with tissue-specific expression. **Biochemical and Biophysical Research Communications**, v. 357, n. 4, p. 991–996, 15 jun. 2007.

SCHLECHT, N. F. et al. Human Papillomavirus Infection and Time to Progression and Regression of Cervical Intraepithelial Neoplasia. v. 95, n. 17, 2003.

SCHMITT, M. et al. Diagnosing Cervical Cancer and High-Grade Precursors by HPV16 Transcription Patterns. p. 249–257, 2010.

SHIMADA, M. et al. Comparison of the outcome between cervical adenocarcinoma and squamous cell carcinoma patients with adjuvant radiotherapy following radical surgery: SGSG/TGCU Intergroup Surveillance. **Molecular and clinical oncology**, v. 1, n. 4, p. 780–784, jul. 2013.

SONG, L. et al. A multi-sample approach increases the accuracy of transcript assembly. **Nature Communications**, v. 10, n. 5000, 1 dez. 2019.

SONG, L.; SABUNCIYAN, S.; FLOREA, L. CLASS2: Accurate and efficient splice variant annotation from RNA-seq reads. **Nucleic Acids Research**, v. 44, n. 10, p. 1–15, 2016.

SONG, X. et al. Alternative splicing in cancers: From aberrant regulation to new therapeutics. **Seminars in Cell and Developmental Biology**, v. 75, p. 13–22, 2018.

SONG, X. H. et al. Expression of a novel alternatively spliced variant of NADP(H)-dependent retinol dehydrogenase/reductase with deletion of exon 3 in cervical squamous carcinoma. **International Journal of Cancer**, v. 120, n. 8, p. 1618–1626, 2007.

TEAM, R. C. **R: A language and environment for statistical computing**Vienna, AustriaR Foundation for Statistical Computing, , 2018. Disponível em: <https://www.r-project.org/>

THORVALDSDÓTTIR, H.; ROBINSON, J. T.; MESIROV, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. **Briefings in Bioinformatics**, 2013.

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature Biotechnology**, v. 28, n. 5, p. 511–515, 2 maio 2010.

TRAPNELL, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, v. 7, n. 3, p. 562–578, 1 mar. 2012.

TSAI, Y. S. et al. Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. **Oncotarget**, v. 6, n. 9, p. 6825–6839, 2015.

UK, **Cancer Research UK**. Disponível em: <https://www.cancerresearchuk.org/about-cancer/cervical-cancer>.

ULITSKY, I.; BARTEL, D. P. lincRNAs: Genomics, Evolution, and Mechanisms. **Cell**, v. 154, n. 1, p. 26–46, 2013.

URBANSKI, L. M.; LECLAIR, N.; ANCZUKÓW, O. Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. **Wiley Interdisciplinary Reviews: RNA**, v. 9, n. 4, p. e1476, jul. 2018.

VALENTI, G. et al. Tumor markers of uterine cervical cancer: A new scenario to guide surgical practice? **Updates in Surgery**, v. 69, n. 4, p. 441–449, 2017.

VENABLES, J. P. Unbalanced alternative splicing and its significance in cancer. **BioEssays**, v. 28, n. 4, p. 378–386, 2006.

VENTURINI, L. et al. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. **GigaScience**, v. 7, n. 8, 1 ago. 2018.

WANG, H. et al. Identification of novel long non-coding and circular RNAs in human papillomavirus-mediated cervical cancer. **Frontiers in Microbiology**, v. 8, n. SEP, p. 1720, 2017.

WANG, J.-L. et al. p16INK4A and p14ARF expression pattern by immunohistochemistry in human papillomavirus-related cervical neoplasia. **Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc**, v. 18, n. 5, p. 629–37, 22 maio 2005.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: A revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, 2009.

WARD, M. et al. Conservation and tissue-specific transcription patterns of long noncoding RNAs. **Journal of human transcriptome**, v. 1, n. 1, p. 2–9, 1 jan. 2015.

WENTZENSEN, N.; VON KNEBEL DOEBERITZ, M. Biomarkers in cervical cancer screening. **Disease markers**, v. 23, n. 4, p. 315–30, 2007.

WILLIAMS, N. L. et al. Adenocarcinoma of the Cervix : Should We Treat It Differently ? 2015.

WITTEN, D. et al. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. **BMC Biology**, v. 8, p. 1–14, 2010.

WOERNER, S. M. et al. Expression of CD44 Splice Variants in Normal, Dysplastic, and Neoplastic Cervical Epithelium. **Clinical Cancer Research**, v. 1, n. October, p. 1125–1132, 1995.

World Health Organization. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer>.

XIAOGUANG, W. et al. LncRNA MEG3 has anti-activity effects of cervical cancer. **Biomedicine et Pharmacotherapy**, v. 94, p. 636–643, 2017.

ZERBINO, D. R. et al. Ensembl 2018. **Nucleic Acids Research**, v. 46, n. D1, p. D754–D761, 4 jan. 2018.

ZHOU, J. et al. Postoperative clinicopathological factors affecting cervical adenocarcinoma. n. October 2006, p. 1–5, 2018.

ZOU, D. et al. Biological databases for human research. **Genomics, Proteomics and Bioinformatics**, v. 13, n. 1, p. 55–63, 2015.

APÊNDICE

MEDICINA
USP
COMITÊ DE ÉTICA EM PESQUISA

APROVAÇÃO

O Comitê de Ética em Pesquisa da Faculdade de Medicina da Universidade de São Paulo, em sessão de 17/02/2016, APROVOU o Protocolo de Pesquisa nº 033/16 intitulado: "AVALIAÇÃO DO PERFIL DE EXPRESSÃO GÊNICA DE DOIS SUBTIPOS DE CÂNCER DE COLO UTERINO: CARCINOMA ESCAMOSO E ADENOCARCINOMA" apresentado pelo Departamento de RADIOLOGIA E ONCOLOGIA

Cabe ao pesquisador elaborar e apresentar ao CEP-FMUSP, os relatórios parciais e final sobre a pesquisa (Resolução do Conselho Nacional de Saúde nº 466/12, inciso IX.2, letra "c").

Pesquisador (a) Responsável: Luisa Lina Villa
Pesquisador (a) Executante: Maria Luiza Nogueira Genta

CEP-FMUSP, 19 de Fevereiro de 2016.

Profa. Dra. Maria Aparecida Azevedo Koike Folgueira
Coordenador
Comitê de Ética em Pesquisa

Comitê de Ética em Pesquisa da Faculdade de Medicina
e-mail: cep.fm@usp.br

Figura 25: Aprovação do Comitê de Ética em Pesquisa da Faculdade de Medicina da USP sob protocolo de pesquisa n° 033/16.